UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Aplicación de técnicas estadísticas multivariantes y de aprendizaje automático en el diagnóstico de cáncer de próstata a través de imágenes médicas

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Gironés Sangüesa, Raquel

Tutor/a: Carot Sierra, José Miguel

Director/a Experimental: CERDA ALBERICH, LEONOR

CURSO ACADÉMICO: 2023/2024

Resumen

El cáncer de próstata es una de las principales causas de morbilidad y mortalidad entre la población masculina a nivel mundial. Su diagnóstico temprano y preciso es crucial para mejorar las tasas de supervivencia y la calidad de vida de los pacientes. En este contexto, las imágenes médicas juegan un papel fundamental, proporcionando una visión detallada y no invasiva del tejido prostático. Sin embargo, la interpretación de estas imágenes es compleja y sujeta a variabilidad inter-observador. El presente trabajo se centra en la aplicación de técnicas estadísticas avanzadas y algoritmos de aprendizaje automático para la predicción y evaluación del desarrollo del cáncer de próstata utilizando imágenes médicas. La metodología que se ha empleado incluye la selección y preprocesamiento de un conjunto de datos de imágenes médicas, la integración de distintas fuentes de datos (datos clínicos y distintos procesados de imágenes) y la aplicación de técnicas estadísticas multivariantes de reducción de la dimensión y clasificación.

Palabras clave: Análisis multivariante, Aprendizaje automático, Imagen médica, Cáncer de próstata, Clasificación

Resum

El càncer de pròstata és una de les principals causes de morbiditat i mortalitat entre la població masculina a nivell mundial. El seu diagnòstic primerenc i precís és crucial per a millorar les taxes de supervivència i la qualitat de vida dels pacients. En este context, les imatges mèdiques juguen un paper fonamental, proporcionant una visió detallada i no invasiva del teixit prostàtic. No obstant això, la interpretació d'estes imatges és complexa i subjecta a variabilitat inter-observador. El present treball se centra en l'aplicació de tècniques estadístiques avançades i algorismes d'aprenentatge automàtic per a la predicció i avaluació del desenvolupament del càncer de pròstata utilitzant imatges mèdiques. La metodologia que s'ha emprat inclou la selecció i preprocessament d'un conjunt de dades d'imatges mèdiques, la integració de diferents fonts de dades (dades clíniques i diferents processaments d'imatges) i l'aplicació de tècniques estadístiques multivariants de reducció de la dimensió i classificació.

Paraules clau: Anàlisi multivariant, Aprenentatge automàtic, Imatge mèdica, Càncer de pròstata, Classificació

Abstract

Prostate cancer is one of the leading causes of morbidity and mortality among the male population worldwide. Early and accurate diagnosis is crucial for improving survival rates and the quality of life of patients. In this context, medical imaging plays a crucial role, providing a detailed and non-invasive view of prostate tissue. However, the interpretation of these images is complex and subject to inter-observer variability. This work focuses on the application of advanced statistical techniques and machine learning algorithms for the prediction and assessment of prostate cancer development using medical images. The methodology employed includes the selection and preprocessing of a *dataset* of medical images, the integration of different data sources (clinical data and various image processing outputs), and the application of multivariate statistical techniques for dimensionality reduction and classification.

Key words: Multivariate Analysis, Machine Learning, Medical Imaging, Prostate Cancer, Classification

Índice general

Índice general v
Índice de figuras ix
Índice de tablas xxi

1 Introducción 1
1.1 Motivación . . . 2
1.2 Objetivos . . . 3
1.3 Impacto esperado . . . 3
1.3.1 Avance en la Medicina de Precisión . . . 3
1.3.2 Optimización de Recursos . . . 3
1.3.3 Mejora en la Eficiencia Diagnóstica . . . 4
1.3.4 Empoderamiento del paciente . . . 4
1.3.5 Contribución a la Investigación del Cáncer de Próstata . . . 4
1.4 Metodología . . . 4
1.5 Estructura del TFG . . . 5

2 Antecedentes 7
2.1 Radiómica y su Relevancia en la Medicina . . . 7
2.2 Machine Learning en Radiómica . . . 7
2.3 Métodos de Clasificación en Cáncer de Próstata mediante Radiómica . . . 8
2.4 Propuesta . . . 9

3 Base de Datos 11
3.1 Obtención de los datos . . . 11
3.2 Variables radiómicas . . . 11
3.3 Variables clínicas . . . 12

4 Evaluación 15
4.1 Preprocesamiento de datos . . . 15
4.1.1 Escalado . . . 15
4.1.2 División del conjunto de datos . . . 15
4.1.3 Remuestreo . . . 16
4.2 Modelos de clasificación . . . 17
4.2.1 Clasificadores lineales . . . 18
4.2.2 Árboles de decisión . . . 19
4.2.3 Clasificadores de ensamblado . . . 19
4.2.4 Clasificadores de máquina de soporte vectorial . . . 19
4.2.5 Clasificadores de vecinos más cercanos . . . 19
4.2.6 Clasificadores de naive Bayes . . . 20
4.2.7 Clasificadores de redes neuronales . . . 20
4.3 Entrenamiento y evaluación de modelos . . . 20
4.3.1 *Accuracy* . . . 21
4.3.2 *Recall* . . . 21
4.3.3 *Precision* . . . 21

4.3.4 *F1 score* . . . 21
5 Resultados **23**
5.1 CLASIFICACIÓN 1 - 1,2 vs 3,4,5 . . . 25
5.1.1 ADASYN . . . 25
5.1.2 SMOTE . . . 30
5.2 CLASIFICACIÓN 2 - 1 vs 2,3 vs 4,5 . . . 33
5.2.1 ADASYN . . . 33
5.2.2 SMOTE . . . 36
5.3 CLASIFICACIÓN 3-1 vs 2 vs 3,4,5 . . . 38
5.3.1 ADASYN . . . 39
5.3.2 SMOTE . . . 41
5.4 CLASIFICACIÓN 4-1,2,3 vs 4,5 . . . 43
5.4.1 ADASYN . . . 44
5.4.2 SMOTE . . . 46
6 Conclusiones **49**
6.1 Conclusiones del trabajo . . . 49
6.2 Conclusiones personales . . . 50
6.3 Análisis marco legal y ético . . . 50
6.4 Relación del trabajo con los estudios cursados . . . 51
6.5 Legado . . . 51
6.6 Trabajo futuro . . . 52
6.6.1 Exploración de Métricas Adicionales . . . 52
6.6.2 Investigación sobre Métodos Avanzados de Selección de Características . . . 52
6.6.3 Exploración de Otros Modelos . . . 52
Referencias . . . 53

Apéndices

A Objetivos de desarrollo sostenible (ODS) **55**
A.1 Objetivo 3: Salud y Bienestar . . . 56
A.2 Objetivo 9: Industria, Innovación e Infraestructura . . . 56
A.3 Objetivo 10: Reducción de las Desigualdades . . . 56
A.4 Objetivo 17: Alianzas para lograr los objetivos . . . 56
B Construcción del conjunto de datos **57**
B.1 Introducción . . . 57
B.2 Segmentación . . . 57
B.3 Co-registro . . . 57
B.4 Evaluación de la calidad de la segmentación . . . 58
B.5 Extracción de características radiómicas . . . 58
B.6 Extracción de características profundas . . . 59
B.7 Características clínicas . . . 59
B.8 Construcción del *dataset* . . . 59
B.9 Pipeline de preprocesamiento . . . 59
B.10 Entrenamiento . . . 60
B.11 Postprocesamiento . . . 60
B.12 Resultados . . . 60
C Análisis exploratorio de datos **63**
C.1 Edad (age) . . . 63
C.2 PSA (psas_0_total) . . . 64
C.3 PI-RADS (lesions_0_pi_rads) . . . 65
C.4 Gleason (lesions_0_gleason1.1 , lesions_0_gleason2.1) . . . 66

C.5 ISUP 67
C.6 TZ, PZ, CZ, AZ 68
C.7 Correlaciones 69

D Clasificación **71**
D.1 Clasificación 1 71
D.1.1 ADASYN 71
D.1.2 SMOTE 83
D.2 Clasificación 2 95
D.2.1 ADASYN 95
D.2.2 SMOTE 107
D.3 Clasificación 3 119
D.3.1 ADASYN 119
D.3.2 SMOTE 131
D.4 Clasificación 4 141
D.4.1 ADASYN 141
D.4.2 SMOTE 153

Índice de figuras

1.1 Metodología en forma de diagrama de flujo. 5

4.1 Desbalanceo de clases 16

5.1 Proporción de las clases en la Clasificación 1 25
5.2 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 1 con ADASYN 25
5.3 Matriz de confusión para Clasificación 1 con ADASYN 27
5.4 Variables más importantes para Clasificación 1 con ADASYN 29
5.5 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 1 con SMOTE 30
5.6 Matriz de confusión para Clasificación 1 con SMOTE 31
5.7 Variables más importantes para Clasificación 1 con SMOTE 32
5.8 Proporción de las clases en la Clasificación 2 33
5.9 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 2 con ADASYN 33
5.10 Matriz de confusión para Clasificación 2 con ADASYN 34
5.11 Variables más importantes para Clasificación 2 con ADASYN 35
5.12 Variables más importantes para Clasificación 2 con ADASYN 35
5.13 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 2 con SMOTE 36
5.14 Matriz de confusión para Clasificación 2 con SMOTE 37
5.15 Variables más importantes para Clasificación 2 con SMOTE 38
5.16 Proporción de las clases en la Clasificación 3 38
5.17 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 3 con ADASYN 39
5.18 Matriz de confusión para Clasificación 3 con ADASYN 40
5.19 Variables más importantes para Clasificación 3 con ADASYN 41
5.20 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 3 con SMOTE 41
5.21 Matriz de confusión para Clasificación 3 con SMOTE 42
5.22 Variables más importantes para Clasificación 3 con SMOTE 43
5.23 Proporción de las clases en la Clasificación 4 43
5.24 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 4 con ADASYN 44
5.25 Matriz de confusión para Clasificación 4 con ADASYN 45
5.26 Variables más importantes para Clasificación 4 con ADASYN 45
5.27 Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 4 con SMOTE 46
5.28 Matriz de confusión para Clasificación 4 con SMOTE 47
5.29 Variables más importantes para Clasificación 4 con SMOTE 47

C.1 Distribución de Edades 63

C.2 Distribución de PSA . . . 64
C.3 Frecuencia de la variable PI-RADS . . . 65
C.4 Frecuencia de los Gleason . . . 66
C.5 Frecuencia de ISUP . . . 67
C.6 Frecuencia de las zonas afectadas . . . 68
C.7 Matriz de correlación . . . 69

D.1 Logistic Regression, Fold 1, C1, ADASYN . . . 71
D.2 Logistic Regression, Fold 2, C1, ADASYN . . . 71
D.2 Logistic Regression, Fold 3, C1, ADASYN . . . 71
D.3 Logistic Regression, Fold 4, C1, ADASYN . . . 71
D.3 Logistic Regression, Fold 5, C1, ADASYN . . . 72
D.4 Logistic Regression, Final, C1, ADASYN . . . 72
D.5 Logistic Regression Lasso, Fold 1, C1, ADASYN . . . 72
D.6 Logistic Regression Lasso, Fold 2, C1, ADASYN . . . 72
D.6 Logistic Regression Lasso, Fold 3, C1, ADASYN . . . 72
D.7 Logistic Regression Lasso, Fold 4, C1, ADASYN . . . 72
D.7 Logistic Regression Lasso, Fold 5, C1, ADASYN . . . 73
D.8 Logistic Regression Lasso, Final, C1, ADASYN . . . 73
D.9 Ridge, Fold 1, C1, ADASYN . . . 73
D.10 Ridge, Fold 2, C1, ADASYN . . . 73
D.10 Ridge, Fold 3, C1, ADASYN . . . 73
D.11 Ridge, Fold 4, C1, ADASYN . . . 73
D.11 Ridge, Fold 5, C1, ADASYN . . . 74
D.12 Ridge, Final, C1, ADASYN . . . 74
D.13 SDG, Fold 1, C1, ADASYN . . . 74
D.14 SDG, Fold 2, C1, ADASYN . . . 74
D.14 SDG, Fold 3, C1, ADASYN . . . 74
D.15 SDG, Fold 4, C1, ADASYN . . . 74
D.15 SDG, Fold 5, C1, ADASYN . . . 75
D.16 SDG, Final, C1, ADASYN . . . 75
D.17 Decision Tree, Fold 1, C1, ADASYN . . . 75
D.18 Decision Tree, Fold 2, C1, ADASYN . . . 75
D.18 Decision Tree, Fold 3, C1, ADASYN . . . 75
D.19 Decision Tree, Fold 4, C1, ADASYN . . . 75
D.19 Decision Tree, Fold 5, C1, ADASYN . . . 76
D.20 Decision Tree, Final, C1, ADASYN . . . 76
D.21 Random Forest, Fold 1, C1, ADASYN . . . 76
D.22 Random Forest, Fold 2, C1, ADASYN . . . 76
D.22 Random Forest, Fold 3, C1, ADASYN . . . 76
D.23 Random Forest, Fold 4, C1, ADASYN . . . 76
D.23 Random Forest, Fold 5, C1, ADASYN . . . 77
D.24 Random Forest, Final, C1, ADASYN . . . 77
D.25 AdaBoost, Fold 1, C1, ADASYN . . . 77
D.26 AdaBoost, Fold 2, C1, ADASYN . . . 77
D.26 AdaBoost, Fold 3, C1, ADASYN . . . 77
D.27 AdaBoost, Fold 4, C1, ADASYN . . . 77
D.27 AdaBoost, Fold 5, C1, ADASYN . . . 78
D.28 AdaBoost, Final, C1, ADASYN . . . 78
D.29 Gradient Boosting, Fold 1, C1, ADASYN . . . 78
D.30 Gradient Boosting, Fold 2, C1, ADASYN . . . 78

D.30 Gradient Boosting, Fold 3, C1, ADASYN . . . 78
D.31 Gradient Boosting, Fold 4, C1, ADASYN . . . 78
D.31 Gradient Boosting, Fold 5, C1, ADASYN . . . 79
D.32 Gradient Boosting, Final, C1, ADASYN . . . 79
D.33 SVC, Fold 1, C1, ADASYN . . . 79
D.34 SVC, Fold 2, C1, ADASYN . . . 79
D.34 SVC, Fold 3, C1, ADASYN . . . 79
D.35 SVC, Fold 4, C1, ADASYN . . . 79
D.35 SVC, Fold 5, C1, ADASYN . . . 80
D.36 SVC, Final, C1, ADASYN . . . 80
D.37 k-Nearest Neighbors, Fold 1, C1, ADASYN . . . 80
D.38 k-Nearest Neighbors, Fold 2, C1, ADASYN . . . 80
D.38 k-Nearest Neighbors, Fold 3, C1, ADASYN . . . 80
D.39 k-Nearest Neighbors, Fold 4, C1, ADASYN . . . 80
D.39 k-Nearest Neighbors, Fold 5, C1, ADASYN . . . 81
D.40 k-Nearest Neighbors, Final, C1, ADASYN . . . 81
D.41 Gaussian Naive Bayes, Fold 1, C1, ADASYN . . . 81
D.42 Gaussian Naive Bayes, Fold 2, C1, ADASYN . . . 81
D.42 Gaussian Naive Bayes, Fold 3, C1, ADASYN . . . 81
D.43 Gaussian Naive Bayes, Fold 4, C1, ADASYN . . . 81
D.43 Gaussian Naive Bayes, Fold 5, C1, ADASYN . . . 82
D.44 Gaussian Naive Bayes, Final, C1, ADASYN . . . 82
D.45 MLP, Fold 1, C1, ADASYN . . . 82
D.46 MLP, Fold 2, C1, ADASYN . . . 82
D.46 MLP, Fold 3, C1, ADASYN . . . 82
D.47 MLP, Fold 4, C1, ADASYN . . . 82
D.47 MLP, Fold 5, C1, ADASYN . . . 83
D.48 MLP, Final, C1, ADASYN . . . 83
D.49 Logistic Regression, Fold 1, C1, SMOTE . . . 83
D.50 Logistic Regression, Fold 2, C1, SMOTE . . . 83
D.50 Logistic Regression, Fold 3, C1, SMOTE . . . 83
D.51 Logistic Regression, Fold 4, C1, SMOTE . . . 83
D.51 Logistic Regression, Fold 5, C1, SMOTE . . . 84
D.52 Logistic Regression, Final, C1, SMOTE . . . 84
D.53 Logistic Regression Lasso, Fold 1, C1, SMOTE . . . 84
D.54 Logistic Regression Lasso, Fold 2, C1, SMOTE . . . 84
D.54 Logistic Regression Lasso, Fold 3, C1, SMOTE . . . 84
D.55 Logistic Regression Lasso, Fold 4, C1, SMOTE . . . 84
D.55 Logistic Regression Lasso, Fold 5, C1, SMOTE . . . 85
D.56 Logistic Regression Lasso, Final, C1, SMOTE . . . 85
D.57 Ridge, Fold 1, C1, SMOTE . . . 85
D.58 Ridge, Fold 2, C1, SMOTE . . . 85
D.58 Ridge, Fold 3, C1, SMOTE . . . 85
D.59 Ridge, Fold 4, C1, SMOTE . . . 85
D.59 Ridge, Fold 5, C1, SMOTE . . . 86
D.60 Ridge, Final, C1, SMOTE . . . 86
D.61 SDG, Fold 1, C1, SMOTE . . . 86
D.62 SDG, Fold 2, C1, SMOTE . . . 86
D.62 SDG, Fold 3, C1, SMOTE . . . 86
D.63 SDG, Fold 4, C1, SMOTE . . . 86
D.63 SDG, Fold 5, C1, SMOTE . . . 87

D.64 SDG, Final, C1, SMOTE . . . 87
D.65 Decision Tree, Fold 1, C1, SMOTE . . . 87
D.66 Decision Tree, Fold 2, C1, SMOTE . . . 87
D.66 Decision Tree, Fold 3, C1, SMOTE . . . 87
D.67 Decision Tree, Fold 4, C1, SMOTE . . . 87
D.67 Decision Tree, Fold 5, C1, SMOTE . . . 88
D.68 Decision Tree, Final, C1, SMOTE . . . 88
D.69 Random Forest, Fold 1, C1, SMOTE . . . 88
D.70 Random Forest, Fold 2, C1, SMOTE . . . 88
D.70 Random Forest, Fold 3, C1, SMOTE . . . 88
D.71 Random Forest, Fold 4, C1, SMOTE . . . 88
D.71 Random Forest, Fold 5, C1, SMOTE . . . 89
D.72 Random Forest, Final, C1, SMOTE . . . 89
D.73 AdaBoost, Fold 1, C1, SMOTE . . . 89
D.74 AdaBoost, Fold 2, C1, SMOTE . . . 89
D.74 AdaBoost, Fold 3, C1, SMOTE . . . 89
D.75 AdaBoost, Fold 4, C1, SMOTE . . . 89
D.75 AdaBoost, Fold 5, C1, SMOTE . . . 90
D.76 AdaBoost, Final, C1, SMOTE . . . 90
D.77 Gradient Boosting, Fold 1, C1, SMOTE . . . 90
D.78 Gradient Boosting, Fold 2, C1, SMOTE . . . 90
D.78 Gradient Boosting, Fold 3, C1, SMOTE . . . 90
D.79 Gradient Boosting, Fold 4, C1, SMOTE . . . 90
D.79 Gradient Boosting, Fold 5, C1, SMOTE . . . 91
D.80 Gradient Boosting, Final, C1, SMOTE . . . 91
D.81 SVC, Fold 1, C1, SMOTE . . . 91
D.82 SVC, Fold 2, C1, SMOTE . . . 91
D.82 SVC, Fold 3, C1, SMOTE . . . 91
D.83 SVC, Fold 4, C1, SMOTE . . . 91
D.83 SVC, Fold 5, C1, SMOTE . . . 92
D.84 SVC, Final, C1, SMOTE . . . 92
D.85 k-Nearest Neighbors, Fold 1, C1, SMOTE . . . 92
D.86 k-Nearest Neighbors, Fold 2, C1, SMOTE . . . 92
D.86 k-Nearest Neighbors, Fold 3, C1, SMOTE . . . 92
D.87 k-Nearest Neighbors, Fold 4, C1, SMOTE . . . 92
D.87 k-Nearest Neighbors, Fold 5, C1, SMOTE . . . 93
D.88 k-Nearest Neighbors, Final, C1, SMOTE . . . 93
D.89 Gaussian Naive Bayes, Fold 1, C1, SMOTE . . . 93
D.90 Gaussian Naive Bayes, Fold 2, C1, SMOTE . . . 93
D.90 Gaussian Naive Bayes, Fold 3, C1, SMOTE . . . 93
D.91 Gaussian Naive Bayes, Fold 4, C1, SMOTE . . . 93
D.91 Gaussian Naive Bayes, Fold 5, C1, SMOTE . . . 94
D.92 Gaussian Naive Bayes, Final, C1, SMOTE . . . 94
D.93 MLP, Fold 1, C1, SMOTE . . . 94
D.94 MLP, Fold 2, C1, SMOTE . . . 94
D.94 MLP, Fold 3, C1, SMOTE . . . 94
D.95 MLP, Fold 4, C1, SMOTE . . . 94
D.95 MLP, Fold 5, C1, SMOTE . . . 95
D.96 MLP, Final, C1, SMOTE . . . 95
D.97 Logistic Regression, Fold 1, C2, ADASYN . . . 95
D.98 Logistic Regression, Fold 2, C2, ADASYN . . . 95

D.98 Logistic Regression, Fold 3, C2, ADASYN . . . 95
D.99 Logistic Regression, Fold 4, C2, ADASYN . . . 95
D.99 Logistic Regression, Fold 5, C2, ADASYN . . . 96
D.100 Logistic Regression, Final, C2, ADASYN . . . 96
D.101 Logistic Regression Lasso, Fold 1, C2, ADASYN . . . 96
D.102 Logistic Regression Lasso, Fold 2, C2, ADASYN . . . 96
D.102 Logistic Regression Lasso, Fold 3, C2, ADASYN . . . 96
D.103 Logistic Regression Lasso, Fold 4, C2, ADASYN . . . 96
D.103 Logistic Regression Lasso, Fold 5, C2, ADASYN . . . 97
D.104 Logistic Regression Lasso, Final, C2, ADASYN . . . 97
D.105 Ridge, Fold 1, C2, ADASYN . . . 97
D.106 Ridge, Fold 2, C2, ADASYN . . . 97
D.106 Ridge, Fold 3, C2, ADASYN . . . 97
D.107 Ridge, Fold 4, C2, ADASYN . . . 97
D.107 Ridge, Fold 5, C2, ADASYN . . . 98
D.108 Ridge, Final, C2, ADASYN . . . 98
D.109 SDG, Fold 1, C2, ADASYN . . . 98
D.110 SDG, Fold 2, C2, ADASYN . . . 98
D.110 SDG, Fold 3, C2, ADASYN . . . 98
D.111 SDG, Fold 4, C2, ADASYN . . . 98
D.111 SDG, Fold 5, C2, ADASYN . . . 99
D.112 SDG, Final, C2, ADASYN . . . 99
D.113 Decision Tree, Fold 1, C2, ADASYN . . . 99
D.114 Decision Tree, Fold 2, C2, ADASYN . . . 99
D.114 Decision Tree, Fold 3, C2, ADASYN . . . 99
D.115 Decision Tree, Fold 4, C2, ADASYN . . . 99
D.115 Decision Tree, Fold 5, C2, ADASYN . . . 100
D.116 Decision Tree, Final, C2, ADASYN . . . 100
D.117 Random Forest, Fold 1, C2, ADASYN . . . 100
D.118 Random Forest, Fold 2, C2, ADASYN . . . 100
D.118 Random Forest, Fold 3, C2, ADASYN . . . 100
D.119 Random Forest, Fold 4, C2, ADASYN . . . 100
D.119 Random Forest, Fold 5, C2, ADASYN . . . 101
D.120 Random Forest, Final, C2, ADASYN . . . 101
D.121 AdaBoost, Fold 1, C2, ADASYN . . . 101
D.122 AdaBoost, Fold 2, C2, ADASYN . . . 101
D.122 AdaBoost, Fold 3, C2, ADASYN . . . 101
D.123 AdaBoost, Fold 4, C2, ADASYN . . . 101
D.123 AdaBoost, Fold 5, C2, ADASYN . . . 102
D.124 AdaBoost, Final, C2, ADASYN . . . 102
D.125 Gradient Boosting, Fold 1, C2, ADASYN . . . 102
D.126 Gradient Boosting, Fold 2, C2, ADASYN . . . 102
D.126 Gradient Boosting, Fold 3, C2, ADASYN . . . 102
D.127 Gradient Boosting, Fold 4, C2, ADASYN . . . 102
D.127 Gradient Boosting, Fold 5, C2, ADASYN . . . 103
D.128 Gradient Boosting, Final, C2, ADASYN . . . 103
D.129 SVC, Fold 1, C2, ADASYN . . . 103
D.130 SVC, Fold 2, C2, ADASYN . . . 103
D.130 SVC, Fold 3, C2, ADASYN . . . 103
D.131 SVC, Fold 4, C2, ADASYN . . . 103
D.131 SVC, Fold 5, C2, ADASYN . . . 104

D.132 SVC, Final, C2, ADASYN . . . 104
D.133 k-Nearest Neighbors, Fold 1, C2, ADASYN . . . 104
D.134 k-Nearest Neighbors, Fold 2, C2, ADASYN . . . 104
D.134 k-Nearest Neighbors, Fold 3, C2, ADASYN . . . 104
D.135 k-Nearest Neighbors, Fold 4, C2, ADASYN . . . 104
D.135 k-Nearest Neighbors, Fold 5, C2, ADASYN . . . 105
D.136 k-Nearest Neighbors, Final, C2, ADASYN . . . 105
D.137 Gaussian Naive Bayes, Fold 1, C2, ADASYN . . . 105
D.138 Gaussian Naive Bayes, Fold 2, C2, ADASYN . . . 105
D.138 Gaussian Naive Bayes, Fold 3, C2, ADASYN . . . 105
D.139 Gaussian Naive Bayes, Fold 4, C2, ADASYN . . . 105
D.139 Gaussian Naive Bayes, Fold 5, C2, ADASYN . . . 106
D.140 Gaussian Naive Bayes, Final, C2, ADASYN . . . 106
D.141 MLP, Fold 1, C2, ADASYN . . . 106
D.142 MLP, Fold 2, C2, ADASYN . . . 106
D.142 MLP, Fold 3, C2, ADASYN . . . 106
D.143 MLP, Fold 4, C2, ADASYN . . . 106
D.143 MLP, Fold 5, C2, ADASYN . . . 107
D.144 MLP, Final, C2, ADASYN . . . 107
D.145 Logistic Regression, Fold 1, C2, SMOTE . . . 107
D.146 Logistic Regression, Fold 2, C2, SMOTE . . . 107
D.146 Logistic Regression, Fold 3, C2, SMOTE . . . 107
D.147 Logistic Regression, Fold 4, C2, SMOTE . . . 107
D.147 Logistic Regression, Fold 5, C2, SMOTE . . . 108
D.148 Logistic Regression, Final, C2, SMOTE . . . 108
D.149 Logistic Regression Lasso, Fold 1, C2, SMOTE . . . 108
D.150 Logistic Regression Lasso, Fold 2, C2, SMOTE . . . 108
D.150 Logistic Regression Lasso, Fold 3, C2, SMOTE . . . 108
D.151 Logistic Regression Lasso, Fold 4, C2, SMOTE . . . 108
D.151 Logistic Regression Lasso, Fold 5, C2, SMOTE . . . 109
D.152 Logistic Regression Lasso, Final, C2, SMOTE . . . 109
D.153 Ridge, Fold 1, C2, SMOTE . . . 109
D.154 Ridge, Fold 2, C2, SMOTE . . . 109
D.154 Ridge, Fold 3, C2, SMOTE . . . 109
D.155 Ridge, Fold 4, C2, SMOTE . . . 109
D.155 Ridge, Fold 5, C2, SMOTE . . . 110
D.156 Ridge, Final, C2, SMOTE . . . 110
D.157 SDG, Fold 1, C2, SMOTE . . . 110
D.158 SDG, Fold 2, C2, SMOTE . . . 110
D.158 SDG, Fold 3, C2, SMOTE . . . 110
D.159 SDG, Fold 4, C2, SMOTE . . . 110
D.159 SDG, Fold 5, C2, SMOTE . . . 111
D.160 SDG, Final, C2, SMOTE . . . 111
D.161 Decision Tree, Fold 1, C2, SMOTE . . . 111
D.162 Decision Tree, Fold 2, C2, SMOTE . . . 111
D.162 Decision Tree, Fold 3, C2, SMOTE . . . 111
D.163 Decision Tree, Fold 4, C2, SMOTE . . . 111
D.163 Decision Tree, Fold 5, C2, SMOTE . . . 112
D.164 Decision Tree, Final, C2, SMOTE . . . 112
D.165 Random Forest, Fold 1, C2, SMOTE . . . 112
D.166 Random Forest, Fold 2, C2, SMOTE . . . 112

D.166 Random Forest, Fold 3, C2, SMOTE . . . 112
D.167 Random Forest, Fold 4, C2, SMOTE . . . 112
D.167 Random Forest, Fold 5, C2, SMOTE . . . 113
D.168 Random Forest, Final, C2, SMOTE . . . 113
D.169 AdaBoost, Fold 1, C2, SMOTE . . . 113
D.170 AdaBoost, Fold 2, C2, SMOTE . . . 113
D.170 AdaBoost, Fold 3, C2, SMOTE . . . 113
D.171 AdaBoost, Fold 4, C2, SMOTE . . . 113
D.171 AdaBoost, Fold 5, C2, SMOTE . . . 114
D.172 AdaBoost, Final, C2, SMOTE . . . 114
D.173 Gradient Boosting, Fold 1, C2, SMOTE . . . 114
D.174 Gradient Boosting, Fold 2, C2, SMOTE . . . 114
D.174 Gradient Boosting, Fold 3, C2, SMOTE . . . 114
D.175 Gradient Boosting, Fold 4, C2, SMOTE . . . 114
D.175 Gradient Boosting, Fold 5, C2, SMOTE . . . 115
D.176 Gradient Boosting, Final, C2, SMOTE . . . 115
D.177 SVC, Fold 1, C2, SMOTE . . . 115
D.178 SVC, Fold 2, C2, SMOTE . . . 115
D.178 SVC, Fold 3, C2, SMOTE . . . 115
D.179 SVC, Fold 4, C2, SMOTE . . . 115
D.179 SVC, Fold 5, C2, SMOTE . . . 116
D.180 SVC, Final, C2, SMOTE . . . 116
D.181 k-Nearest Neighbors, Fold 1, C2, SMOTE . . . 116
D.182 k-Nearest Neighbors, Fold 2, C2, SMOTE . . . 116
D.182 k-Nearest Neighbors, Fold 3, C2, SMOTE . . . 116
D.183 k-Nearest Neighbors, Fold 4, C2, SMOTE . . . 116
D.183 k-Nearest Neighbors, Fold 5, C2, SMOTE . . . 117
D.184 k-Nearest Neighbors, Final, C2, SMOTE . . . 117
D.185 Gaussian Naive Bayes, Fold 1, C2, SMOTE . . . 117
D.186 Gaussian Naive Bayes, Fold 2, C2, SMOTE . . . 117
D.186 Gaussian Naive Bayes, Fold 3, C2, SMOTE . . . 117
D.187 Gaussian Naive Bayes, Fold 4, C2, SMOTE . . . 117
D.187 Gaussian Naive Bayes, Fold 5, C2, SMOTE . . . 118
D.188 Gaussian Naive Bayes, Final, C2, SMOTE . . . 118
D.189 MLP, Fold 1, C2, SMOTE . . . 118
D.190 MLP, Fold 2, C2, SMOTE . . . 118
D.190 MLP, Fold 3, C2, SMOTE . . . 118
D.191 MLP, Fold 4, C2, SMOTE . . . 118
D.191 MLP, Fold 5, C2, SMOTE . . . 119
D.192 MLP, Final, C2, SMOTE . . . 119
D.193 Logistic Regression, Fold 1, C3, ADASYN . . . 119
D.194 Logistic Regression, Fold 2, C3, ADASYN . . . 119
D.194 Logistic Regression, Fold 3, C3, ADASYN . . . 119
D.195 Logistic Regression, Fold 4, C3, ADASYN . . . 119
D.195 Logistic Regression, Fold 5, C3, ADASYN . . . 120
D.196 Logistic Regression, Final, C3, ADASYN . . . 120
D.197 Logistic Regression Lasso, Fold 1, C3, ADASYN . . . 120
D.198 Logistic Regression Lasso, Fold 2, C3, ADASYN . . . 120
D.198 Logistic Regression Lasso, Fold 3, C3, ADASYN . . . 120
D.199 Logistic Regression Lasso, Fold 4, C3, ADASYN . . . 120
D.199 Logistic Regression Lasso, Fold 5, C3, ADASYN . . . 121

D.200 Logistic Regression Lasso, Final, C3, ADASYN . . . 121
D.201 Ridge, Fold 1, C3, ADASYN . . . 121
D.202 Ridge, Fold 2, C3, ADASYN . . . 121
D.202 Ridge, Fold 3, C3, ADASYN . . . 121
D.203 Ridge, Fold 4, C3, ADASYN . . . 121
D.203 Ridge, Fold 5, C3, ADASYN . . . 122
D.204 Ridge, Final, C3, ADASYN . . . 122
D.205 SDG, Fold 1, C3, ADASYN . . . 122
D.206 SDG, Fold 2, C3, ADASYN . . . 122
D.206 SDG, Fold 3, C3, ADASYN . . . 122
D.207 SDG, Fold 4, C3, ADASYN . . . 122
D.207 SDG, Fold 5, C3, ADASYN . . . 123
D.208 SDG, Final, C3, ADASYN . . . 123
D.209 Decision Tree, Fold 1, C3, ADASYN . . . 123
D.210 Decision Tree, Fold 2, C3, ADASYN . . . 123
D.210 Decision Tree, Fold 3, C3, ADASYN . . . 123
D.211 Decision Tree, Fold 4, C3, ADASYN . . . 123
D.211 Decision Tree, Fold 5, C3, ADASYN . . . 124
D.212 Decision Tree, Final, C3, ADASYN . . . 124
D.213 Random Forest, Fold 1, C3, ADASYN . . . 124
D.214 Random Forest, Fold 2, C3, ADASYN . . . 124
D.214 Random Forest, Fold 3, C3, ADASYN . . . 124
D.215 Random Forest, Fold 4, C3, ADASYN . . . 124
D.215 Random Forest, Fold 5, C3, ADASYN . . . 125
D.216 Random Forest, Final, C3, ADASYN . . . 125
D.217 AdaBoost, Fold 1, C3, ADASYN . . . 125
D.218 AdaBoost, Fold 2, C3, ADASYN . . . 125
D.218 AdaBoost, Fold 3, C3, ADASYN . . . 125
D.219 AdaBoost, Fold 4, C3, ADASYN . . . 125
D.219 AdaBoost, Fold 5, C3, ADASYN . . . 126
D.220 AdaBoost, Final, C3, ADASYN . . . 126
D.221 Gradient Boosting, Fold 1, C3, ADASYN . . . 126
D.222 Gradient Boosting, Fold 2, C3, ADASYN . . . 126
D.222 Gradient Boosting, Fold 3, C3, ADASYN . . . 126
D.223 Gradient Boosting, Fold 4, C3, ADASYN . . . 126
D.223 Gradient Boosting, Fold 5, C3, ADASYN . . . 127
D.224 Gradient Boosting, Final, C3, ADASYN . . . 127
D.225 SVC, Fold 1, C3, ADASYN . . . 127
D.226 SVC, Fold 2, C3, ADASYN . . . 127
D.226 SVC, Fold 3, C3, ADASYN . . . 127
D.227 SVC, Fold 4, C3, ADASYN . . . 127
D.227 SVC, Fold 5, C3, ADASYN . . . 128
D.228 SVC, Final, C3, ADASYN . . . 128
D.229 k-Nearest Neighbors, Fold 1, C3, ADASYN . . . 128
D.230 k-Nearest Neighbors, Fold 2, C3, ADASYN . . . 128
D.230 k-Nearest Neighbors, Fold 3, C3, ADASYN . . . 128
D.231 k-Nearest Neighbors, Fold 4, C3, ADASYN . . . 128
D.231 k-Nearest Neighbors, Fold 5, C3, ADASYN . . . 129
D.232 k-Nearest Neighbors, Final, C3, ADASYN . . . 129
D.233 Gaussian Naive Bayes, Fold 1, C3, ADASYN . . . 129
D.234 Gaussian Naive Bayes, Fold 2, C3, ADASYN . . . 129

D.234 Gaussian Naive Bayes, Fold 3, C3, ADASYN . . . 129
D.235 Gaussian Naive Bayes, Fold 4, C3, ADASYN . . . 129
D.235 Gaussian Naive Bayes, Fold 5, C3, ADASYN . . . 130
D.236 Gaussian Naive Bayes, Final, C3, ADASYN . . . 130
D.237 MLP, Fold 1, C3, ADASYN . . . 130
D.238 MLP, Fold 2, C3, ADASYN . . . 130
D.238 MLP, Fold 3, C3, ADASYN . . . 130
D.239 MLP, Fold 4, C3, ADASYN . . . 130
D.239 MLP, Fold 5, C3, ADASYN . . . 131
D.240 MLP, Final, C3, ADASYN . . . 131
D.241 Ridge, Fold 1, C3, SMOTE . . . 131
D.242 Ridge, Fold 2, C3, SMOTE . . . 131
D.242 Ridge, Fold 3, C3, SMOTE . . . 131
D.243 Ridge, Fold 4, C3, SMOTE . . . 131
D.243 Ridge, Fold 5, C3, SMOTE . . . 132
D.244 Ridge, Final, C3, SMOTE . . . 132
D.245 SDG, Fold 1, C3, SMOTE . . . 132
D.246 SDG, Fold 2, C3, SMOTE . . . 132
D.246 SDG, Fold 3, C3, SMOTE . . . 132
D.247 SDG, Fold 4, C3, SMOTE . . . 132
D.247 SDG, Fold 5, C3, SMOTE . . . 133
D.248 SDG, Final, C3, SMOTE . . . 133
D.249 Decision Tree, Fold 1, C3, SMOTE . . . 133
D.250 Decision Tree, Fold 2, C3, SMOTE . . . 133
D.250 Decision Tree, Fold 3, C3, SMOTE . . . 133
D.251 Decision Tree, Fold 4, C3, SMOTE . . . 133
D.251 Decision Tree, Fold 5, C3, SMOTE . . . 134
D.252 Decision Tree, Final, C3, SMOTE . . . 134
D.253 Random Forest, Fold 1, C3, SMOTE . . . 134
D.254 Random Forest, Fold 2, C3, SMOTE . . . 134
D.254 Random Forest, Fold 3, C3, SMOTE . . . 134
D.255 Random Forest, Fold 4, C3, SMOTE . . . 134
D.255 Random Forest, Fold 5, C3, SMOTE . . . 135
D.256 Random Forest, Final, C3, SMOTE . . . 135
D.257 AdaBoost, Fold 1, C3, SMOTE . . . 135
D.258 AdaBoost, Fold 2, C3, SMOTE . . . 135
D.258 AdaBoost, Fold 3, C3, SMOTE . . . 135
D.259 AdaBoost, Fold 4, C3, SMOTE . . . 135
D.259 AdaBoost, Fold 5, C3, SMOTE . . . 136
D.260 AdaBoost, Final, C3, SMOTE . . . 136
D.261 Gradient Boosting, Fold 1, C3, SMOTE . . . 136
D.262 Gradient Boosting, Fold 2, C3, SMOTE . . . 136
D.262 Gradient Boosting, Fold 3, C3, SMOTE . . . 136
D.263 Gradient Boosting, Fold 4, C3, SMOTE . . . 136
D.263 Gradient Boosting, Fold 5, C3, SMOTE . . . 137
D.264 Gradient Boosting, Final, C3, SMOTE . . . 137
D.265 SVC, Fold 1, C3, SMOTE . . . 137
D.266 SVC, Fold 2, C3, SMOTE . . . 137
D.266 SVC, Fold 3, C3, SMOTE . . . 137
D.267 SVC, Fold 4, C3, SMOTE . . . 137
D.267 SVC, Fold 5, C3, SMOTE . . . 138

D.268 SVC, Final, C3, SMOTE . . . 138
D.269 k-Nearest Neighbors, Fold 1, C3, SMOTE . . . 138
D.270 k-Nearest Neighbors, Fold 2, C3, SMOTE . . . 138
D.270 k-Nearest Neighbors, Fold 3, C3, SMOTE . . . 138
D.271 k-Nearest Neighbors, Fold 4, C3, SMOTE . . . 138
D.271 k-Nearest Neighbors, Fold 5, C3, SMOTE . . . 139
D.272 k-Nearest Neighbors, Final, C3, SMOTE . . . 139
D.273 Gaussian Naive Bayes, Fold 1, C3, SMOTE . . . 139
D.274 Gaussian Naive Bayes, Fold 2, C3, SMOTE . . . 139
D.274 Gaussian Naive Bayes, Fold 3, C3, SMOTE . . . 139
D.275 Gaussian Naive Bayes, Fold 4, C3, SMOTE . . . 139
D.275 Gaussian Naive Bayes, Fold 5, C3, SMOTE . . . 140
D.276 Gaussian Naive Bayes, Final, C3, SMOTE . . . 140
D.277 MLP, Fold 1, C3, SMOTE . . . 140
D.278 MLP, Fold 2, C3, SMOTE . . . 140
D.278 MLP, Fold 3, C3, SMOTE . . . 140
D.279 MLP, Fold 4, C3, SMOTE . . . 140
D.279 MLP, Fold 5, C3, SMOTE . . . 141
D.280 MLP, Final, C3, SMOTE . . . 141
D.281 Logistic Regression, Fold 1, C4, ADASYN . . . 141
D.282 Logistic Regression, Fold 2, C4, ADASYN . . . 141
D.282 Logistic Regression, Fold 3, C4, ADASYN . . . 141
D.283 Logistic Regression, Fold 4, C4, ADASYN . . . 141
D.283 Logistic Regression, Fold 5, C4, ADASYN . . . 142
D.284 Logistic Regression, Final, C4, ADASYN . . . 142
D.285 Logistic Regression Lasso, Fold 1, C4, ADASYN . . . 142
D.286 Logistic Regression Lasso, Fold 2, C4, ADASYN . . . 142
D.286 Logistic Regression Lasso, Fold 3, C4, ADASYN . . . 142
D.287 Logistic Regression Lasso, Fold 4, C4, ADASYN . . . 142
D.287 Logistic Regression Lasso, Fold 5, C4, ADASYN . . . 143
D.288 Logistic Regression Lasso, Final, C4, ADASYN . . . 143
D.289 Ridge, Fold 1, C4, ADASYN . . . 143
D.290 Ridge, Fold 2, C4, ADASYN . . . 143
D.290 Ridge, Fold 3, C4, ADASYN . . . 143
D.291 Ridge, Fold 4, C4, ADASYN . . . 143
D.291 Ridge, Fold 5, C4, ADASYN . . . 144
D.292 Ridge, Final, C4, ADASYN . . . 144
D.293 SDG, Fold 1, C4, ADASYN . . . 144
D.294 SDG, Fold 2, C4, ADASYN . . . 144
D.294 SDG, Fold 3, C4, ADASYN . . . 144
D.295 SDG, Fold 4, C4, ADASYN . . . 144
D.295 SDG, Fold 5, C4, ADASYN . . . 145
D.296 SDG, Final, C4, ADASYN . . . 145
D.297 Decision Tree, Fold 1, C4, ADASYN . . . 145
D.298 Decision Tree, Fold 2, C4, ADASYN . . . 145
D.298 Decision Tree, Fold 3, C4, ADASYN . . . 145
D.299 Decision Tree, Fold 4, C4, ADASYN . . . 145
D.299 Decision Tree, Fold 5, C4, ADASYN . . . 146
D.300 Decision Tree, Final, C4, ADASYN . . . 146
D.301 Random Forest, Fold 1, C4, ADASYN . . . 146
D.302 Random Forest, Fold 2, C4, ADASYN . . . 146

D.302 Random Forest, Fold 3, C4, ADASYN . . . 146
D.303 Random Forest, Fold 4, C4, ADASYN . . . 146
D.303 Random Forest, Fold 5, C4, ADASYN . . . 147
D.304 Random Forest, Final, C4, ADASYN . . . 147
D.305 AdaBoost, Fold 1, C4, ADASYN . . . 147
D.306 AdaBoost, Fold 2, C4, ADASYN . . . 147
D.306 AdaBoost, Fold 3, C4, ADASYN . . . 147
D.307 AdaBoost, Fold 4, C4, ADASYN . . . 147
D.307 AdaBoost, Fold 5, C4, ADASYN . . . 148
D.308 AdaBoost, Final, C4, ADASYN . . . 148
D.309 Gradient Boosting, Fold 1, C4, ADASYN . . . 148
D.310 Gradient Boosting, Fold 2, C4, ADASYN . . . 148
D.310 Gradient Boosting, Fold 3, C4, ADASYN . . . 148
D.311 Gradient Boosting, Fold 4, C4, ADASYN . . . 148
D.311 Gradient Boosting, Fold 5, C4, ADASYN . . . 149
D.312 Gradient Boosting, Final, C4, ADASYN . . . 149
D.313 SVC, Fold 1, C4, ADASYN . . . 149
D.314 SVC, Fold 2, C4, ADASYN . . . 149
D.314 SVC, Fold 3, C4, ADASYN . . . 149
D.315 SVC, Fold 4, C4, ADASYN . . . 149
D.315 SVC, Fold 5, C4, ADASYN . . . 150
D.316 SVC, Final, C4, ADASYN . . . 150
D.317 k-Nearest Neighbors, Fold 1, C4, ADASYN . . . 150
D.318 k-Nearest Neighbors, Fold 2, C4, ADASYN . . . 150
D.318 k-Nearest Neighbors, Fold 3, C4, ADASYN . . . 150
D.319 k-Nearest Neighbors, Fold 4, C4, ADASYN . . . 150
D.319 k-Nearest Neighbors, Fold 5, C4, ADASYN . . . 151
D.320 k-Nearest Neighbors, Final, C4, ADASYN . . . 151
D.321 Gaussian Naive Bayes, Fold 1, C4, ADASYN . . . 151
D.322 Gaussian Naive Bayes, Fold 2, C4, ADASYN . . . 151
D.322 Gaussian Naive Bayes, Fold 3, C4, ADASYN . . . 151
D.323 Gaussian Naive Bayes, Fold 4, C4, ADASYN . . . 151
D.323 Gaussian Naive Bayes, Fold 5, C4, ADASYN . . . 152
D.324 Gaussian Naive Bayes, Final, C4, ADASYN . . . 152
D.325 MLP, Fold 1, C4, ADASYN . . . 152
D.326 MLP, Fold 2, C4, ADASYN . . . 152
D.326 MLP, Fold 3, C4, ADASYN . . . 152
D.327 MLP, Fold 4, C4, ADASYN . . . 152
D.327 MLP, Fold 5, C4, ADASYN . . . 153
D.328 MLP, Final, C4, ADASYN . . . 153
D.329 Ridge, Fold 1, C4, SMOTE . . . 153
D.330 Ridge, Fold 2, C4, SMOTE . . . 153
D.330 Ridge, Fold 3, C4, SMOTE . . . 153
D.331 Ridge, Fold 4, C4, SMOTE . . . 153
D.331 Ridge, Fold 5, C4, SMOTE . . . 154
D.332 Ridge, Final, C4, SMOTE . . . 154
D.333 SDG, Fold 1, C4, SMOTE . . . 154
D.334 SDG, Fold 2, C4, SMOTE . . . 154
D.334 SDG, Fold 3, C4, SMOTE . . . 154
D.335 SDG, Fold 4, C4, SMOTE . . . 154
D.335 SDG, Fold 5, C4, SMOTE . . . 155

D.336 SDG, Final, C4, SMOTE . . . 155
D.337 Decision Tree, Fold 1, C4, SMOTE . . . 155
D.338 Decision Tree, Fold 2, C4, SMOTE . . . 155
D.338 Decision Tree, Fold 3, C4, SMOTE . . . 155
D.339 Decision Tree, Fold 4, C4, SMOTE . . . 155
D.339 Decision Tree, Fold 5, C4, SMOTE . . . 156
D.340 Decision Tree, Final, C4, SMOTE . . . 156
D.341 Random Forest, Fold 1, C4, SMOTE . . . 156
D.342 Random Forest, Fold 2, C4, SMOTE . . . 156
D.342 Random Forest, Fold 3, C4, SMOTE . . . 156
D.343 Random Forest, Fold 4, C4, SMOTE . . . 156
D.343 Random Forest, Fold 5, C4, SMOTE . . . 157
D.344 Random Forest, Final, C4, SMOTE . . . 157
D.345 AdaBoost, Fold 1, C4, SMOTE . . . 157
D.346 AdaBoost, Fold 2, C4, SMOTE . . . 157
D.346 AdaBoost, Fold 3, C4, SMOTE . . . 157
D.347 AdaBoost, Fold 4, C4, SMOTE . . . 157
D.347 AdaBoost, Fold 5, C4, SMOTE . . . 158
D.348 AdaBoost, Final, C4, SMOTE . . . 158
D.349 Gradient Boosting, Fold 1, C4, SMOTE . . . 158
D.350 Gradient Boosting, Fold 2, C4, SMOTE . . . 158
D.350 Gradient Boosting, Fold 3, C4, SMOTE . . . 158
D.351 Gradient Boosting, Fold 4, C4, SMOTE . . . 158
D.351 Gradient Boosting, Fold 5, C4, SMOTE . . . 159
D.352 Gradient Boosting, Final, C4, SMOTE . . . 159
D.353 SVC, Fold 1, C4, SMOTE . . . 159
D.354 SVC, Fold 2, C4, SMOTE . . . 159
D.354 SVC, Fold 3, C4, SMOTE . . . 159
D.355 SVC, Fold 4, C4, SMOTE . . . 159
D.355 SVC, Fold 5, C4, SMOTE . . . 160
D.356 SVC, Final, C4, SMOTE . . . 160
D.357 k-Nearest Neighbors, Fold 1, C4, SMOTE . . . 160
D.358 k-Nearest Neighbors, Fold 2, C4, SMOTE . . . 160
D.358 k-Nearest Neighbors, Fold 3, C4, SMOTE . . . 160
D.359 k-Nearest Neighbors, Fold 4, C4, SMOTE . . . 160
D.359 k-Nearest Neighbors, Fold 5, C4, SMOTE . . . 161
D.360 k-Nearest Neighbors, Final, C4, SMOTE . . . 161
D.361 Gaussian Naive Bayes, Fold 1, C4, SMOTE . . . 161
D.362 Gaussian Naive Bayes, Fold 2, C4, SMOTE . . . 161
D.362 Gaussian Naive Bayes, Fold 3, C4, SMOTE . . . 161
D.363 Gaussian Naive Bayes, Fold 4, C4, SMOTE . . . 161
D.363 Gaussian Naive Bayes, Fold 5, C4, SMOTE . . . 162
D.364 Gaussian Naive Bayes, Final, C4, SMOTE . . . 162
D.365 MLP, Fold 1, C4, SMOTE . . . 162
D.366 MLP, Fold 2, C4, SMOTE . . . 162
D.366 MLP, Fold 3, C4, SMOTE . . . 162
D.367 MLP, Fold 4, C4, SMOTE . . . 162
D.367 MLP, Fold 5, C4, SMOTE . . . 163
D.368 MLP, Final, C4, SMOTE . . . 163

Índice de tablas

3.1 Información variables clínicas. 12

4.1 Clasificadores 18

5.1 Clasificación 1: ADASYN 26
5.2 Clasificación 1: SMOTE 30
5.3 Clasificación 2: ADASYN 34
5.4 Clasificación 2: SMOTE 36
5.5 Clasificación 3: ADASYN 39
5.6 Clasificación 3: SMOTE 42
5.7 Clasificación 4: ADASYN 44
5.8 Clasificación 4: SMOTE 46

C.1 Estadísticas de la edad 64
C.2 Estadísticas del PSA 65
C.3 Frecuencia de lesiones para gleason1.1 y gleason2.1 66

CAPÍTULO 1
Introducción

No es ninguna novedad que la tecnología está cada vez más presente en nuestras vidas. Desde la inteligencia artificial hasta el *machine learning*, se ha transformado la forma en que interactuamos con el mundo que nos rodea. El *machine learning*, o aprendizaje automático, destaca como una rama de la inteligencia artificial que permite a las máquinas adquirir conocimientos a partir de datos, mejorando su rendimiento mediante la experiencia, sin necesidad de ser programadas explícitamente. Este avance ha propiciado mejoras significativas en una variedad de campos, desde el procesamiento de lenguaje natural hasta la visión por computadora, así como la toma de decisiones automatizada.

A día de hoy, el *machine learning* se está aplicando en todos los sectores que nos podamos imaginar. Sin embargo, uno en el que destaca de forma revolucionaria es en el campo de la salud y la atención médica.

> "El sector sanitario lleva años trabajando para sacar el máximo provecho del gran *machine learning*. Y es que este avance tecnológico juega un papel determinante en el desarrollo de procedimientos médicos o la gestión de datos. Un sector que se está transformando a pasos agigantados gracias a un aprendizaje automático que, entre otras cosas, permite predecir, tratar y diagnosticar enfermedades a los pacientes." (Etkho, 2021a)

Esta transformación tecnológica ha permitido identificar patrones y tendencias en datos médicos que antes pasaban totalmente desapercibidos. Todo ello va abriendo puertas hacia la medicina de precisión, que se centra en la salud de cada paciente como un individual, adaptando todo tratamiento a sus necesidades y su caso. Su objetivo, al igual que el de muchos avances médicos, es mejorar la calidad de vida de los pacientes a la vez que se facilita la labor del personal médico (Etkho, 2021b). En este contexto, la radiómica ha emergido como una disciplina fascinante.

La radiómica surge como una herramienta fundamental para interpretar las complejidades visuales de una imagen médica. Esta ciencia se encarga de traducir cada imagen en un gran número de características, proporcionando una riqueza de información que va más allá de la representación visual superficial. Así pues, si ya de normal 'una imagen vale más que mil palabras', en este campo su valor se potencia significativamente gracias a la interpretación detallada. (Sierra, s.f.)

Es cierto que una imagen aporta confianza tanto a pacientes como a especialistas. Sin embargo, una imagen por sí sola no es más que eso, una imagen. Es por esto que es muy importante la interpretación que se haga de ella.

Así pues, este estudio se centra en utilizar características radiómicas y variables clínicas de distintos pacientes para construir un modelo de clasificación. El objetivo es predecir la gravedad del cáncer de próstata, fusionando la información visual con datos clínicos para avanzar en la precisión diagnóstica y mejorar la toma de decisiones médicas.

1.1 Motivación

En los últimos años, el análisis de datos se ha vuelto una herramienta esencial en diversas áreas de nuestra vida. Los datos son omnipresentes y utilizarlos para la toma de decisiones se ha convertido en una práctica esencial. En la actualidad, organizaciones de distintos sectores confían en el análisis de datos como un recurso poderoso y eficaz. A través de esta disciplina, es posible identificar tanto las fortalezas como las debilidades en un campo determinado y emprender acciones para mejorar. Este trabajo tiene como objetivo contribuir a esas mejoras, enfocándose especialmente en el campo de la salud.

En el ámbito de la salud, la utilidad del análisis de datos se extiende a la mejora de la calidad de la atención médica y, por ende, a la vida de los pacientes. La capacidad para extraer información significativa a partir de conjuntos de datos complejos ha propiciado avances notables en el diagnóstico y tratamiento de diversas enfermedades. Sin embargo, aún persisten desafíos significativos, como en el ámbito de la oncología, donde la precisión en la evaluación de la gravedad de las enfermedades sigue siendo una tarea compleja.

La importancia de avanzar en la investigación del cáncer es innegable. La mejora en la detección temprana y tratamiento de esta enfermedad es un objetivo primordial, y el análisis de datos puede desempeñar un papel crítico en esta tarea.

El impacto potencial de este trabajo se traduce en una mejora significativa en la capacidad de los profesionales médicos para realizar evaluaciones más precisas y objetivas. Al centrarnos en el análisis radiómico, buscamos aprovechar la riqueza de información presente en las imágenes médicas para desarrollar un modelo de clasificación que complemente y mejore las prácticas actuales. Este enfoque tiene el potencial no solo de proporcionar diagnósticos más tempranos y precisos, sino también de allanar el camino hacia tratamientos personalizados, permitiendo mejorar la calidad de vida de los pacientes afectados por el cáncer de próstata.

En conclusión, al contribuir a la convergencia entre la medicina y las tecnologías de análisis de datos, este trabajo aspira a ser un paso adelante en esta revolución médica. Apoyar el progreso de esta disciplina resulta esencial para impulsar la toma de decisiones clínicas informadas, potenciar la investigación médica y, en última instancia, elevar la calidad de la atención médica.

1.2 Objetivos

Conforme a lo dicho anteriormente, definimos un objetivo general seguido de otros más específicos.

El objetivo general de este Trabajo de Fin de Grado es desarrollar un modelo de clasificación mediante técnicas estadísticas multivariantes y de aprendizaje automático para predecir la variable ISUP, que refleja la gravedad del cáncer de próstata a partir de imágenes médicas. Este enfoque busca mejorar la precisión en el diagnóstico y clasificación de la enfermedad, contribuyendo así al avance de la medicina de precisión y la toma de decisiones clínicas informadas.

Para poder conseguirlo, definimos otros objetivos específicos más concretos, factibles y alcanzables:

- Realizar una revisión exhaustiva de la literatura sobre técnicas estadísticas multivariantes y aprendizaje automático aplicadas a la radiómica del cáncer de próstata.
- Explorar y comprender la estructura y las características del conjunto de datos existente, identificando variables médicas clave y su relación con la variable objetivo (ISUP).
- Desarrollar y entrenar modelos de aprendizaje automático utilizando el conjunto de datos existente, con el objetivo de predecir con precisión los niveles de ISUP.
- Evaluar la efectividad y la robustez del modelo propuesto mediante métricas de rendimiento y validación cruzada, utilizando el conjunto de datos proporcionado.
- Interpretar los resultados obtenidos, destacando las características clave identificadas por el modelo y su relevancia clínica en el contexto del cáncer de próstata.
- Elaborar conclusiones y recomendaciones basadas en los hallazgos, con énfasis en su aplicabilidad para el personal médico y los investigadores involucrados en la medicina de precisión.

1.3 Impacto esperado

El presente trabajo de fin de Grado en Ciencia de Datos, que se enfoca en la clasificación del grado de severidad del cáncer de próstata, busca tener un impacto significativo en varios aspectos clave:

1.3.1. Avance en la Medicina de Precisión

La implementación de técnicas de radiómica en datos médicos busca impulsar la medicina de precisión. Este enfoque permitirá personalizar los tratamientos y diagnósticos de los pacientes, mejorando la calidad de vida de los mismos y optimizando la toma de decisiones clínicas. La medicina de precisión es el futuro de la atención médica, y este trabajo representa un paso crucial en esa dirección.

1.3.2. Optimización de Recursos

Un modelo de clasificación preciso puede contribuir a la optimización de recursos al permitir una asignación más eficiente de tratamientos y seguimientos. Al identificar con mayor preci-

sión la gravedad del cáncer, se pueden priorizar los casos más urgentes, reduciendo tiempos de espera y maximizando la eficacia de los recursos médicos disponibles.

1.3.3. Mejora en la Eficiencia Diagnóstica

La aplicación de técnicas avanzadas de análisis de datos busca mejorar la eficiencia diagnóstica al proporcionar resultados más rápidos y precisos. Un modelo de clasificación robusto puede agilizar el proceso de diagnóstico, permitiendo intervenciones médicas más tempranas y aumentando las posibilidades de éxito en el tratamiento.

1.3.4. Empoderamiento del paciente

Al mejorar la precisión en la evaluación del cáncer de próstata, este trabajo tiene el potencial de empoderar a los pacientes al proporcionar información más clara sobre su condición. Una comprensión más completa y precisa puede ayudar a los pacientes a tomar decisiones informadas sobre su tratamiento y participar activamente en su atención médica.

1.3.5. Contribución a la Investigación del Cáncer de Próstata

El cáncer de próstata es una enfermedad devastadora, y la detección temprana y el tratamiento son esenciales para mejorar la supervivencia de los pacientes. Este trabajo tiene el potencial de mejorar la precisión en la detección y el monitoreo de esta enfermedad, lo que puede marcar una diferencia significativa en la lucha contra el cáncer de próstata. Además, los hallazgos derivados de este trabajo pueden contribuir a la literatura científica, proporcionando información relevante para futuras investigaciones y avances en la comprensión y tratamiento del cáncer de próstata. Cada avance en esta área es una contribución valiosa a la investigación médica.

Por todo ello, este trabajo de fin de grado busca marcar una diferencia real en la aplicación de la ciencia de datos en el ámbito de la salud, con un enfoque particular en la radiómica y el cáncer de próstata. Se espera que sus resultados tengan un impacto duradero en la medicina de precisión, la investigación médica y la calidad de la atención médica, contribuyendo así al avance de la ciencia y la mejora de la salud de las personas.

1.4 Metodología

Definimos una serie pasos que nos permitan alcanzar los objetivos propuestos. A continuación se muestra el diagrama de flujo de la metodología:

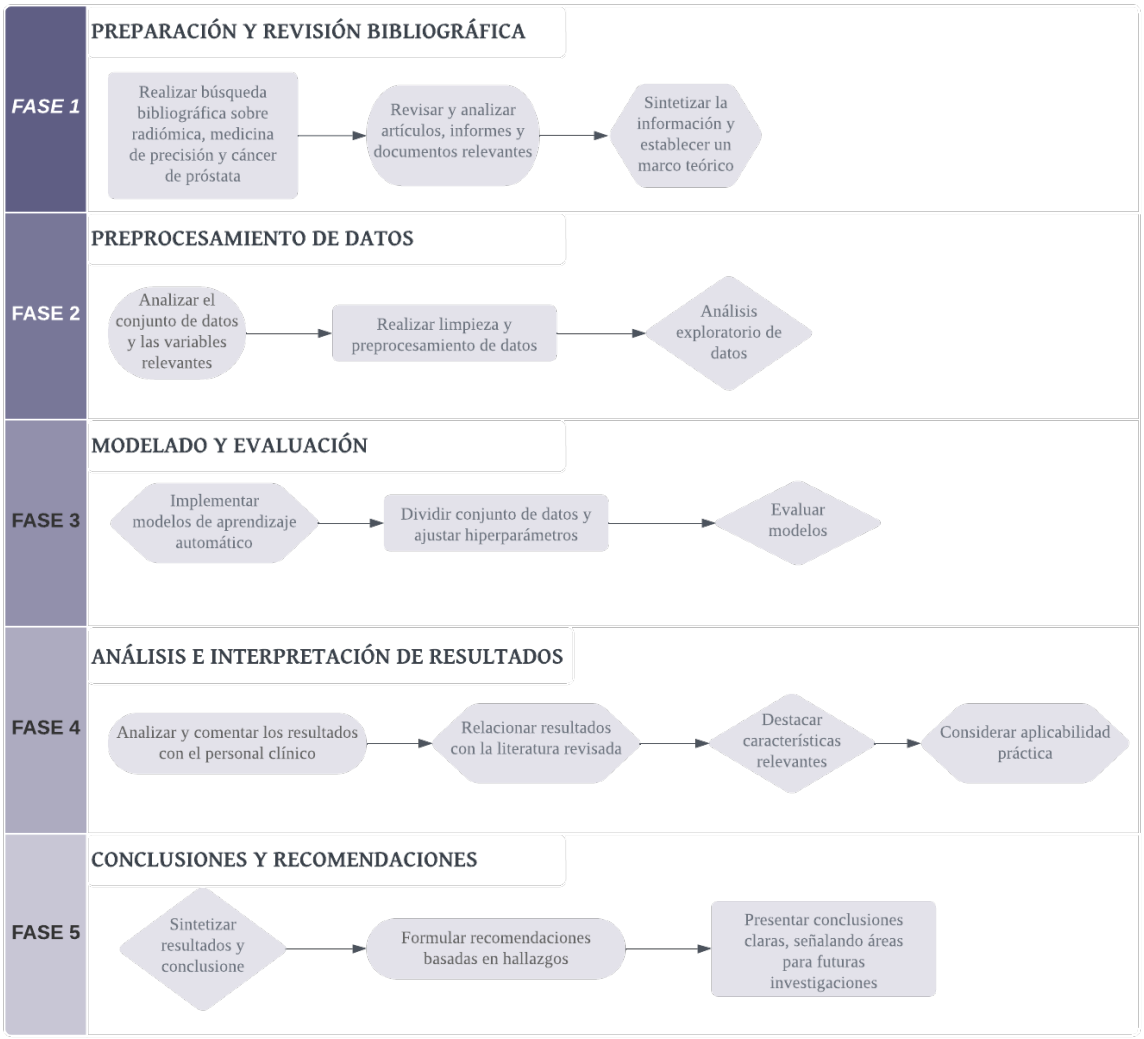

Figura 1.1: Metodología en forma de diagrama de flujo.

1.5 Estructura del TFG

Este Trabajo de Fin de Grado (TFG) se organiza en seis capítulos esenciales, cada uno contribuyendo a la comprensión integral de la investigación llevada a cabo:

En el capítulo inicial, se exponen las motivaciones, el impacto y los objetivos que guían la investigación. Se establece una visión general de la relevancia del estudio y se delimita el alcance del trabajo.

El segundo capítulo contextualiza la temática del cáncer de próstata en relación con las técnicas estadísticas multivariantes y el aprendizaje automático. Se definen los conceptos clave y se presenta un análisis detallado de la literatura existente, proporcionando una base sólida para la investigación.

El tercer capítulo se centra en la descripción completa de la base de datos utilizada en el estudio. Se detallan los procesos de adquisición de datos, se exploran las características fundamen-

tales y se explica el procedimiento de limpieza y preprocesamiento de datos aplicado.

En el cuarto capítulo, se detallan los enfoques y técnicas de evaluación aplicados durante el desarrollo y entrenamiento de los modelos de aprendizaje automático.

El quinto capítulo presenta y analiza de manera detallada los resultados obtenidos a través de la implementación de los modelos de aprendizaje automático. Se acompañan los hallazgos con gráficos, tablas y visualizaciones pertinentes para facilitar la comprensión.

El capítulo final sintetiza los objetivos alcanzados, los resultados obtenidos y las implicaciones clave de la investigación. Se discuten las limitaciones del estudio y se proporcionan recomendaciones para futuras investigaciones, destacando la aplicabilidad de los hallazgos para el personal médico y los investigadores en el ámbito de la medicina de precisión.

CAPÍTULO 2
Antecedentes

2.1 Radiómica y su Relevancia en la Medicina

La radiómica es una disciplina en auge que, por medio de algoritmos computacionales, se encarga de extraer y analizar una amplia gama de características cuantitativas de las imágenes médicas, transformando las imágenes en datos numéricos. Estas características, imperceptibles al ojo humano, poseen información muy valiosa para la detección, diagnóstico, y tratamiento de diversas enfermedades (Soteras, 2022). Por tanto, el objetivo fundamental de esta ciencia es establecer conexiones sólidas entre estas características y estados fisiológicos específicos.

La radiómica ha llegado para revolucionar el sector médico, tanto en el ámbito asistencial como en el de la investigación. Aunque su utilidad y aplicaciones siguen siendo objeto de estudio, existen investigaciones que ya han demostrado su relevancia en ensayos clínicos (Higgins, 2020) (Fundación Instituto Roche, 2022). Como ya se ha mencionado previamente, la necesidad de una medicina personalizada que se ajuste a las necesidades de cada paciente es cada vez mayor. En este contexto, la radiómica desempeña un papel fundamental al proporcionar información precisa y detallada. Esto ha supuesto un avance significativo en el estudio y tratamiento de enfermedades muy comunes y a la vez muy desconocidas, como puede ser el cáncer (IM Médico, 2023).

2.2 Machine Learning en Radiómica

Sin embargo, para dar sentido a esta gran cantidad de datos y aprovechar su potencial, es esencial contar con un aliado: el *machine learning*. La radiómica, por una parte, se especializa en la extracción meticulosa de innumerables características cuantitativas a partir de imágenes médicas, como textura, forma, intensidad, y muchas otras. Estas características, en su mayoría invisibles para el ojo humano, contienen información valiosa. Por otra parte, el *machine learning* es quien asume el papel de analizar estos datos detallados y encontrar patrones y relaciones significativas. Ambos campos trabajan en conjunto para aprovechar al máximo la información contenida en las imágenes médicas y "mejorar nuestras vidas y la salud del planeta" (Etkho, 2021b). Para entenderlo mejor, podríamos decir que la radiómica proporciona la materia prima, mientras que el *machine learning* la procesa para desvelar conocimientos clínicamente relevantes.

Esta colaboración entre la radiómica y el *machine learning* ha sido fundamental en la evolución del análisis de imágenes médicas en los últimos años. Son varios los estudios que, ilustran cómo estas dos disciplinas se han combinado para extraer información valiosa de imágenes médicas y mejorar la toma de decisiones clínicas. A continuación, se habla de algunos de ellos.

En primer lugar, (García Galindo, 2021) destaca la utilidad de la radiómica en el diagnóstico y pronóstico de pacientes con glioblastoma multiforme (GBM), el tumor cerebral primario más común y agresivo. La aplicación de técnicas de radiómica a las imágenes de resonancia magnética (RM) prequirúrgicas permitió identificar características de textura que se correlacionaron con la progresión tumoral y la supervivencia. Esto subraya la capacidad de la radiómica para extraer información pronóstica de las imágenes médicas, lo que podría ayudar a mejorar la atención a los pacientes con GBM.

Por otro lado, el artículo (Salhab Ibáñez, 2021), se centró en el uso de la radiómica en el contexto de los tumores pancreáticos. En este estudio, se destacó cómo la combinación de características radiómicas y clínicas condujo a un modelo de predicción de supervivencia más eficiente en comparación con el uso exclusivo de características clínicas. Esto destaca la sinergia entre la información visual extraída de las imágenes médicas y los datos clínicos en el proceso de toma de decisiones médicas.

Otro ejemplo es el artículo (Martí Bonmatí y cols., 2012), donde se resalta la importancia de los biomarcadores de imagen, que son características identificadas en imágenes médicas relacionadas con enfermedades o tratamientos. Detalla los pasos necesarios para su desarrollo y validación, desde la concepción inicial hasta su aplicación en hospitales. Destaca la necesidad crucial de adquirir imágenes de alta calidad y de analizarlas utilizando técnicas avanzadas de aprendizaje automático para obtener información precisa y relevante para el diagnóstico y tratamiento médico.

Por último, en (González Vilanova, 2019) resalta la importancia del *machine learning* en el análisis de datos de imágenes médicas, especialmente en radiómica. Exploró tres modelos de aprendizaje automático: k-vecinos más cercanos, máquinas de soporte vectorial y bosques aleatorios en datos de radiómica e imágenes morfológicas. Estos métodos subrayan la relevancia de datos de alta calidad y cantidad en modelos predictivos y su relación con resultados clínicos, respaldando el potencial de estas técnicas en la medicina personalizada.

Es por esto que el uso combinado de la radiómica y el *machine learning* en la interpretación de imágenes médicas se ha convertido en una herramienta esencial para los profesionales de la salud. Esta sinergia entre la extracción de características cuantitativas y el análisis de datos ha demostrado su eficacia en diversos contextos clínicos, lo que acentúa la importancia de estas tecnologías en la toma de decisiones médicas y la mejora de la atención al paciente.

2.3 Métodos de Clasificación en Cáncer de Próstata mediante Radiómica

A pesar de sus raíces ancestrales, el cáncer persiste como un tema que despierta inquietudes y esfuerzos continuos en la sociedad, dado su impacto significativo en la salud global y la calidad de vida de las personas. Además, a medida que avanzamos en el siglo XXI, esta preocupación se intensifica, evidenciada por el aumento constante de nuevos casos. En 2016, se registraron

262.343 nuevos casos, y para 2022, esta cifra ascendió a 290.175, subrayando la creciente urgencia de abordar y comprender esta enfermedad que continúa desafiando a la medicina y la investigación médica (Asociación Española Contra el Cáncer, 2023).

Pero, entre todos los tipos, el cáncer de próstata sigue siendo la neoplasias más común entre los hombres, y representa un gran desafío en el ámbito de la oncología. En el año 2022, se diagnosticaron 35.862 casos clínicos de cáncer de próstata (Asociación Española Contra el Cáncer, 2023).

Esto destaca la relevancia crucial de este estudio, que se centra en la detección temprana e investigación de la enfermedad, contribuyendo así a reducir la mortalidad y mejorar la calidad de vida de los pacientes (Asociación Nacional de Cáncer de Próstata, s.f.).

Adicionalmente, (Jaén Lorites, 2019) utiliza técnicas avanzadas de clasificación y resonancia magnética para mejorar la precisión en la clasificación de lesiones de próstata.

En resumen, el continuo aumento de casos de cáncer, en particular de próstata, subraya la urgencia de estrategias efectivas. Este estudio, enfocado en la detección temprana, marca un avance significativo en el campo de la medicina.

2.4 Propuesta

El actual marco contextual crea una oportunidad de investigación única en la intersección de la radiómica y el *machine learning*. La capacidad de la radiómica para extraer características cuantitativas de imágenes médicas, combinada con la destreza del *machine learning* en el análisis de grandes conjuntos de datos, ha impulsado avances notables en la detección y clasificación de enfermedades.

Este proyecto se centra en aplicar métodos de clasificación en cáncer de próstata mediante la radiómica y el *machine learning*. El cáncer de próstata sigue siendo un desafío importante con un aumento constante de casos. Por ello, esta investigación busca mejorar la precisión en la detección temprana de la enfermedad, aprovechando la sinergia entre la radiómica y el *machine learning*. Se sitúa en la vanguardia de la lucha contra esta enfermedad, aprovechando las oportunidades proporcionadas por el actual marco contextual.

CAPÍTULO 3
Base de Datos

El conjunto de datos fue proporcionado por el Hospital La Fe de Valencia. Este conjunto parte de un proyecto anterior que se explica más detalladamente a continuación.

3.1 Obtención de los datos

En primer lugar, los datos se recopilaron a partir de exámenes de resonancia magnética T2W, DWI y ADC, provenientes del archivo de imágenes creado en el proyecto ProCancer-I.

ProCancer-I es un proyecto europeo en el que colaboran varios centros clínicos. Su objetivo es entrenar diferentes modelos de inteligencia artificial para la predicción de distintos resultados clínicos en cáncer de próstata.

Así pues, esas imágenes fueron adquiridas en las etapas iniciales del diagnóstico de la enfermedad y provienen de 13 socios clínicos que utilizaron 4 fabricantes de escáneres y 27 modelos de escáneres distintos.

Para ello, contaron con la aprobación de un comité ético y el consentimiento de los pacientes por parte de cada socio involucrado en la recopilación.

Todo esto conforma un conjunto de datos representativo que considera la variabilidad inherente en las imágenes médicas del cáncer de próstata. Este se dividió en variables radiómicas y variables clínicas.

Para entender mejor cómo se hizo la extracción de datos se puede revisar el Apéndice B.

3.2 Variables radiómicas

Por una parte, contamos con un conjunto de datos que contiene las variables radiómicas. Este consta de 323 filas y 111 columnas. De todas las columnas, 4 son identificadores, y, por ello, son de tipo objeto. Las otras 107 son todo variables cuantitativas extraídas de los datos radiómicos de las imágenes.

En este caso, el gran número de variables dificulta su visualización.

3.3 Variables clínicas

Este segundo conjunto de datos está formado por 323 filas y 19 columnas que describen información clínica sobre los pacientes.

La siguiente tabla muestra el tipo de cada variable y qué representa.

Variable	Descripción	Tipo
patient_id	identificador del paciente	objeto
age	edad	entero
psas_0_total	prueba de sangre que mide el nivel de PSA total del baseline	decimal
lesions_0_pi_rads	probabilidad de que una lesión sea cancerígena	entero
lesions_0_location_0	ubicación de las lesiones en la próstata	objeto
lesions_0_location_1	ubicación de las lesiones en la próstata	objeto
lesions_0_location_2	ubicación de las lesiones en la próstata	objeto
lesions_0_location_3	ubicación de las lesiones en la próstata	objeto
lesions_0_location_4	ubicación de las lesiones en la próstata	objeto
lesions_0_location_5	ubicación de las lesiones en la próstata	objeto
lesions_0_location_6	ubicación de las lesiones en la próstata	objeto
lesions_0_location_7	ubicación de las lesiones en la próstata	objeto
lesions_0_location_gleason1.1	puntuación que se asigna después de examinar las muestras de tejido en una biopsia. Evalua la agresividad del cáncer	entero
lesions_0_location_gleason2.1	puntuación que se asigna después de examinar las muestras de tejido en una biopsia. Evalua la agresividad del cáncer	entero
ISUP	clasifica el grado de severidad del cáncer. Está correlacionado con el Gleason	entero
TZ	indica si hay infiltración en la zona de transición de la próstata	bool
PZ	indica si hay infiltración en la zona periférica de la próstata	bool
CZ	indica si hay infiltración en la zona central de la próstata	bool
AS	indica si hay infiltración en la zona anterior de la próstata	bool

Tabla 3.1: Información variables clínicas.

Sin embargo, dado que las variables de tipo lesions_0_location_x no las vamos a usar en este estudio, las eliminamos.

De igual forma, también vamos a prescindir de las variables de tipo lesions_0_location_gleasonx. El objetivo del estudio es predecir la variable ISUP, que se obtiene a partir de la unión de los dos valores Gleason. Por ello, eliminar estas variables del *dataset* puede ser una decisión estratégica para mejorar la calidad y la interpretación del modelo de clasificación, así como para evitar problemas potenciales como sobreajuste.

En este caso, los datos permiten estudiar cada variable individualmente. Este análisis más profundo se puede ver en el Apéndice C.

CAPÍTULO 4
Evaluación

4.1 Preprocesamiento de datos

En primer lugar, se combinaron ambas bases de datos; la radiómica y la clínica. De esta manera, se integraron las 107 variables radiómicas con las 8 variables clínicas que quedaron después del proceso de depuración. A continuación, se realizó el preprocesado necesario para garantizar la calidad y adecuación de los datos.

Sin embargo, antes de explicar en detalle este proceso, cabe destacar que, inicialmente, se intentó clasificar utilizando la variable ISUP original, con sus 5 categorías. No obstante, los resultados obtenidos fueron significativamente pobres, lo que condujo a una reconsideración de la estrategia de clasificación. Se optó por explorar distintas recategorizaciones de la variable ISUP con el objetivo de mejorar el rendimiento de los modelos de clasificación. Estos enfoques son:

- 1,2 vs 3,4,5
- 1 vs 2,3 vs 4,5
- 1 vs 2 vs 3,4,5
- 1,2,3 vs 4,5

Cabe destacar que estas agrupaciones no son aleatorias, sino que se han elegido en diálogo con los expertos en cácer de próstata. Así pues, están justificadas clínicamente.

En cada una de estas recalibraciones de la variable ISUP, se aplicaron los siguientes pasos de preprocesamiento de datos:

4.1.1. Escalado

Se utilizó la técnica de escalado estándar (StandardScaler) para estandarizar las características numéricas y evitar que alguna variable tenga un peso desproporcionado en el modelo. Este paso es crucial para muchos algoritmos, ya que algunos pueden ser sensibles a la escala de las características.

4.1.2. División del conjunto de datos

El conjunto de datos se dividió en entrenamiento y prueba. Esto es fundamental para evaluar, ajustar y validar los modelos de forma efectiva; ayuda a evitar el sobreajuste, garantizando

que los modelos funcionarán para nuevos conjuntos de datos.

Para ello, se utilizó la técnica train_test_split de scikit-learn. Se asignó un tamaño de prueba del 20 %. También se estratificó para mantener la proporción de clases en ambos conjuntos.

4.1.3. Remuestreo

Finalmente, se observó un desbalanceo de clases que era necesario corregir. Así se puede ver en la siguiente imagen.

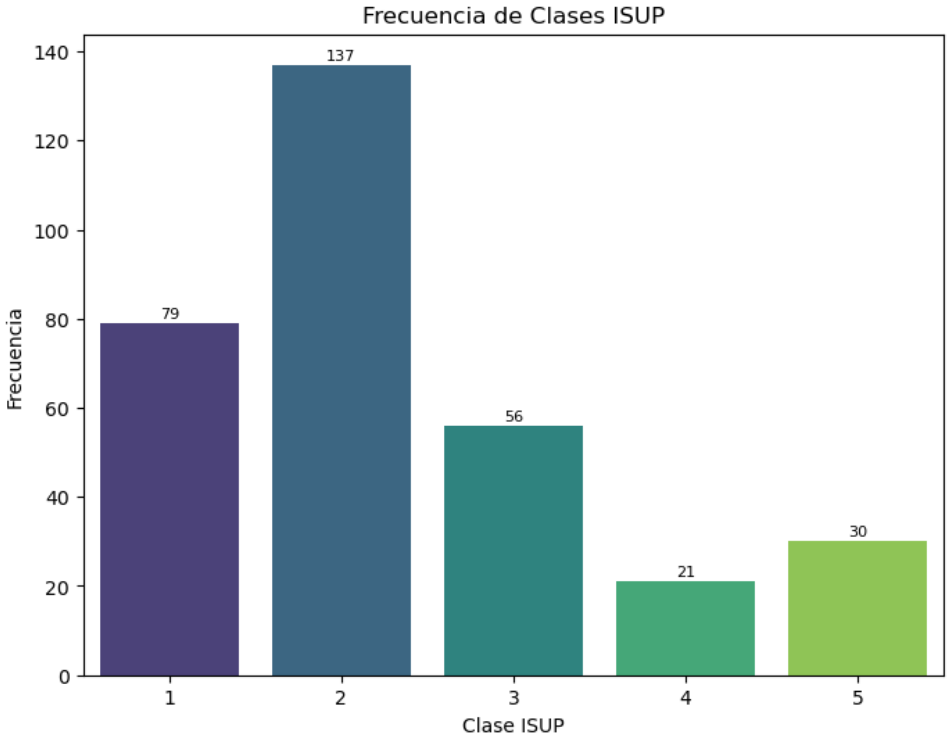

Figura 4.1: Desbalanceo de clases

Así pues, se aplicaron técnicas de remuestreo para abordar este problema. Se emplearon las técnicas de ADASYN (*Adaptive Synthetic Sampling*) y SMOTE (*Synthetic Minority Over-sampling Technique*) para generar nuevas muestras sintéticas de las clases minoritarias y así equilibrar el conjunto de datos. Esta estrategia permitió mejorar la capacidad de los modelos para generalizar patrones de las clases subrepresentadas.

Para entender mejor estas técnicas de remuestreo, es fundamental comprender sus diferencias y cómo operan. Aunque ambas generan muestras sintéticas mediante la interpolación, difieren en cómo seleccionan las muestras originales para la generación de muestras sintéticas y en cómo adaptan su enfoque para abordar el desequilibrio de clases en el conjunto de datos.

ADASYN (*Adaptive Synthetic Sampling*)

ADASYN (He, Bai, Garcia, y Li, 2008) se centra en las muestras que son difíciles de clasificar según una regla de vecinos más cercanos. Esto significa que ADASYN asigna más peso a las instancias minoritarias que están en regiones de baja densidad o cercanas a la frontera de decisión del clasificador.

De esta froma, un clasificador que depende fuertemente de las distancias entre las instancias, como los vecinos más cercanos, podría beneficiarse más de ADASYN, ya que genera muestras sintéticas en áreas de difícil clasificación, mejorando así la capacidad del modelo para generalizar en esas regiones.

SMOTE (*Synthetic Minority Over-sampling Technique*)

Sin embargo, SMOTE (Han, Wang, y Mao, 2005) genera nuevas muestras sintéticas para las clases minoritarias al tomar puntos entre instancias de la misma clase. No se hace ninguna distinción entre las muestras fáciles y difíciles de clasificar. Simplemente se generan muestras sintéticas mediante la interpolación entre instancias de la clase minoritaria existente.

Ambas técnicas tienen como objetivo principal abordar el desequilibrio de clases en conjuntos de datos, permitiendo a los modelos aprender de manera más efectiva las características de las clases minoritarias. Sin embargo, es importante tener en cuenta que el uso de estas técnicas debe ser cuidadosamente considerado, ya que la generación excesiva de muestras sintéticas podría conducir a un sobreajuste en el modelo. Es fundamental encontrar un equilibrio adecuado entre la corrección del desbalance y la preservación de la generalización del modelo.

4.2 Modelos de clasificación

Para la predicción de la variable ISUP del cáncer de próstata, se consideran varios modelos de clasificación, que serán entrenados y evaluados, para luego poder elegir el mejor. Los modelos son:

Tabla 4.1: Clasificadores

CLASIFICADORES LINEALES
Regresión Logística
Regresión Logística con regularización Lasso
Regresión Ridge
Stochastic Gradient Descent (SGD)
CLASIFICADORES BASADOS EN ÁRBOLES
Árbol de Decisión
CLASIFICADORES DE ENSAMBLADO
Bosques Aleatorios (Random Forest)
AdaBoost
Gradient Boosting
CLASIFICADORES SVC
Support Vector Classifier (SVC)
CLASIFICADORES DE VECINOS MÁS CERCANOS
k-Nearest Neighbors (kNN)
CLASIFICADORES DE NAIVE BAYES
Naive Bayes Gaussiano
CLASIFICADORES DE RED NEURONAL
Perceptrón Multicapa (MLP)

A continuación, una breve explicación. (García, 2023)

4.2.1. Clasificadores lineales

Estos tratan de separar los datos en diferentes clases encontrando una combinación lineal de las características que separe máximamente las clases. Los datos han de ser separables de manera lineal, es decir, deben poder ser separados por una línea recta. Se caracteriza por su simplicidad y eficiencia computacionales.

Sin embargo, cada modelo tiene una particularidad distinta (Gonzalez, 2019) (Jain, 2023):

- **Regresión Logística**. Un modelo estadístico que utiliza una función logística, llamada sigmoide, para modelar la probabilidad de que una variable dependiente sea una clase particular.

- **Regresión Logística con regularización Lasso**. Similar a la regresión logística, pero con una penalización adicional para los coeficientes. Esto ayuda a prevenir el sobreajuste.

- **Regresión Ridge**. Otra variante de la regresión logística que utiliza una penalización diferente para los coeficientes.

- **Stochastic Gradient Descent (SGD)**. Es un algoritmo de optimización utilizado para entrenar modelos de aprendizaje automático, incluida la regresión logística, de manera eficiente. Busca encontrar el límite de decisión óptimo (un hiperplano) para separar los puntos de datos que pertenecen a diferentes clases en un espacio de características. Opera mediante el descenso por gradiente, de ahí su nombre.

4.2.2. Árboles de decisión

El modelo tiene forma de árbol con nodos, ramas y hojas. Cada nodo representa una prueba en una característica; cada rama representa el resultado de la prueba; y cada hoja representa la etiqueta de clase.

El modelo se entrena dividiendo recursivamente los datos en conjuntos más pequeños en base a una prueba. Esto se hace de manera recursiva hasta que se alcanza un criterio de parada.

Son modelos simples de entender e interpretar.

4.2.3. Clasificadores de ensamblado

Combinan las predicciones de varios clasificadores base para mejorar la precisión general del modelo.

Se basan en la idea de que las predicciones combinadas de múltiples modelos son, a menudo, más precisas que las de cualquier modelo individual.

Pueden ser útiles para reducir el sobreajuste (*overfitting*) y mejorar la generalización del modelo.

- **Bosques Aleatorios** (Random Forest). Un conjunto de árboles de decisión entrenados en subconjuntos aleatorios de los datos, cuyas predicciones se combinan para mejorar la precisión.
- **AdaBoost**. Un algoritmo de ensamblaje que entrena iterativamente clasificadores débiles, asignando mayor peso a los ejemplos mal clasificados en cada iteración.
- **Gradient Boosting**. Otro algoritmo de ensamblaje que construye un modelo aditivo de forma secuencial, minimizando el error residual en cada etapa.
  (AIML, 2023)

4.2.4. Clasificadores de máquina de soporte vectorial

Se centran en encontrar el hiperplano de decisión óptimo que mayor separa las instancias de las diferentes clases. El hiperplano se elige de forma que la distancia desde él hasta los ejemplos más cercanos de cada clase sea máxima.

Si las clases no son linealmente separables, las máquinas de soporte vectorial utilizan transformaciones no lineales para llevarlas a un espacio donde se puedan separar.

Se caracterizan por un buen rendimiento y por la capacidad para manejar datos de alta dimensionalidad. Pueden ser ser sensibles al ajuste de hiperparámetros y pueden ser lentas al entrenar grandes conjuntos de datos.

4.2.5. Clasificadores de vecinos más cercanos

Almacenan los ejemplos de entrenamiento y predicen la clase de un ejemplo de prueba encontrando los ejemplos almacenados que son más similares a él, basándose en una medida de

distancia.

Son simples y fáciles de interpretar. Pueden ser lentos para clasificar nuevos ejemplos. Requieren gran cantidad de memoria para almacenar los ejemplos de entrenamiento.

4.2.6. Clasificadores de naive Bayes

Utilizan el teorema de Bayes para predecir la clase de un ejemplo de prueba basándose en la probabilidad de que el ejemplo pertenezca a cada clase, dado las características de ese ejemplo.

Además, suponen que las características del ejemplo son independientes entre sí, lo que significa que la probabilidad de cada característica no se ve afectada por la presencia o ausencia de otras características.

Destacan por su simplicidad y por los tiempos rápidos de entrenamiento y predicción.

- **Naive Bayes Gaussiano**. Una variante de Naive Bayes que asume que las características se distribuyen de acuerdo a una distribución gaussiana.

4.2.7. Clasificadores de redes neuronales

Estos modelos están inspirados en la estructura y funcionamiento del cerebro humano. Consisten en capas de neuronas artificiales interconectadas, donde cada neurona realiza una operación matemática en la entrada que recibe y produce una salida. Aprenden patrones complejos en los datos ajustando los pesos de las conexiones entre neuronas durante el entrenamiento.

Son efectivos para problemas complejos con grandes conjuntos de datos, pero pueden requerir mucho tiempo de entrenamiento y ajuste de hiperparámetros. Son menos interpretables que otros modelos.

- **Perceptrón Multicapa** (MLP). Un tipo de red neuronal artificial con múltiples capas ocultas entre la capa de entrada y la capa de salida, que utiliza retropropagación para aprender los pesos de las conexiones entre neuronas. (Gamco, s.f.)

4.3 Entrenamiento y evaluación de modelos

Ahora que los datos estaban preparados, comenzamos el proceso de entrenamiento de los modelos. Cada modelo viene con un conjunto específico de hiperparámetros, y es crucial encontrar la combinación óptima que maximice su rendimiento predictivo. Para lograr esto, utilizamos una técnica llamada búsqueda en cuadrícula (Grid Search) con validación cruzada.

La búsqueda en cuadrícula implica probar cada modelo con diferentes configuraciones de hiperparámetros y evaluar su rendimiento utilizando la métrica deseada. En nuestro caso, utilizamos el F1-score como métrica de evaluación. Esto nos permite identificar la combinación de hiperparámetros que optimiza el rendimiento del modelo.

Además, la validación cruzada es una técnica que nos ayuda a evaluar la capacidad de generalización del modelo al dividir el conjunto de datos en múltiples subconjuntos; en este caso 5.

De esta forma, se entrena el modelo en una parte y se valida en otra, de manera iterativa. Esto nos proporciona una estimación más precisa del rendimiento del modelo en datos no vistos.

Una vez que hemos encontrado esta combinación óptima de hiperparámetros, procedemos a entrenar los modelos utilizando el conjunto de datos de entrenamiento. Luego, evaluamos el rendimiento de los modelos entrenados utilizando el conjunto de datos de prueba, lo que nos proporciona una medida objetiva de su capacidad para generalizar a nuevos datos.

Finalmente, calculamos diferentes métricas de evaluación para medir el rendimiento de los modelos. Dado que se trata de un caso clínico que predice el nivel de gravedad del cáncer, preferimos usar métricas que penalicen los falsos negativos para no perder a ningún paciente enfermo. Estas métricas son *accuracy*, *recall*, *precision* y *f1 score* ponderado. Estas métricas proporcionan una visión completa del rendimiento de los modelos en la predicción de la variable ISUP del cáncer de próstata.

Antes de analizar los resultados, es fundamental conocer cada una de las métricas utilizadas para así entender mejor cómo de bueno o malo es el rendimiento de los modelos. (Barrios, 2022)

4.3.1. *Accuracy*

El *accuracy* o exactitud mide la proporción de predicciones correctas realizadas por el modelo sobre el total de casos. Indica como de exacto es un modelo. No obstante, no es la mejor métrica cuando las clases están desbalanceadas. Aunque el re-muestreo hace que sea más confiable, es mejor usar otras métricas para tener una evaluación más completa del rendimeinto del modelo.

4.3.2. *Recall*

El *recall* o sensibilidad mide la proporción de casos positivos verdaderos identificados por el modelo sobre los positivos reales. Indica cómo de bueno es el modelo para predecir los casos positivos. Dado que queremos minimizar los casos negativos, buscamos que este número sea lo más alto posible, pues eso significará que la cantidad de falsos negativos es muy baja.

4.3.3. *Precision*

En español precisión, mide cuántos de los casos positivos identificados por el modelo son positivos reales. Muestra cómo de precisos son los casos positivos del modelo.

4.3.4. *F1 score*

Es una medida que combina precisión y sensibilidad en un solo número. Es la media armónica que tiene en cuenta tanto los falsos positivos como los falsos negativos. Además, en este caso usamos el *F1 score weighted* o ponderado, que tiene en cuenta el desequilibrio de clases. Calcular un promedio ponderado del puntaje F1 para cada clase. Este promedio se calcula considerando la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos de cada clase. Una vez más, buscamos que este número sea lo más alto posible.

CAPÍTULO 5
Resultados

La puntuación ISUP (*International Society of Urological Pathology*) es una medida que se utiliza para clasificar el cáncer de próstata en diferentes categorías según su agresividad. Esta variable toma valores desde 1 hasta 5, donde 1 indica un cáncer poco agresivo y 5 indica un cáncer altamente agresivo.

Dada la información proporcionada en el proyecto Pro-Cancer mencionado anteriormente, se postula que las clases ISUP 1 y 2 podrían ser más difíciles de separar en la clasificación, ya que las diferencias en la supervivencia global entre estas dos clases no son estadísticamente significativas. Por otro lado, se espera que a partir de la clase ISUP 3, se observe una peor supervivencia global, lo que podría facilitar la distinción entre estas clases y las clases más agresivas (ISUP 4 y 5).

En base a esta hipótesis, procedemos a analizar cómo se distribuyen y separan las clases en los modelos de clasificación desarrollados y qué recategorización funciona mejor.

Inicialmente, exploramos dos recategorizaciones: agrupando las categorías 1 y 2 frente a las categorías 3, 4 y 5, y luego agrupando la categoría 1 frente a la 2 y las categorías 3, 4 y 5. Sin embargo, los resultados obtenidos no alcanzaron nuestras expectativas en términos de rendimiento predictivo. Así pues, probamos con otras estrategias distintas.

Cabe recordar que, para cada clasificación, se utilizan dos técnicas de remuestreo, ADASYN y SMOTE, con el propósito de abordar de manera exhaustiva y comparativa el problema del desequilibrio de las clases en el conjunto de datos usado.

Como se ha mencionado previamente, estas técnicas difieren en su enfoque y aplicación. ADASYN tiene la capacidad de enfocar su esfuerzo de generación de muestras en las regiones donde se requiere mayor atención, lo que puede resultar en una representación más precisa de la distribución subyacente de los datos y, por ende, en una mayor capacidad de generalización del modelo.

Por otro lado, SMOTE destaca por su simplicidad y efectividad al generar muestras sintéticas a través de la interpolación de instancias minoritarias, lo que contribuye a una mayor diversidad en los datos generados. Esta diversidad adicional puede ayudar a prevenir el sobreajuste al proporcionar al modelo una representación más equilibrada de las clases minoritarias.

La utilización de ambas técnicas busca explorar la influencia de estas estrategias de remuestreo en la calidad de los modelos resultantes, así como identificar posibles escenarios en los que una técnica pueda superar a la otra en términos de desempeño predictivo y estabilidad del modelo.

No obstante al aplicar el remuestreo, las proporciones de las clases en el conjunto de entrenamiento se ven alteradas. A pesar de haber realizado una estratificación al dividir los datos para mantener la proporción original de las clases, el proceso de sobremuestreo cambia estas proporciones en el conjunto de entrenamiento.

Este cambio en la distribución de clases solo se aplica al conjunto de datos de entrenamiento. El objetivo es influir en el ajuste de los modelos para que estos puedan aprender de forma óptima y capturar los patrones subyacentes en los datos. (Brownlee, 2021) El remuestreo no se aplica al conjunto de datos de prueba, pues este es para evaluar el rendimiento del modelo y debe reflejar datos del mundo real.

Cuando se realiza el sobremuestreo solo en el conjunto de datos de entrenamiento, se generan nuevas muestras sintéticas solo a partir de las instancias existentes en el conjunto de entrenamiento, sin involucrar ninguna información del conjunto de datos de validación. (Becker, 2016)

Esto significa que las nuevas muestras generadas durante el sobremuestreo no se basan en ninguna información proporcionada por el conjunto de datos de validación, lo que asegura que este conjunto de datos de validación se mantenga intacto y sin manipulación.

De esta forma, cuando evaluamos el modelo utilizando el conjunto de datos de validación, estamos probando su rendimiento en datos que no han sido alterados artificialmente por el proceso de remuestreo. Esta estrategia garantiza una evaluación imparcial y precisa del modelo en datos reales, sin introducir sesgos artificiales.

5.1 CLASIFICACIÓN 1 - 1,2 vs 3,4,5

En esta clasificación, se están intentando distinguir los cánceres poco agresivos (clases ISUP 1 y 2) de los más agresivos (clases ISUP 3, 4 y 5).

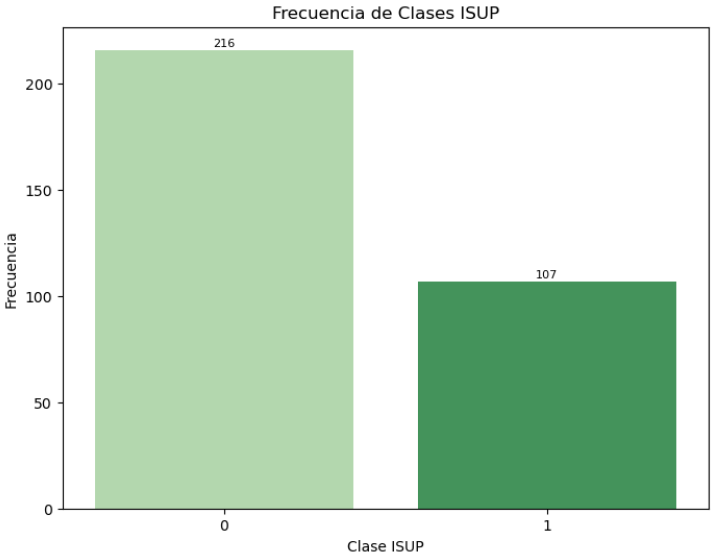

Figura 5.1: Proporción de las clases en la Clasificación 1

5.1.1. ADASYN

N (entrenamiento): 342 N (test): 65

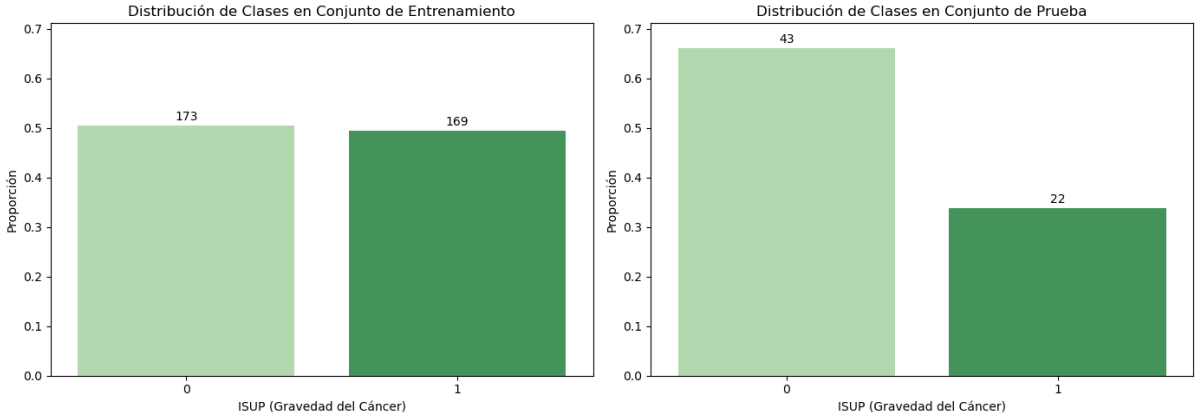

Figura 5.2: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 1 con ADASYN

En este gráfico podemos ver como, el sobremuestreo trata de tener el mismo número de muestras para cada clase en el conjunto de entrenamiento. Sin embargo, en el conjunto de prueba se mantienen las proporciones originales.

Tabla 5.1: Clasificación 1: ADASYN

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.6308	0.6308	0.6392	0.6344
Logistic Regression Lasso	0.5231	0.5231	0.5383	0.5297
Ridge Classifier	0.6308	0.6308	0.6392	0.6344
SGD Classifier	0.6308	0.6308	0.6683	0.6402
Decision Tree Classifier	0.5385	0.5385	0.5485	0.5430
Random Forest Classifier	0.7385	0.7385	0.7289	0.7289
AdaBoost Classifier	0.6154	0.6154	0.6111	0.6131
Gradient Boosting Classifier	0.6615	0.6615	0.6544	0.6573
Support Vector Classifier	0.6308	0.6308	0.6071	0.6137
k-Nearest Neighbors	0.5538	0.5538	0.5878	0.5647
Gaussian Naive Bayes	0.3692	0.3692	0.4945	0.3218
Multi-layer Perceptron	0.6462	0.6462	0.6210	0.6260

Esta tabla muestra los resultados de las métricas utilizadas.

No obstante, dado que estamos trabajando con un *dataset* desbalanceado, hemos optado por trabajar con las métricas ponderadas; *weighted average*. Esto quiere decir que, se calcula cada métrica por separado para cada clase. Luego, se multiplica cada una por la proporción de instancias de esa clase. El resultado final es la suma de estos dos números. Es decir, cada clase contribuye con un peso proporcional a su soporte en el conjunto de datos.

Esto ayuda a dar una idea más equilibrada del rendimiento del modelo considerando el impacto relativo de cada clase.

Por tanto, a excepción del *accuracy*, los resultados de las tablas muestran las métricas ponderadas. El *accuracy* no tiene una media ponderada porque simplemente mide la proporción de predicciones correctas sobre el total de casos. No tiene en cuenta la distribución de clases en los datos, por lo que no tiene sentido asignar pesos a diferentes clases.

Podemos agrupar a los modelos según su rendimiento.

- Random Forest Classifier es el mejor modelo con resultados superiores al 70 % en todas las métricas. Destaca por tener una mayor precisión y *recall* en comparación con otros modelos. Esto sugiere que es capaz de identificar correctamente una proporción significativa de casos positivos, que es lo que estamos buscando en general, pero luego haremos algunas observaciones sobre esto.
- Otros modelos como Logistic Regression, Ridge Classifier, SDG, AdaBoost, Gradient Boosting, SVC y Multi-layer Perceptron tienen un rendimiento similar, por encima del 60 %. Sus valores de *recall* son más bajos, lo que indica que pueden estar perdiendo algunos casos positivos.
- Gaussian Naive Bayes tiene un *recall* muy bajo, lo que indica que está clasificando incorrectamente muchos casos positivos.

Random Forest es el mejor modelo en este caso, con un *recall* de 0.74. Esto quiere decir que, el modelo logra recuperar correctamente el 74 % de los casos positivos en ambas las clases. Para entender mejor como funciona, vamos a analizar su matriz de confusión.

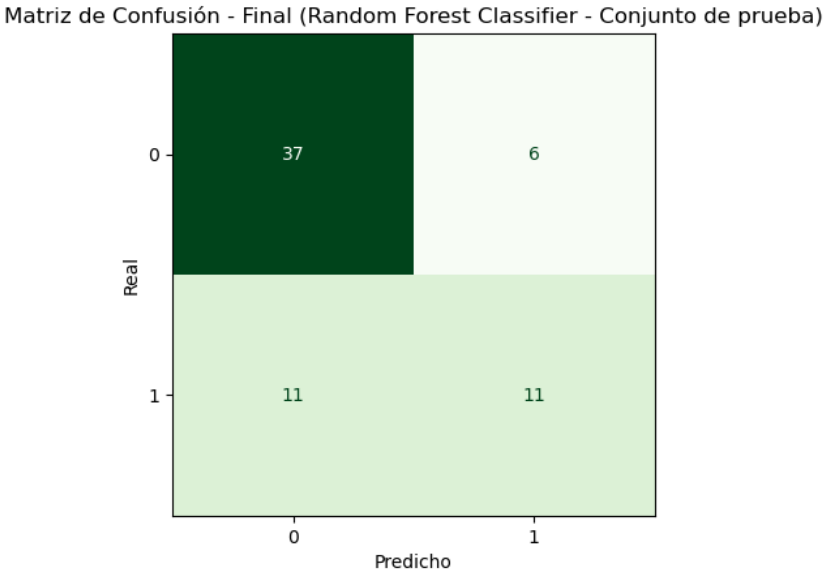

Figura 5.3: Matriz de confusión para Clasificación 1 con ADASYN

Una matriz de confusión funciona de la siguiente manera; en la diagonal tenemos los casos predichos correctamente. Si atendemos a las filas, encontramos el total de casos actuales; mientras que las columnas son los casos predichos. Ahora vamos a dividir por clases:

Clase 1 como positiva

El primer valor de la matriz se lee como: 'Pertenece a la clase 0 (ISUP 1, 2) y se ha predicho como clase 0'. Esto es lo que se entiende como TN, *true negative* o negativo verdadero. De igual forma, su valor de la derecha sería: 'Pertenece a la clase 0 pero es predicho como clase 1'. Esto es un FP, *false positive* o falso positivo.

El primer valor de la segunda fila sería: 'Pertenece a la clase 1 pero se ha predicho como 0'. Esto es un falso negativo o *false negative*, FN. Finalmente, el último valor es TP, *true positive*: 'Es de la clase 1 y se clasifica como tal'.

Entonces, si sabemos que el *recall* es la proporción de casos positivos que se han clasificado correctamente, se calcula como TP / (TP + FN).

Para esta clase será: 11 / (11 + 11) = 0.5

También tenemos que calcular la proporción de instancias de esta clase. Como hemos dicho, las filas representan los valores actuales, así que hay 22 instancias de la clase 1. Entre el total de casos; 22 / 65 = 0.338.

Clase 0 como positiva

Ahora los valores de positivo y negativo están invertidos. El primer valor de la matriz será ahora los TP (*true positive*). El siguiente valor representa los FN. Los últimos dos valores son FP y TN, respectivamente.

El cálculo del *recall* queda como: 37 / (37 + 6) = 0.86

La proporción es 43 / 65 = 0.661

Recall ponderado

Por tanto, el *recall* final se calcula como:

0.5 * 0.338 + 0.86 * 0.661 = 0.737

Esto explica por qué el *accuracy* y *recall* son iguales. Al tener en cuenta ambas clases las fórmulas acaban siendo iguales.

Tal y como hemos mencionado antes, este es el rendimiento ponderado, pero es importante parar a analizar lo que esto quiere decir. Este modelo clasifica mejor a la clase 0. Es decir, tiene un mejor rendimiento en la clasificación de los ISUP 1 y 2 en comparación con los otros. Según los números obtenidos, podemos saber que solo clasifica correctamente el 50 % de los casos de las ISUP 3, 4 y 5; mientras que con las otras ISUP clasifica correctamente el 86 % de los casos. Esto sugiere que el modelo puede tener dificultades para distinguir entre las clases más altas de ISUP (3, 4 y 5), lo que destaca la disparidad en el rendimiento del modelo para diferentes clases.

Esto puede deberse a la diferencia en el número de muestras; el modelo podría estar teniendo dificultades para aprender patrones en las clase minoritaria debido a la falta de ejemplos de entrenamiento.

Aunque se ha intentado corregir el desbalance con las técnicas de remuestreo, no ha sido suficiente.

Para calcular *precision* y *f1-score*, se hace igual.

También es interesante analizar las variables que más aportan al modelo.

Esto nos proporciona información valiosa sobre qué características tienen el mayor impacto en la predicción del modelo. Puede ayudar a comprender mejor la relación entre las características de los datos y la variable objetivo. Asimismo, puede orientar la selección de características para mejorar la precisión y la interpretabilidad del modelo, así como identificar posibles sesgos o problemas en los datos.

En un modelo de Random Forest, la importancia de las variables se calcula mediante la contribución de cada variable a la reducción del error en las predicciones del modelo. Es decir, cuánto más disminuye la precisión del modelo si se elimina una variable durante la contrucción de árboles, más importante es la variable.

No es necesario que la variable más importante tenga un valor de 1, porque la importancia está normalizada. Hay que fijarse en la relación relativa entre ellas.

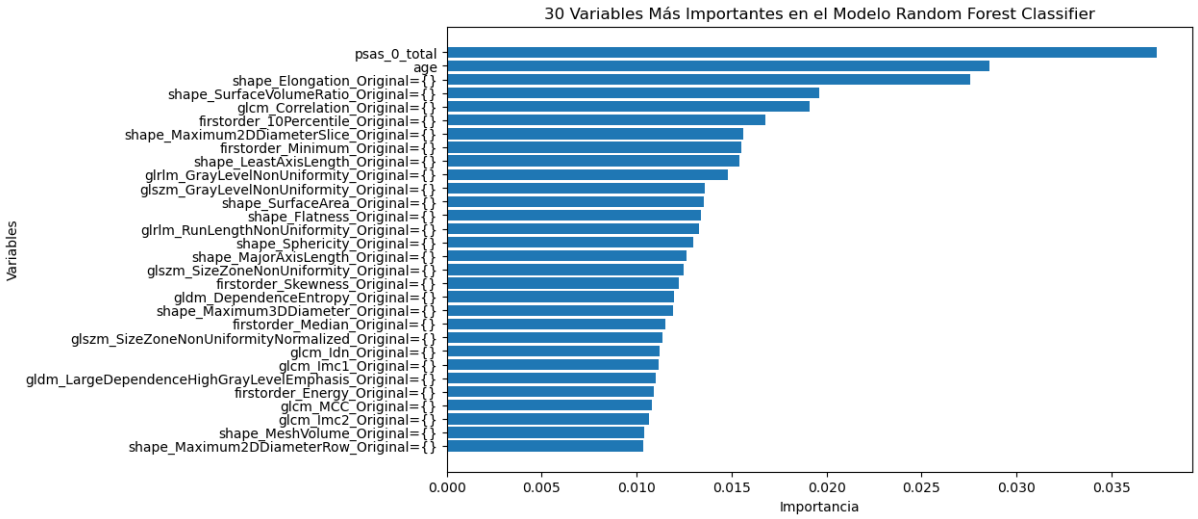

Figura 5.4: Variables más importantes para Clasificación 1 con ADASYN

Las dos primeras son variables clínicas. En primer lugar, el PSA. Este se extrae a partir de la sangre. Además, es un factor pronóstico relacionado con la agresividad tumoral. Con lo que tiene sentido que sea la variable que más contribuye.

A continuación, encontramos la edad. Se cree que, a menor edad, el cáncer es más agresivo porque parte de antecedentes familiares, factores genéticos, etc.

Luego ya son todo variables radiómicas.

- Las variables de entropía reflejan el nivel de desorden celular. Con desorden nos referimos a una discrepancia respecto a las células normales. La clasificación Gleason, que es a partir de lo que se clacula el ISUP, se basa en este desorden. Por tanto, a mayor desorden y mayor entropía, se espera un cáncer más agresivo.
- Las variables de forma indican que a medida que aumenta el volumen, área y elongación del tumor, se espera una mayor agresividad.
- Las características de intensidad de los vóxeles se extraen de la secuencia de T2-w de RM. Las intensidades bajas o los valores más bajos corresponden a zonas más hipointensas en la RM, lo que sugiere mayor agresividad tumoral. Además, la falta de simetría en la distribución de intensidades, como se observa en el skewness, también indica focos de mayor desorden celular y, por ende, tumores más agresivos.

5.1.2. SMOTE

Los siguientes resultados siguen siendo de la primera clasificación pero, en este caso, el remuestreo se hizo con SMOTE. Se van a comentar estos resultados y compararlos con los de ADASYN.

N (entrenamiento): 346 N (test): 65

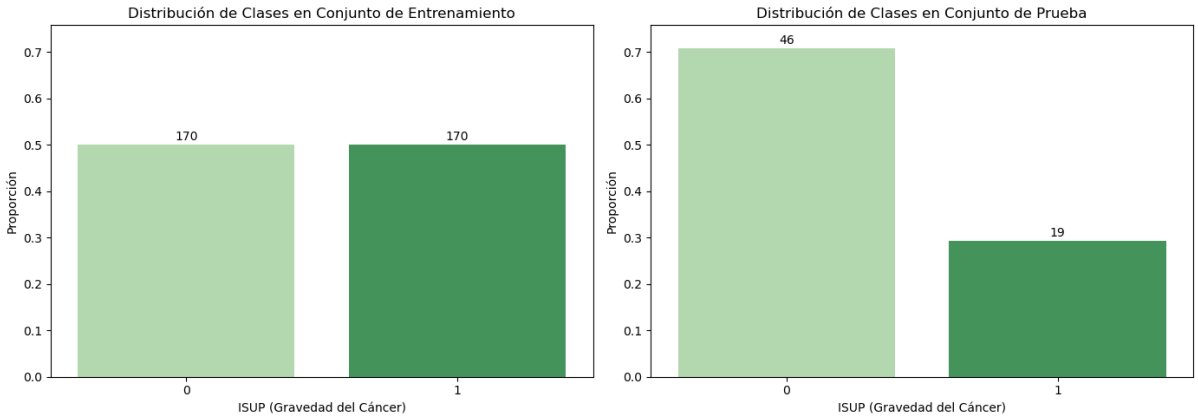

Figura 5.5: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 1 con SMOTE

Tabla 5.2: Clasificación 1: SMOTE

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.6769	0.6769	0.6734	0.6750
Logistic Regression Lasso	0.5385	0.5385	0.5485	0.5430
Ridge Classifier	0.6308	0.6308	0.6227	0.6262
SGD Classifier	0.7385	0.7385	0.7577	0.7437
Decision Tree Classifier	0.6615	0.6615	0.6694	0.6649
Random Forest Classifier	0.6923	0.6923	0.6733	0.6712
AdaBoost Classifier	0.5692	0.5692	0.5595	0.5639
Gradient Boosting Classifier	0.6000	0.6000	0.5820	0.5889
Support Vector Classifier	0.7077	0.7077	0.6926	0.6910
k-Nearest Neighbors	0.6615	0.6615	0.6615	0.6615
Gaussian Naive Bayes	0.4154	0.4154	0.5561	0.3879
Multi-layer Perceptron	0.6923	0.6923	0.6733	0.6712

- Los modelos con un mayor rendimiento son SGD Classifier, Support Vector Classifier, Random Forest Classifier y Multi-layer Perceptron, con un *recall* aproximado de 0.7.
- Gaussian Naive Bayes vuelve a ser el peor modelo pero con unos resultados más altos.

En general, aunque los resultados son bastante consistentes, SMOTE presenta unos valores generalmente más elevados.

En este caso, el modelo con el *recall* más alto es SDG classifier. Aunque el *accuracy* y el *recall* son iguales que los de Random Forest de ADASYN, las otras dos métricas son mejores en esta tabla.

El modelo es capaz de identificar correctamente el 74 % de todos los casos positivos en el conjunto de datos (*recall*). Además, el 76 % de las instancias clasificadas como positivas por el modelo son realmente positivas (*precision*).

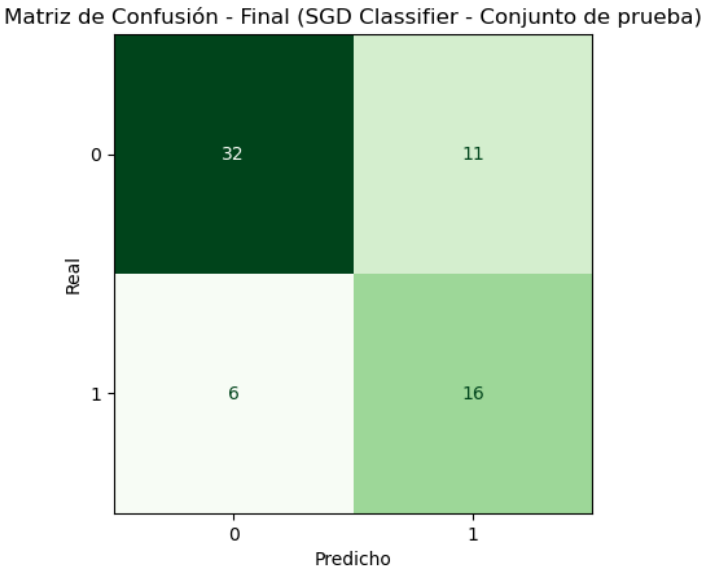

Figura 5.6: Matriz de confusión para Clasificación 1 con SMOTE

Las métricas se calculan de igual forma que se ha explicado antes.

De igual forma que hemos hecho antes, podemos ver cómo de bueno es el rendimiento del modelo según las clases. Ahora parece que el modelo clasifica un poco peor la clase 0, con un *recall* de 0.74 frente a 0.86 de antes. Sin embargo, predice mucho mejor la clase 1. El 72 % de los casos de los ISUPs más agresivos son clasificados correctamente.

Esto es otro indicativo de que este modelo es mejor que el Random Forest de SMOTE, ya que demuestra un desempeño más consistente en la clasificación de la clase agresiva.

En un modelo SGD, la importancia de las variables puede evaluarse mediante la magnitud de los coeficientes asociados a cada variable en la función de pérdida que está siendo optimizada. Las variables con coeficientes de mayor magnitud se consideran más importantes en términos de su contribución a la predicción del modelo.

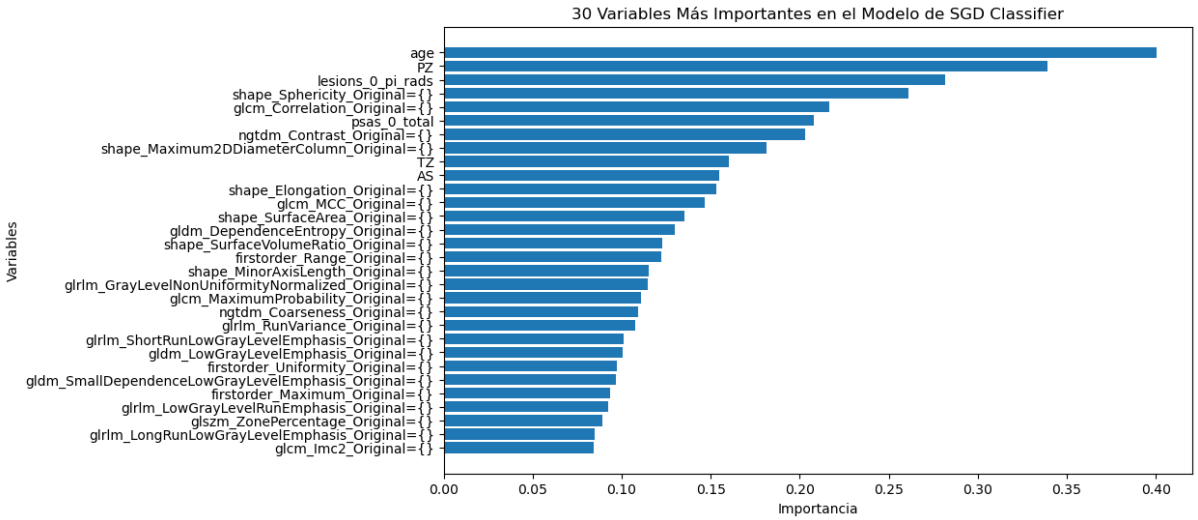

Figura 5.7: Variables más importantes para Clasificación 1 con SMOTE

Al graficar las variables que más contribuyen podemos observar que la edad se mantiene y el PSA contribuye un poco menos que antes. De variables clínicas se añaden aquellas que indican si hay infiltración en diferentes zonas de la próstata y, también el PI-RADS. Es un sistema utilizado para interpretar los resultados de resonancias magnéticas de la próstata cuando se sospecha la presencia de cáncer en pacientes sin tratamiento previo. Dado que los radiólogos lo utilizan para evaluar posibles indicios de cáncer, tiene sentido que el modelo lo use como variable predictora. Un PI-RADS más alto se relaciona con un mayor riesgo de cáncer de próstata clínicamente significativo y se asocia con un cáncer más agresivo.

5.2 CLASIFICACIÓN 2 - 1 vs 2,3 vs 4,5

Aquí, la idea es separar aún más las clases, agrupando la ISUP 1 en una categoría, sin riesgo; las ISUP 2 y 3 en otra, bajo riesgo; y las ISUP 4 y 5 en una tercera, alto riesgo.

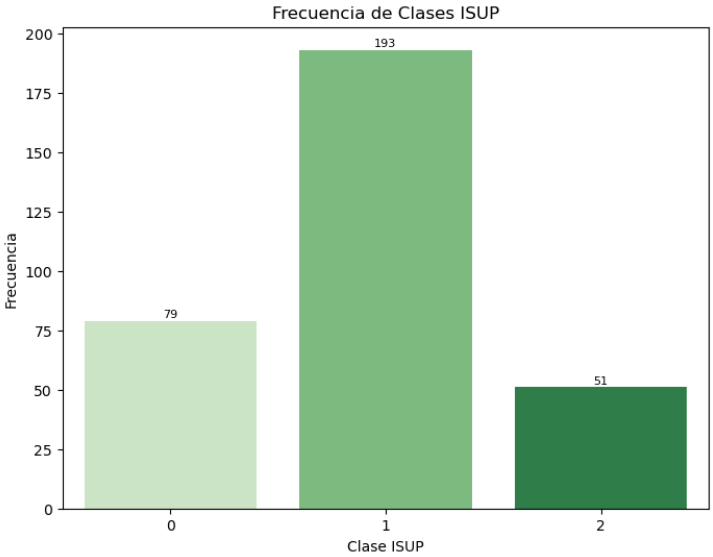

Figura 5.8: Proporción de las clases en la Clasificación 2

5.2.1. ADASYN

N (entrenamiento): 464 N (test): 65

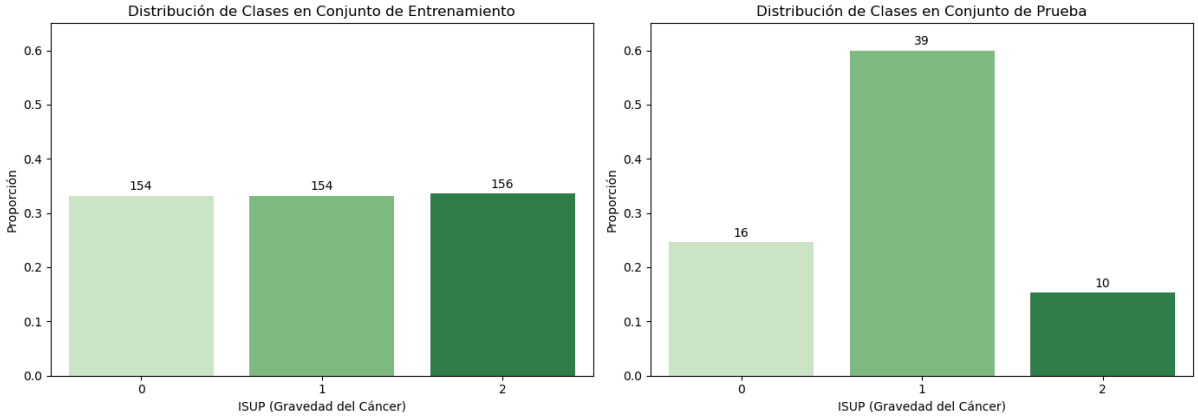

Figura 5.9: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 2 con ADASYN

Tabla 5.3: Clasificación 2: ADASYN

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.4154	0.4154	0.4792	0.4261
Logistic Regression Lasso	0.4000	0.4000	0.4569	0.4186
Ridge Classifier	0.5077	0.5077	0.5835	0.5231
SGD Classifier	0.3692	0.3692	0.4830	0.3601
Decision Tree Classifier	0.4615	0.4615	0.4528	0.4568
Random Forest Classifier	0.5231	0.5231	0.5367	0.5287
AdaBoost Classifier	0.4462	0.4462	0.4580	0.4512
Gradient Boosting Classifier	0.5231	0.5231	0.5229	0.5148
Support Vector Classifier	0.4769	0.4769	0.5009	0.4875
k-Nearest Neighbors	0.4000	0.4000	0.4867	0.4035
Gaussian Naive Bayes	0.2462	0.2462	0.2481	0.1650
Multi-layer Perceptron	0.5385	0.5385	0.5605	0.5464

A primera vista, los resultados muestran un rendimiento bastante peor. Aunque, se trata de las métricas ponderas, el *recall* más alto está alrededor de 0.5. Esto sugiere que los modelos son bastante aleatorios, ya que solo el 50 % de las veces se predice correctamente. Como consecuencia, podemos concluir que los modelos no son fiables.

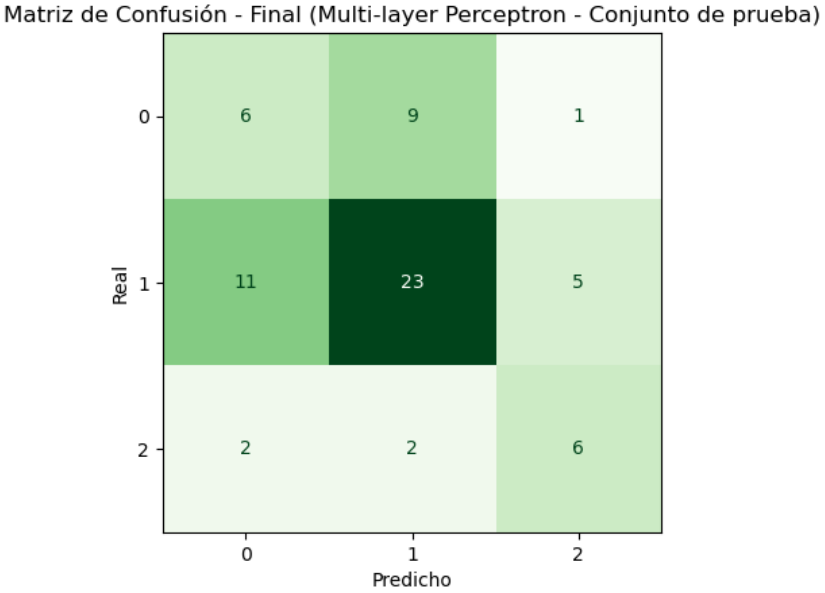

Figura 5.10: Matriz de confusión para Clasificación 2 con ADASYN

Si atendemos a la matriz de confusión del mejor modelo, MLP, podemos entender como rinde el modelo para las distintas clases.

- Para la clase 0, el *recall* es 0.375
- Para la clase 1, 0.589
- Para la clase 2, 0.6

Esto implica que el modelo clasifica mejor las clases más agresivas, que es lo que estamos buscando. Sin embargo, ahora le cuesta más clasificar la clase 0. Esto puede deberse a la hipótesis inicial, que decía que los ISUP 1 y 2 son más difíciles de separar.

Aunque la importancia de las variables en un MLP es más difícil de representar, algunas técnicas pueden proporcionar información sobre cómo las diferentes variables afectan la salida de la red neuronal.

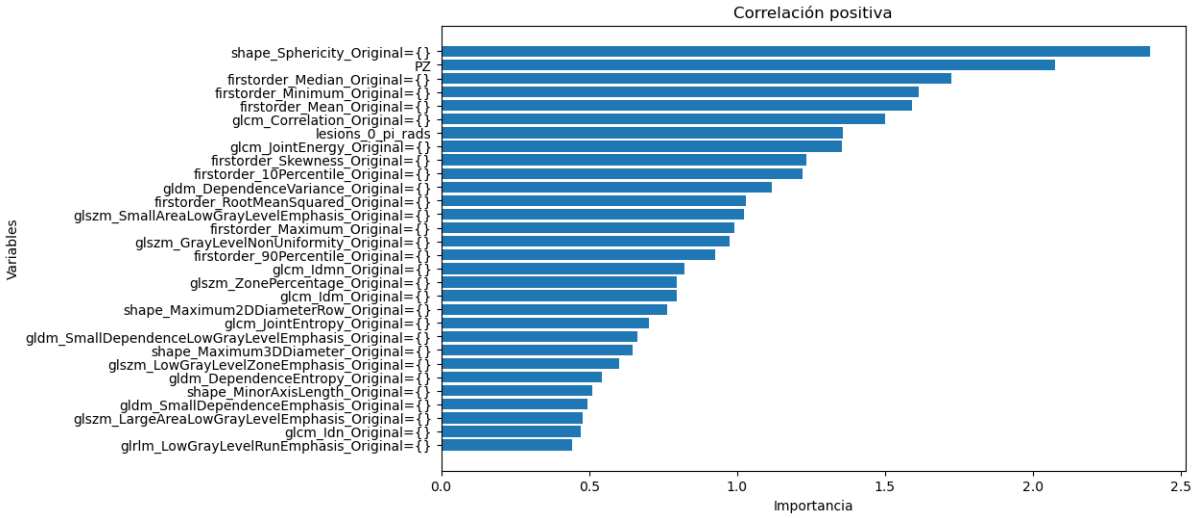

Figura 5.11: Variables más importantes para Clasificación 2 con ADASYN

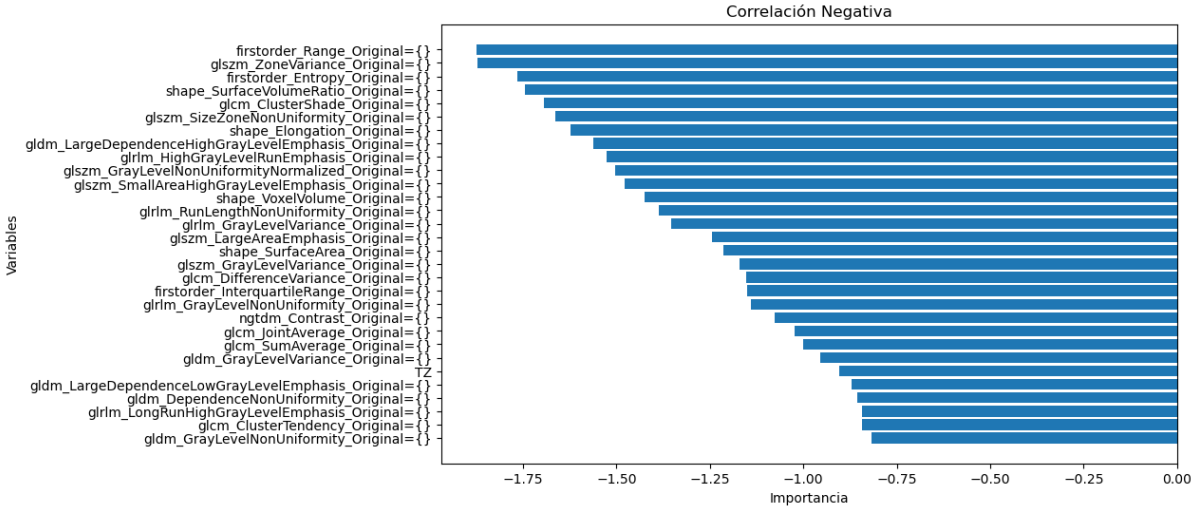

Figura 5.12: Variables más importantes para Clasificación 2 con ADASYN

En este modelo, vemos que hay variables que influyen de forma positiva y otras que influyen de forma negativa.

La variable PI-RADS se mantiene y, por tanto, un PI-RADS más alto se sigue relacionando con un cáncer más agresivo.

La variable PZ indica si hay infiltración en la zona periférica de la próstata. Esta influye de forma positiva. Dado que es una variable booleana, si hay infiltración, entonces el cáncer será más agresivo.

Sin embargo, la variable TZ que es de la zona de transición, se correlaciona negativamente. Si hay infiltranción en esa zona, el cáncer será menos agresivo.

Las variables radiómicas influyen tanto positiva, como negativamente.

5.2.2. SMOTE

N (entrenamiento): 462 N (test): 65

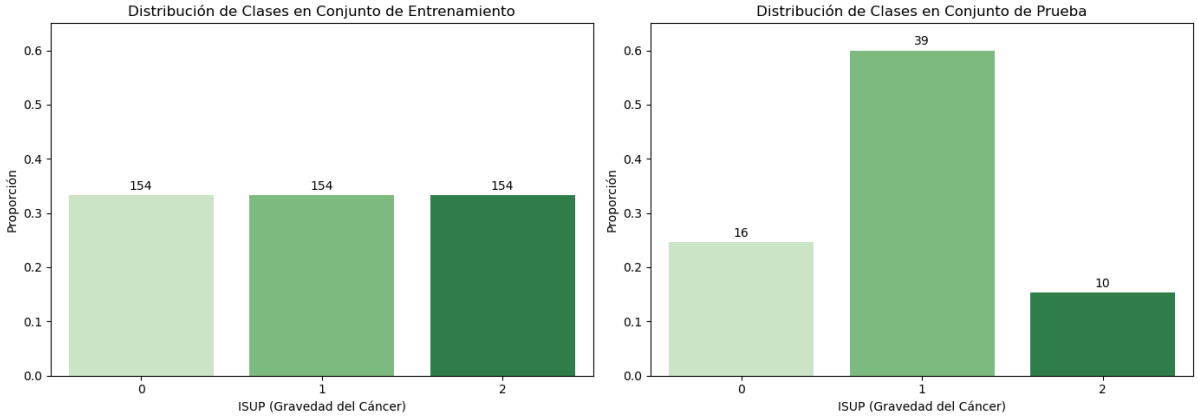

Figura 5.13: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 2 con SMOTE

Tabla 5.4: Clasificación 2: SMOTE

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.4308	0.4308	0.5067	0.4473
Logistic Regression Lasso	0.4308	0.4308	0.4874	0.4483
Ridge Classifier	0.5231	0.5231	0.5925	0.5385
SGD Classifier	0.4154	0.4154	0.5923	0.4051
Decision Tree Classifier	0.4923	0.4923	0.4996	0.4955
Random Forest Classifier	0.5077	0.5077	0.5329	0.5175
AdaBoost Classifier	0.4615	0.4615	0.4819	0.4662
Gradient Boosting Classifier	0.5538	0.5538	0.5202	0.5344
Support Vector Classifier	0.4462	0.4462	0.4671	0.4557
k-Nearest Neighbors	0.4000	0.4000	0.4676	0.4103
Gaussian Naive Bayes	0.2615	0.2615	0.3036	0.1739
Multi-layer Perceptron	0.5077	0.5077	0.5577	0.5260

Los resultados obtenidos muestran un rendimiento generalmente consistente con los modelos previos, aunque se observa una ligera mejora, igual que pasaba en la clasificación 1.

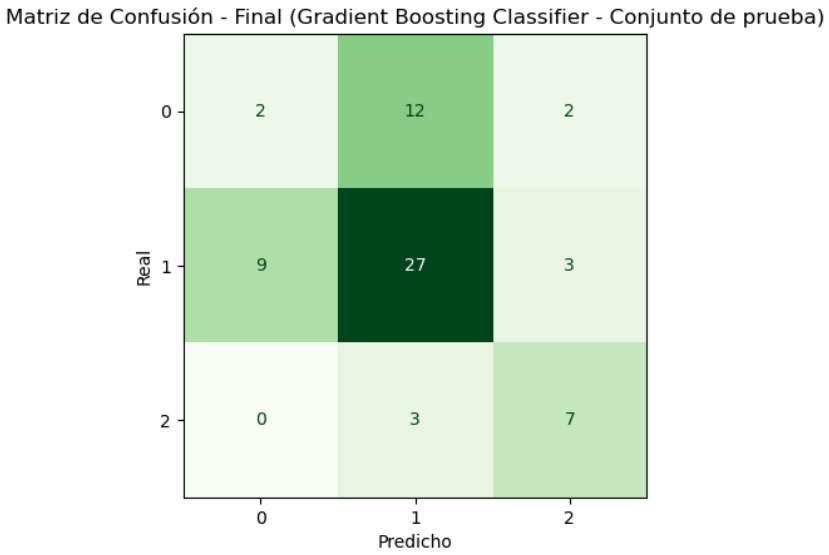

Figura 5.14: Matriz de confusión para Clasificación 2 con SMOTE

Este modelo, con un *recall* ponderado más alto que el modelo MLP de ADASYN, parece que predice peor el ISUP 1 y mejor los otros dos. Para corrobarlo, calculamos su *recall* individual:

- ISUP 1: 0.125
- ISUPs 2,3: 0.692
- ISUPs 4,5: 0.7

Aunque el *recall* ponderado no es muy alto, el modelo predice mejor las clases más agresivas que los primeros dos modelos con un *recall* ponderado aproximado de 0.7.

En un modelo de Gradient Boosting, la importancia de las variables se calcula de manera similar a como se hace en un Random Forest. Se evalúa la contribución de cada variable a la reducción del error en las predicciones del modelo. En este caso, durante el proceso de construcción de árboles de decisión, se observa cuánto mejora la precisión del modelo al incluir una variable específica en comparación con otras variables. La importancia de las variables se determina según la magnitud de esta contribución relativa.

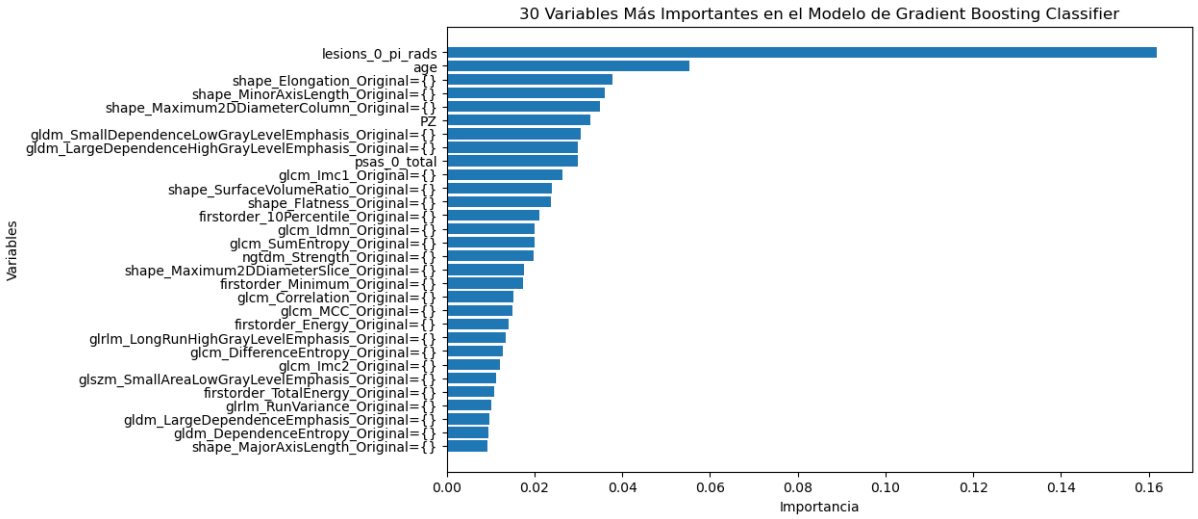

Figura 5.15: Variables más importantes para Clasificación 2 con SMOTE

Según el gráfico vemos como PI-RADS contribuye mucho más que el resto de variables.

5.3 CLASIFICACIÓN 3-1 vs 2 vs 3,4,5

No obstante, los resultados seguían sin ser los esperados así que se intentó situar la clase 3 en la categoría de alto riesgo.

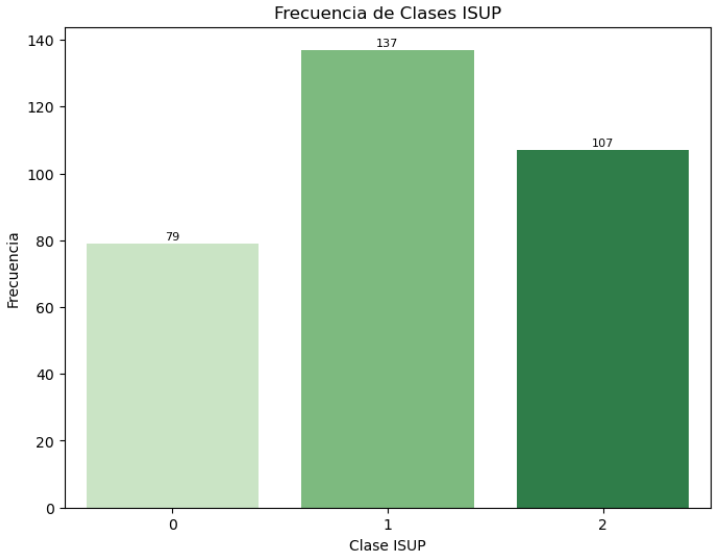

Figura 5.16: Proporción de las clases en la Clasificación 3

5.3.1. ADASYN

N (entrenamiento): 304 N (test): 65

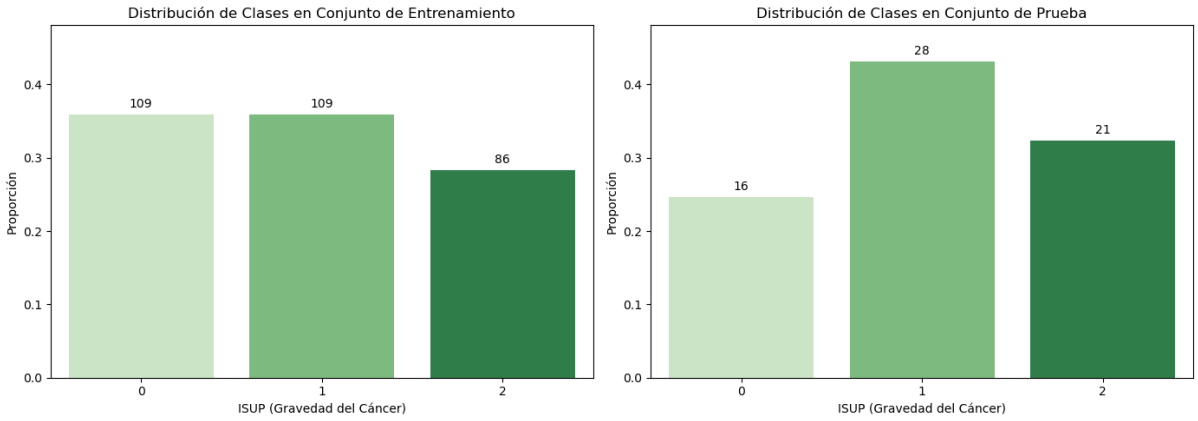

Figura 5.17: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 3 con ADASYN

Tabla 5.5: Clasificación 3: ADASYN

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.4769	0.4769	0.4786	0.4748
Logistic Regression Lasso	0.3846	0.3846	0.3866	0.3852
Ridge Classifier	0.4154	0.4154	0.4328	0.4208
SGD Classifier	0.4308	0.4308	0.4770	0.4264
Decision Tree Classifier	0.4308	0.4308	0.4261	0.4279
Random Forest Classifier	0.4308	0.4308	0.4572	0.4329
AdaBoost Classifier	0.4308	0.4308	0.4231	0.4108
Gradient Boosting Classifier	0.4769	0.4769	0.4725	0.4723
Support Vector Classifier	0.3846	0.3846	0.3571	0.3693
k-Nearest Neighbors	0.3231	0.3231	0.3575	0.3058
Gaussian Naive Bayes	0.2308	0.2308	0.3024	0.1709
Multi-layer Perceptron	0.3692	0.3692	0.3663	0.3658

Pero situar el ISUP 3 en la categoría de alto riesgo y dejar a los ISUP 1 y 2 por separado, parece que empeoró aún más el rendimiento del modelo.

Una vez más, esto puede deberse a la hipótesis formulada al principio del capítulo. Los ISUP 1 y 2 no pueden difernciarse bien entre ellas. Asimismo, el ISUP 3 es un poco ambiguo y está un poco en el limbo entre lo agresivo y lo no agresivo.

De esta manera, mientras en la clasificación 2 se confundían las clases 0 y 1, ahora se confunden las 3 clases entre ellas; y esto es por haber separado los ISUP más conflictivos (en términos de clasificación en el modelo) entre las 3 clases.

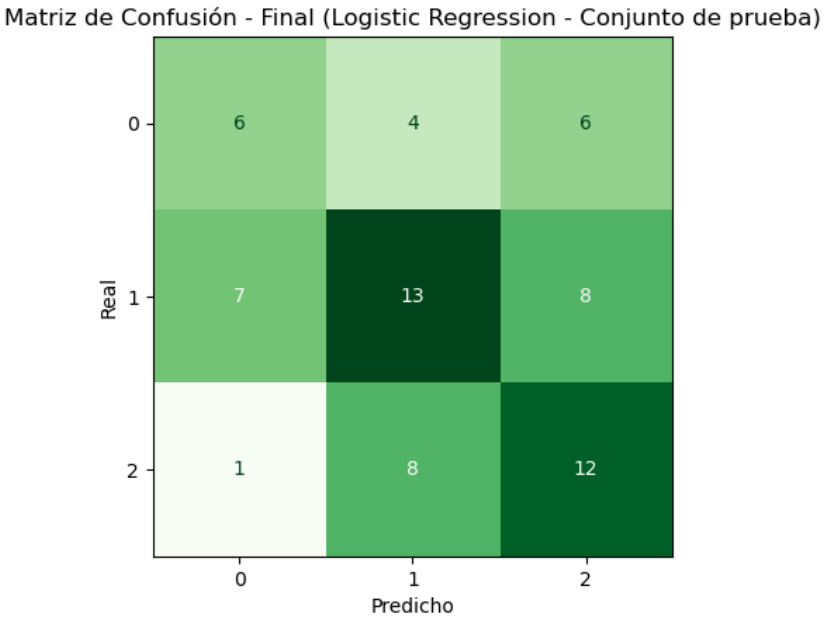

Figura 5.18: Matriz de confusión para Clasificación 3 con ADASYN

Así lo podemos ver en la matriz de confusión.

- Para la clase 0 (ISUP 1), el modelo clasifica correctamente solo 6 de los 16 casos. De los 10 casos restantes, 4 se clasifican como ISUP 2 y 6 como ISUP 3, 4, 5.
- En la clase 1 (ISUP 2), hay una mejora leve, con el modelo prediciendo correctamente 13 de los 28 casos. Los 15 casos restantes se distribuyen equitativamente entre las clases 0 y 2.
- Finalmente, para la clase 2 (ISUP 3, 4, 5), el modelo muestra el mejor *recall*, aunque aún no es suficientemente alto. Clasifica correctamente 12 de los 21 casos, mientras que los 9 casos restantes se clasifican mayoritariamente como ISUP 2. Esto puede explicarse por la ambigüedad entre ISUP 2 y 3, lo que sugiere que el modelo tiene dificultades para distinguir entre estas categorías.

En la Regresión Logística, la importancia de las variables se puede evaluar mediante la magnitud de los coeficientes de regresión asociados a cada variable en el modelo. Al igual que en el SDG, las variables con coeficientes de mayor magnitud tienen un mayor impacto en la predicción de la variable objetivo.

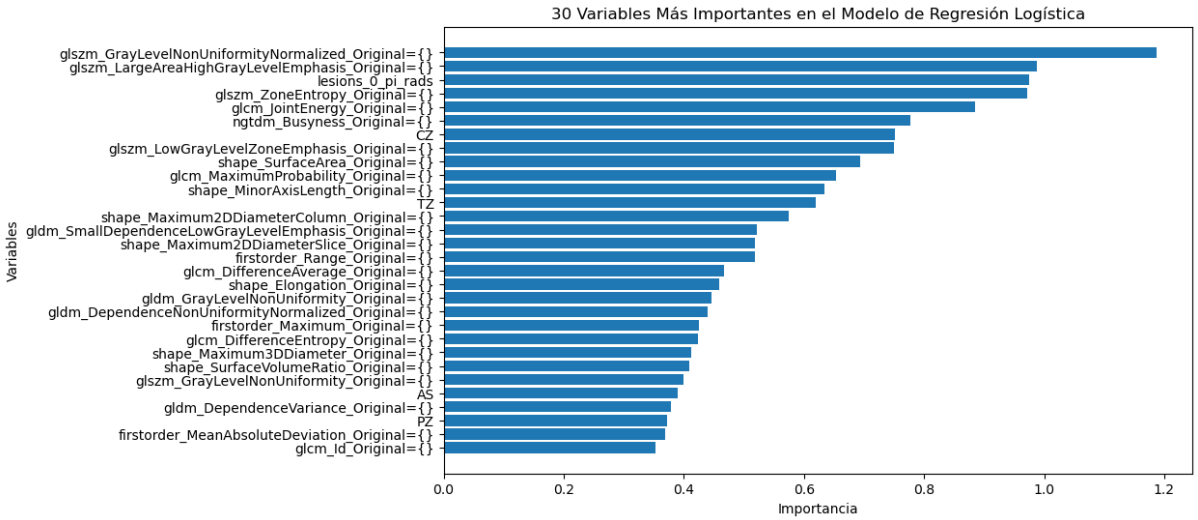

Figura 5.19: Variables más importantes para Clasificación 3 con ADASYN

Entre las 30 variables que más aportan, volvemos a encontrar PI-RADS, variables radiómicas y, en este caso, las 4 variables que indican si hay infiltración en distintas zonas de la próstata.

5.3.2. SMOTE

N (entrenamiento): 304 N (test): 65

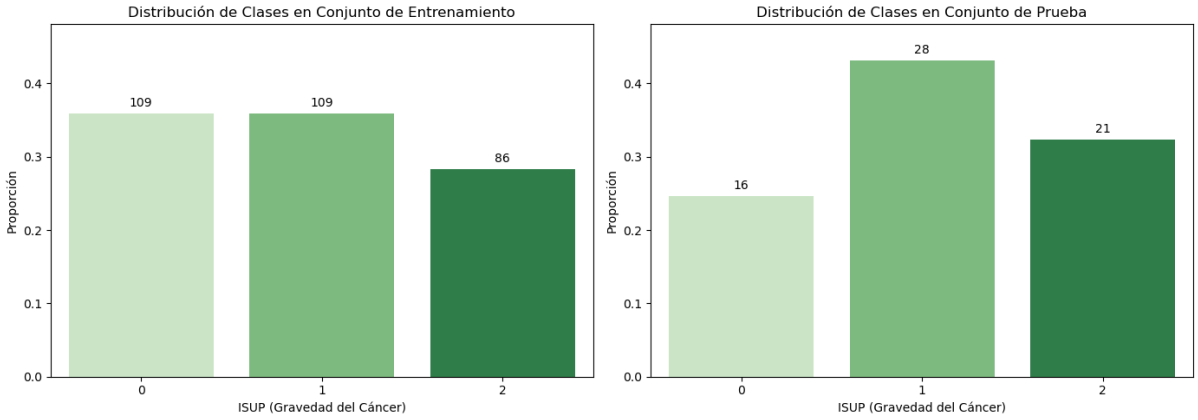

Figura 5.20: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 3 con SMOTE

Tabla 5.6: Clasificación 3: SMOTE

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.5077	0.5077	0.5142	0.5087
Logistic Regression Lasso	0.4923	0.4923	0.4991	0.4952
Ridge Classifier	0.4462	0.4462	0.4504	0.4472
SGD Classifier	0.3692	0.3692	0.3956	0.3143
Decision Tree Classifier	0.3846	0.3846	0.3832	0.3838
Random Forest Classifier	0.4769	0.4769	0.4945	0.4732
AdaBoost Classifier	0.4000	0.4000	0.2414	0.2961
Gradient Boosting Classifier	0.4308	0.4308	0.4264	0.4228
Support Vector Classifier	0.4308	0.4308	0.4162	0.4206
k-Nearest Neighbors	0.3692	0.3692	0.4158	0.3655
Gaussian Naive Bayes	0.2154	0.2154	0.2627	0.1639
Multi-layer Perceptron	0.4154	0.4154	0.4189	0.4086

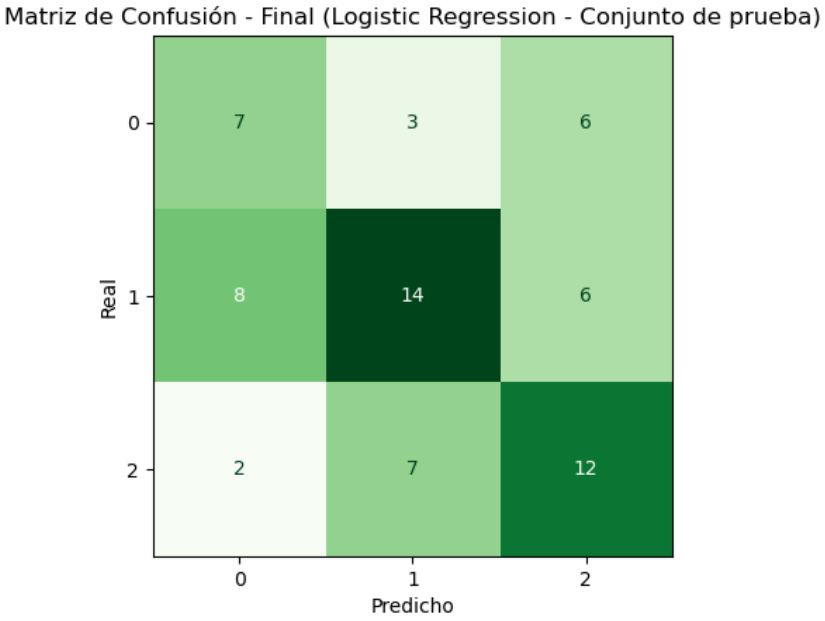

Figura 5.21: Matriz de confusión para Clasificación 3 con SMOTE

En este caso, se observa que el *recall* de la clase más agresiva se mantiene, mientras que las clases 0 y 1 mejoran un poco.

Aunque esta mejora es alentadora, aún no alcanza el nivel deseado. El modelo ideal sería aquel que predice correctamente todas las clases. Dado que esto es difícil de lograr, priorizamos la identificación precisa de las clases más agresivas. En este sentido, el modelo actual cumple con este criterio, ya que presenta el *recall* más alto para la clase 2. Sin embargo, un *recall* de 0.57 sigue sin alcanzar el umbral deseado.

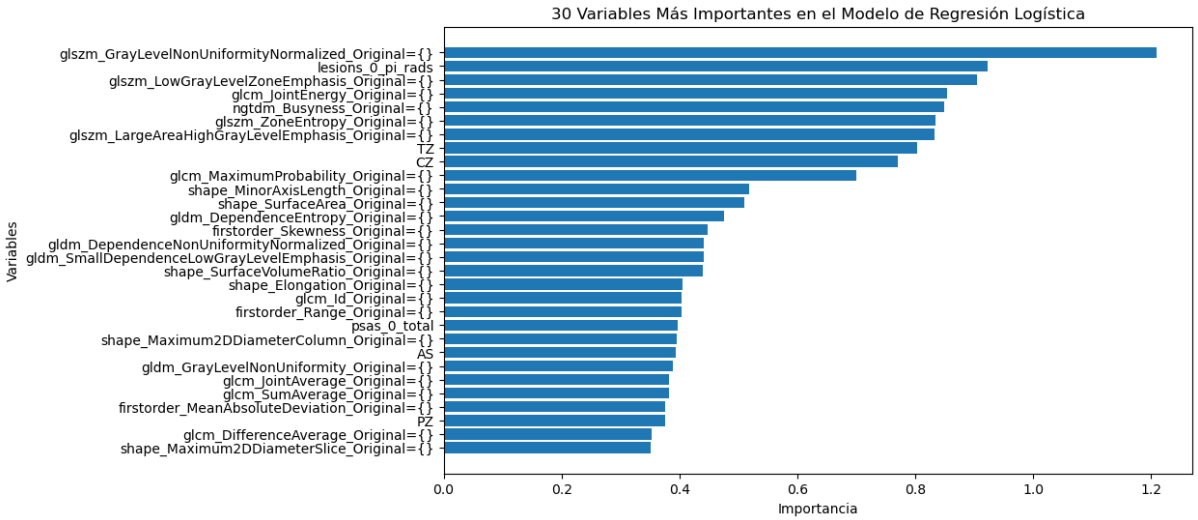

Figura 5.22: Variables más importantes para Clasificación 3 con SMOTE

A pesar de las diferencias, los gráficos de importancia se asemejan bastante entre sí.

5.4 CLASIFICACIÓN 4-1,2,3 vs 4,5

Nos dimos cuenta que al separar las clases 1 y 2, los resultados empeoraban. De esta forma volvimos a una clasificación binaria pero ahora distinguiendo entre riesgo intermedio-bajo (ISUP 1,2,3) vs riesgo alto (ISUP 4,5).

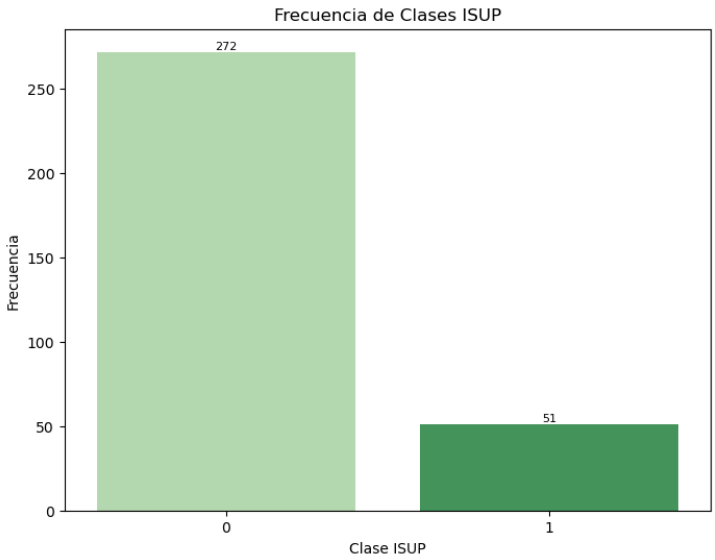

Figura 5.23: Proporción de las clases en la Clasificación 4

5.4.1. ADASYN

N (entrenamiento): 421 N (test): 65

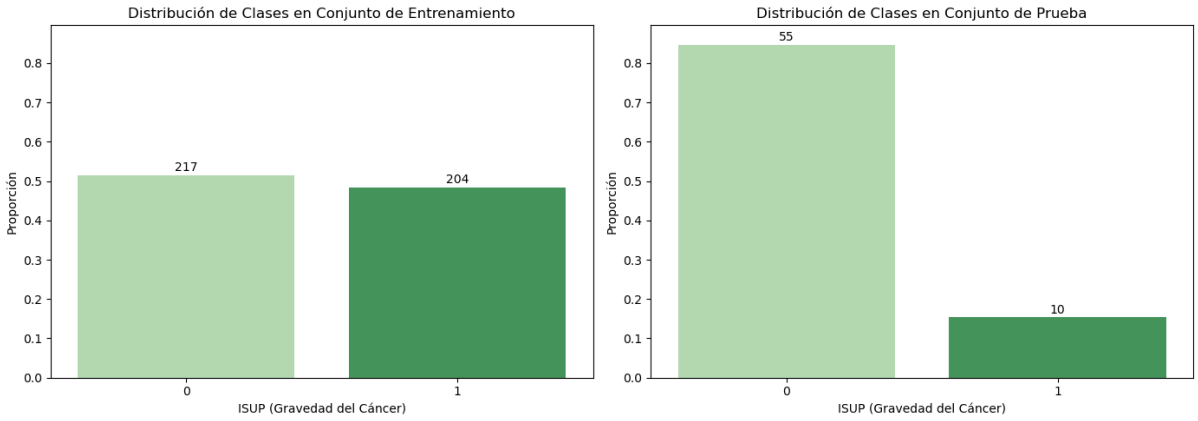

Figura 5.24: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 4 con ADASYN

Tabla 5.7: Clasificación 4: ADASYN

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.7385	0.7385	0.7678	0.7519
Logistic Regression Lasso	0.7077	0.7077	0.7585	0.7300
Ridge Classifier	0.6923	0.6923	0.7541	0.7189
SGD Classifier	0.7077	0.7077	0.7947	0.7400
Decision Tree Classifier	0.6923	0.6923	0.7153	0.7034
Random Forest Classifier	0.8308	0.8308	0.8239	0.8271
AdaBoost Classifier	0.6462	0.6462	0.7232	0.6800
Gradient Boosting Classifier	0.7385	0.7385	0.7678	0.7519
Support Vector Classifier	0.7538	0.7538	0.7318	0.7424
k-Nearest Neighbors	0.6462	0.6462	0.7601	0.6893
Gaussian Naive Bayes	0.4615	0.4615	0.8803	0.5072
Multi-layer Perceptron	0.7538	0.7538	0.7538	0.7538

Observando esta tabla, parece que el rendimiento ponderado ha mejorado. Sin embargo, es necesario estudiar individualmente las clases para comprobar si predice mejor la clase 0 o la clase 1.

Al igual que en la primera clasificación ADASYN, el mejor modelo vuelve a ser Random Forest, en este caso con un *recall* ponderado de 0.83.

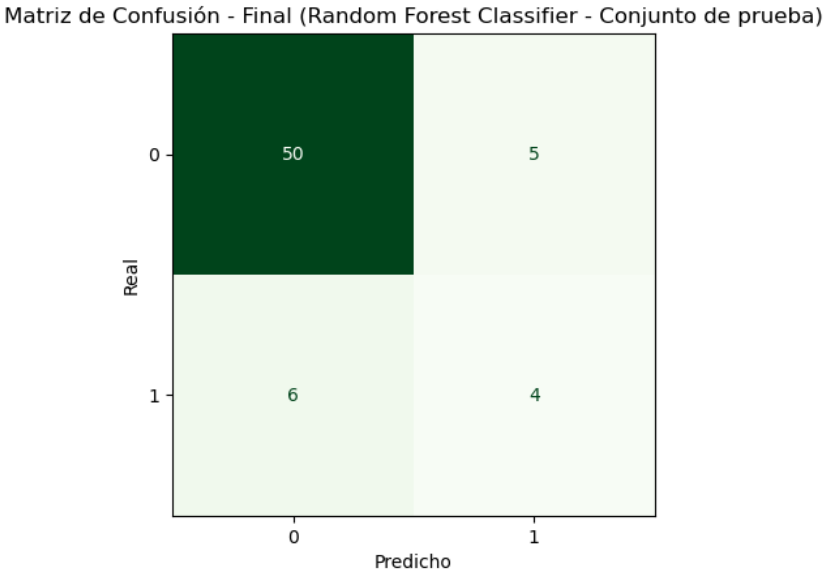

Figura 5.25: Matriz de confusión para Clasificación 4 con ADASYN

Analizando detenidamente su matriz confusión, podemos ver como el *recall* de la clase agresiva vuelve a empeorar drásticamente.

Mientras que el *recall* de la clase 0 es 0.9, el de la clase 1 es de 0.4.

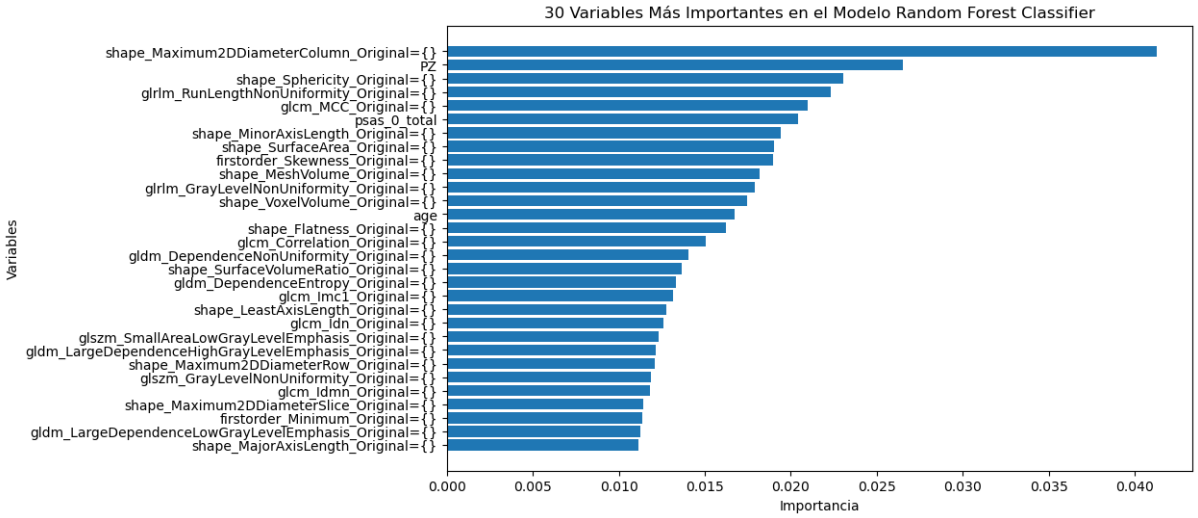

Figura 5.26: Variables más importantes para Clasificación 4 con ADASYN

El gráfico de importancia vuelve a tener un aspecto como el de la primera clasificación, con las variables PSA y edad.

Es cierto que, también se incluye la variable de infiltración de la zona periférica.

5.4.2. SMOTE

N (entrenamiento): 434 N (test): 65

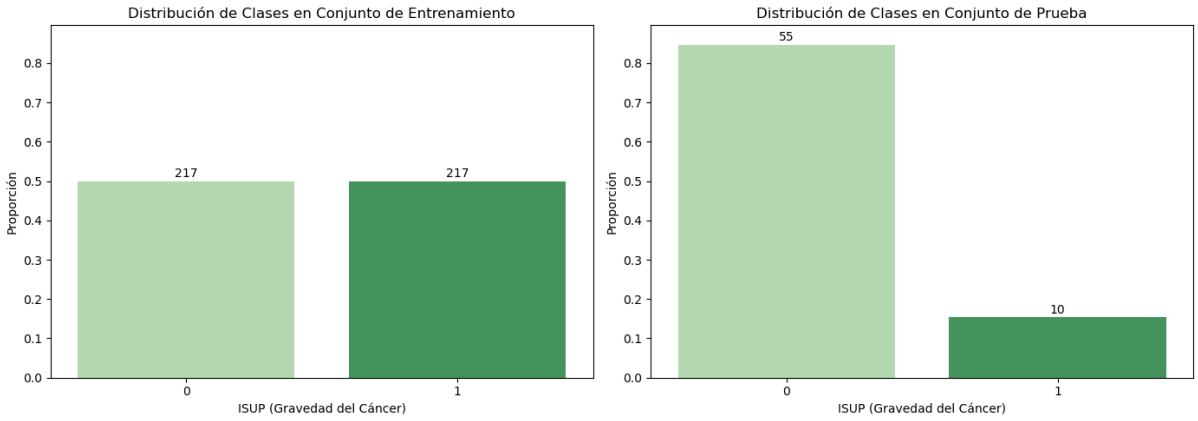

Figura 5.27: Proporción de clases para los conjuntos de entrenamiento y prueba en la Clasificación 4 con SMOTE

Tabla 5.8: Clasificación 4: SMOTE

Modelo	Accuracy	Recall	Precision	F1 Score
Logistic Regression	0.7538	0.7538	0.7729	0.7628
Logistic Regression Lasso	0.7538	0.7538	0.7729	0.7628
Ridge Classifier	0.6923	0.6923	0.7354	0.7121
SGD Classifier	0.7538	0.7538	0.7318	0.7424
Decision Tree Classifier	0.6615	0.6615	0.7458	0.6966
Random Forest Classifier	0.8154	0.8154	0.7999	0.8068
AdaBoost Classifier	0.6923	0.6923	0.7354	0.7121
Gradient Boosting Classifier	0.7538	0.7538	0.7538	0.7538
Support Vector Classifier	0.7692	0.7692	0.7368	0.7519
k-Nearest Neighbors	0.6462	0.6462	0.7601	0.6893
Gaussian Naive Bayes	0.4769	0.4769	0.8811	0.5246
Multi-layer Perceptron	0.7385	0.7385	0.7488	0.7435

Random Forest se vuelve a repetir como mejor modelo aunque con un *recall* ponderado de 0.81.

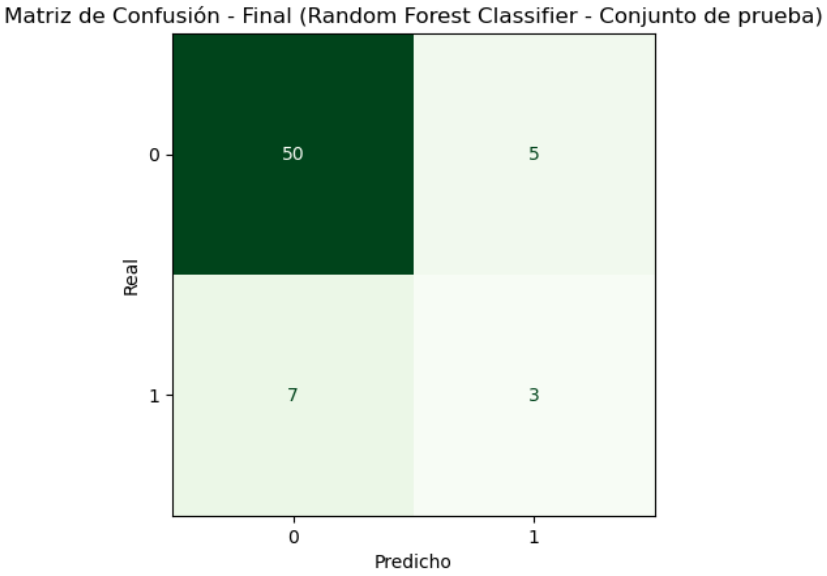

Figura 5.28: Matriz de confusión para Clasificación 4 con SMOTE

Si atendemos a la matriz, entendemos que el *recall* ponderado es un poco peor porque el *recall* de la clase 1 ha empeorado también.

Ahora el modelo solo clasifica correctamente 3 de los 10 casos de las ISUP 4,5.

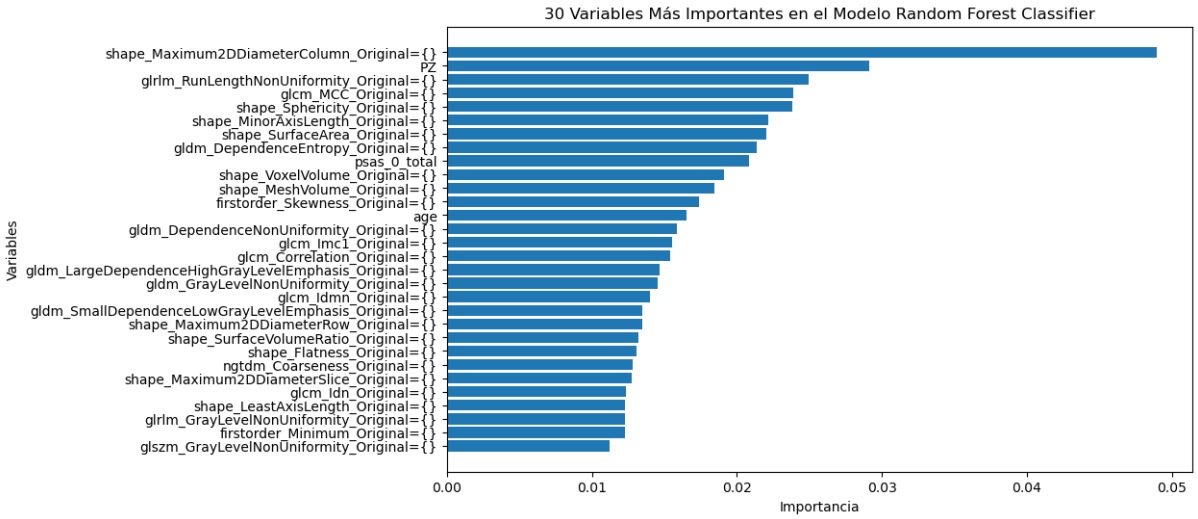

Figura 5.29: Variables más importantes para Clasificación 4 con SMOTE

Este gráfico sigue siendo muy similar al anterior.

CAPÍTULO 6
Conclusiones

6.1 Conclusiones del trabajo

Para sintetizar este Trabajo de Fin de Grado, se hará uso de los objetivos específicos del trabajo, comentando los resultados obtenidos.

Uno de los objetivos específicos fue realizar una revisión exhaustiva de la literatura sobre técnicas estadísticas multivariantes y de aprendizaje automático aplicadas a la radiómica. Aunque algunos de los estudios encontrados no estaban enfocados específicamente en el cáncer de próstata, proporcionaron un valioso marco teórico para el desarrollo del proyecto. Estas investigaciones permitieron una comprensión profunda del estado del arte en el campo de la radiómica y sentaron las bases para aplicar estas técnicas al estudio del cáncer de próstata.

Otro objetivo fue explorar y comprender la estructura y las características del conjunto de datos existente, identificando variables clave. Para ello, se realizó un análisis exploratorio de las variables clínicas. Esto nos permitió entender su relación con la variable ISUP y su importancia para la clasificación.

También se desarrollaron y entrenaron modelos de aprendizaje automático utilizando el conjunto de datos existente, con el objetivo de predecir con precisión los niveles de ISUP. Para ello, se usaron técnicas como Grid Search, validación cruzada o remuestreo.

Los resultados de este paso proporcionaron información sobre la capacidad predictiva de los modelos y su eficacia en la tarea de clasificación. De esta manera, se evaluó la efectividad y la robustez de los modelos mediante métricas de rendimiento. Estos resultados permitieron una evaluación objetiva del desempeño del modelo y su capacidad para generalizar a datos no vistos.

A continuación se exponen algunas de las observaciones a destacar sobre los resultados obtenidos:

- Aunque las clasificaciones 1 y 2 mostraron un rendimiento ponderado mayor, esto se debió a su capacidad para predecir mejor las clases menos agresivas, que son mayoritarias en el conjunto de datos.

- Las clasificaciones 2 y 3 mostraron un rendimiento ponderado ligeramente inferior, pero lograron predecir las clases más agresivas de manera más efectiva, lo cual es fundamental en el contexto clínico.

- El rendimiento ponderado de estas clasificaciones (2 y 3) también era peor por la hipótesis formulada, según la cual los ISUP 1 y 2 apenas se diferencian entre ellas. Esto dificulta su clasificación.
- También se observó que la técnica de *oversampling* SMOTE mostró, generalmente, un desempeño mejor en el manejo del desequilibrio de clases en comparación con ADASYN, lo que destaca su utilidad en futuras investigaciones y aplicaciones prácticas.

En conclusión, este trabajo representa un avance significativo en el campo del diagnóstico del cáncer de próstata, proporcionando herramientas y metodologías para mejorar la precisión en la clasificación de la enfermedad. Las conclusiones y recomendaciones derivadas de este estudio tienen el potencial de beneficiar tanto al personal médico como a los investigadores involucrados en la medicina de precisión, contribuyendo así al avance de la atención médica y la toma de decisiones clínicas informadas.

6.2 Conclusiones personales

Este trabajo de fin de grado ha sido todo un reto personal para mí. Es la primera vez que me enfrento a un proyecto de investigación de esta magnitud de forma individual. Aunque al principio me sentía un poco intimidada, ha acabado yendo mucho mejor de lo esperado.

Durante el proceso, he aprendido mucho sobre lo que implica llevar a cabo una investigación en profundidad, desde el proceso de investigación, hasta la redacción de la memoria. Me ha sorprendido descubrir mi capacidad para superar obstáculos y resolver problemas de forma serena y metódica.

Una de las cosas más importantes que he aprendido es la gestión del tiempo y la importancia de cumplir con los plazos. Aunque siempre se nos advierte sobre este aspecto al comenzar un proyecto, no fue hasta que me enfrenté a ello que comprendí su verdadero valor. He tenido que aprender a organizar mi tiempo de manera más eficiente para poder cumplir con todas las tareas.

También he aprendido la importancia del trabajo en equipo. Para llevar a cabo este proyecto, he tenido que colaborar con distintos expertos en el tema: oncólogos, radiólogos, científicos de datos y otros especialistas en el cáncer de próstata. Esto no solo mejoró mi comprensión del tema, sino que también me permitió ver cómo diferentes perspectivas pueden enriquecer el proceso de investigación y mejorar la calidad de los resultados.

En general, ha sido una experiencia mucho más positiva de lo esperado, que me ha permitido aplicar los conocimientos que he adquirido a lo largo del grado. Esto me da fuerzas a enfrentarme a nuevos proyectos que se me presenten en el futuro.

6.3 Análisis marco legal y ético

Como ya se ha comentado, el presente proyecto se basa en dos conjuntos de datos: variables radiómicas y variables clínicas, recopilados como parte del proyecto ProCancer-I, una colaboración entre varios centros clínicos europeos.

ProCancer-I obtuvo la aprobación de un comité ético y aseguró el consentimiento informado de los pacientes por parte de cada socio clínico involucrado en la recopilación de datos. Aunque los datos incluyen un identificador único para cada paciente y la edad, se han tomado medidas para proteger la privacidad y confidencialidad de los pacientes.

Estas medidas garantizan el cumplimiento de los principios éticos fundamentales, como el respeto a la autonomía y la privacidad de los pacientes, así como la protección de su confidencialidad. Esto asegura la integridad y la ética en la realización de la investigación y el desarrollo de modelos de predicción en el ámbito médico.

6.4 Relación del trabajo con los estudios cursados

A continuación, se muestran las asignaturas que han proporcionado la base teórica y práctica necesaria para llevar a cabo el trabajo. El conocimiento y habilidades adquiridas en estas materias han sido fundamentales para el desarrollo exitoso de tu proyecto:

- **Análisis exploratorio de datos**. Entender los datos con los que se cuenta es fundamental para realizar cualquier proyecto. Esta asignatura nos enseña a analizar y entender los datos con los que estamos trabajando, para obtener información útil.
- **Fundamentos de programación y Programación**. Todo el proyecto ha sido programado con Python, desde el análisis exploratorio hasta los modelos de clasificación. Estas asignaturas nos brindan esas habilidades.
- **Modelos estadísticos para la toma de decisiones I y II**. Estas asignaturas son la continuación de 'Análisis exploratorio de datos'. Nos dan las nociones básicas de algunos de los modelos usados.
- **Modelos descriptivos y predictivos I y II**. Estas asignaturas explican modelos más complejos, como SVC o MLP.
- **Visualización**. En este proyecto son fundamentales las gráficas que nos permiten entender mejor los resultados. Esta asignatura nos descubre técnicas y herramientas para visualizar datos y comunicar los resultados del análisis de manera clara y comprensible.
- **Proyecto I, II y III**: Estas asignaturas se cursan en los 3 primeros años del grado. En ellas se nos enseña a realizar un proyecto de ciencia de datos. Se realizan en grupos pero nos preparan para este moemnto y el futuro.

6.5 Legado

Se deja el siguiente enlace con todo el código del proyecto

https://github.com/rgirsan/TFG

en él se encuentra el código de preparación de los datos, el análisis exploratorio, las clasificaciones ADASYN y SMOTE y un código para mostrar las gráficas de las proporciones de los datos. Por cuestión de privacidad, no se pueden adjuntar las bases de datos y, por tanto, el código no se puede reproducir, solo revisar.

6.6 Trabajo futuro

A pesar de la implementación de técnicas avanzadas como Grid Search para la optimización de hiperparámetros, *oversampling* para tratar el desequilibrio de clases y validación cruzada para evaluar el rendimiento de los modelos, persisten oportunidades para la mejora continua del proyecto. El tiempo nos ha impedido explorar a fondo hasta y, por lo tanto, sugerimos las siguientes líneas de investigación para futuros trabajos:

6.6.1. Exploración de Métricas Adicionales

Aún haber elegido las métricas cuidadosamente, no hemos encontrado un modelo que sea lo suficientemente bueno según nuestro criterio. Por eso, podría ser buena idea explorar métricas adicionales. Aunque el *recall* es justo lo que necesitamos para priorizar falsos positivos sobre falsos negativos, igual podríamos descartar el ponderado que hemos usado hasta ahora y probar con la medida macro para dar el mismo peso a todas las clases, independientemente de su soporte. Otra opción puede ser también el *f2 score*, que penaliza más los falsos negativos que los falsos positivos.

6.6.2. Investigación sobre Métodos Avanzados de Selección de Características

Dado el volumen de variables radiómicas disponibles, sería interesante investigar técnicas más avanzadas de selección de características para identificar aquellas que contribuyen de manera significativa a la predicción de la agresividad del cáncer de próstata. Métodos como Recursive Feature Elimination (RFE), L1-based feature selection o técnicas de reducción de dimensionalidad podrían ayudar a mejorar la interpretabilidad y eficacia de los modelos.

6.6.3. Exploración de Otros Modelos

En este trabajo hemos probado 12 modelos de clasificación distintos y de distintos tipos para poder encontrar una solución óptima. Sin embargo, puede ser que probando otros se llegue a lo que buscábamos.

Referencias

AIML. (2023, October 3). *What is the difference between adaboost and gradient boost?* https://aiml.com/what-is-the-difference-between-adaboost-vs-gradient-boost/.

Asociación Española Contra el Cáncer. (2023, February 3). *El cáncer en españa, datos y estadísticas.* https://www.epdata.es/datos/cancer-espana-datos-estadisticas/289.

Asociación Nacional de Cáncer de Próstata. (s.f.). *Cáncer de próstata — ancap.* https://ancap.es/cancer-de-prostata/.

Barrios, J. I. (2022, May 11). *Evaluando los algoritmos de clasificación – juan barrios.* https://www.juanbarrios.com/evaluando-los-algoritmos-de-clasificacion/.

Becker, N. (2016, December 23). *The right way to oversample in predictive modeling.* https://beckernick.github.io/oversampling-modeling/.

Brownlee, J. (2021, January 5). *Random oversampling and undersampling for imbalanced classification.* https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/.

Etkho. (2021a, April 15). *Machine learning en hospitales.* https://www.etkho.com/machine-learning-en-hospitales-que-es-aplicaciones-y-ventajas/. (etkho)

Etkho. (2021b, February 16). *Nuevos avances tecnológicos en medicina - etkho hospital engineering.* https://www.etkho.com/nuevos-avances-tecnologicos-en-medicina/. (etkho)

Fundación Instituto Roche. (2022, July 11). *Informes anticipando radiÓmica.* https://www.institutoroche.es/recursos/publicaciones/203/Informes_Anticipando_RADIOMICA.

Gamco. (s.f.). *Qué es perceptrón multicapa - mlp | concepto y definición.* https://gamco.es/glosario/percepatron-multicapa-mlp/.

Gamez, M. J. (2022, May 24). *Objetivos y metas de desarrollo sostenible - desarrollo sostenible.* https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/.

García, A. (2023, January 5). *Selección de algoritmos de clasificación de machine learning en python.* https://panamahitek.com/seleccion-de-algoritmos-de-clasificacion-de-machine-learning-en-python/.

García Galindo, M. (2021). *Análisis de las características radiómicas de la resonancia magnética en glioblastomas y su relación con la progresión tumoral y supervivencia* .

Gonzalez, L. (2019, June 28). *Regresión logística - teoría.* https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/.

González Vilanova, A. (2019). *Métodos de machine learning en estudios biomédicos* . Universitat Politècnica de València, Valencia, España.

Han, H., Wang, W.-Y., y Mao, B.-H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. En *International conference on intelligent computing* (pp. 878–887). Springer.

He, H., Bai, Y., Garcia, E. A., y Li, S. (2008). ADASYN: adaptive synthetic sampling approach for imbalanced learning. En *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322–1328).

Higgins, R. (2020, November 11). What radiomics can do for clinical trials. *Applied Clinical Trials*. https://www.appliedclinicaltrialsonline.com/view/radiomics-what-it-is-and-what-it-can-do-for-clinical-trials.

IM Médico. (2023, September 20). *La radiómica transforma la lucha contra el cáncer.* https://www.immedicohospitalario.es/noticia/41331/la-radiomica-transforma-la-lucha-contra-el-cancer.html.

Jain, S. (2023, November 4). *Stochastic gradient descent classifier.* https://www.geeksforgeeks.org/stochastic-gradient-descent-classifier/.

Jaén Lorites, J. M. (2019). *Diseño de una aproximación basada en radiómica para una clasificación de tumores de próstata mediante análisis de texturas en imágenes de resonancia magnética* . Universitat Politècnica de València, València, València, España.

Martí Bonmatí, L., Alberich-bayarri, A., García-Martí, G., Sanz Requena, R., Pérez Castillo, C., Carot Sierra, J., y Manjón Herrera, J. (2012). Imaging biomarkers, quantitative imaging, and bioengineering. *Radiologia*, *54*(3), 269–278.

Salhab Ibáñez, N. (2021, November 16). Uso de radiómica para la predicción preoperatoria en los pacientes con adenocarcinoma ductal pancreático. *American Journal of Roentgenology*. https://cbseram.com/2021/11/16/uso-de-radiomica-para-la-prediccion-preoperatoria-en-los-pacientes-con-adenocarcinoma-ductal-pancreatico/.

Sierra, R. (s.f.). *La imagen, base de la precisión.* https://statics-diariomedico.uecdn.es/diariomedico/aniversario/la_imagen_base_de_la_precision.html.

Soteras, A. (2022, September 11). *Radiómica, la ciencia que estudia las imágenes médicas imperceptibles al ojo humano.* https://efesalud.com/radiomica-ciencia-imagenes-medicas-imperceptibles-ojo-humano/. (EFE Salud)

APÉNDICE A
Objetivos de desarrollo sostenible (ODS)

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de la Agenda 2030				
Objetivo	**Alto**	**Medio**	**Bajo**	**N/A**
1. Fin de la Pobreza				X
2. Hambre Cero				X
3. Salud y Bienestar	X			
4. Educación de Calidad				X
5. Igualdad de Género				X
6. Agua Limpia y Saneamiento				X
7. Energía Asequible y no Contaminante				X
8. Trabajo Decente y Crecimiento Económico				X
9. Industria, Innovación e Infraestructura	X			
10. Reducción de las Desigualdades			X	
11. Ciudades y Comunidades Sostenibles				X
12. Producción y Consumo Responsables				X
13. Acción por el Clima				X
14. Vida Submarina				X
15. Vida de Ecosistemas Terrestres				X
16. Paz, Justicia e Instituciones Sólidas				X
17. Alianzas para lograr los Objetivos		X		

Reflexión sobre la relación del TFG con los ODS más relacionados.

En septiembre de 2015, los líderes mundiales aprobaron una Agenda de Desarrollo Sostenible que establece 17 objetivos globales para abordar desafíos cruciales como la pobreza, la protección del medio ambiente y el fomento de la prosperidad para todos. Cada objetivo viene acompañado de metas específicas a alcanzar en los próximos 15 años, en un esfuerzo conjunto por mejorar la calidad de vida en todo el mundo. (Gamez, 2022)

Este Trabajo de Fin de Grado tiene el potencial de contribuir a la consecución de varios de los Objetivos de Desarrollo Sostenible.

A.1 Objetivo 3: Salud y Bienestar

Este objetivo se centra en garantizar una vida saludable y promover el bienestar para todos en todas las edades. Este proyecto contribuye a este objetivo al trabajar en la detección temprana y la predicción del cáncer de próstata, lo que puede mejorar el pronóstico y la calidad de vida de los pacientes.

A.2 Objetivo 9: Industria, Innovación e Infraestructura

Este objetivo busca promover la innovación y construir infraestructuras resilientes para el desarrollo sostenible. Este trabajo representa una aplicación innovadora de tecnologías de imagen médica y análisis de datos para mejorar la atención médica y el diagnóstico del cáncer de próstata.

A.3 Objetivo 10: Reducción de las Desigualdades

Al desarrollar métodos más precisos para predecir la agresividad del cáncer de próstata, se está contribuyendo a reducir las desigualdades en la atención médica relacionadas con la enfermedad. Esto se debe a que una mejor predicción de la agresividad del cáncer de próstata puede ayudar a garantizar que todos los pacientes reciban un tratamiento más adecuado y oportuno.

A.4 Objetivo 17: Alianzas para lograr los objetivos

La colaboración entre diferentes sectores, incluidos el académico, el médico y el gubernamental, es crucial para abordar desafíos complejos como el cáncer de próstata. Este trabajo puede fomentar la colaboración entre estos sectores para avanzar en la investigación y mejorar los resultados para los pacientes.

APÉNDICE B
Construcción del conjunto de datos

El siguiente apéndice se centra en explicar detenidamente como se partió de las imágenes hasta conseguir la base de datos tal y como la conocemos.

B.1 Introducción

En primer lugar, cabe destacar que el trabajo previo del cual partimos, se centra en abordar diferentes desafíos. Es por ello que definen cada uno de estos escenarios como un caso de uso, Use Case. Su objetivo es llevar a cabo tareas de clasificación y segmentación del cáncer de próstata. Sin embargo, este trabajo solo se centra en el UC2, que está enfocado en la clasificación de la agresividad del cáncer de próstata.

Es importante mencionar también el UC5, que es una clasificación de la agresividad y la calidad de vida, y el UC7, que es una clasificación de la calidad de vida, pues se nombran a lo largo de este documento.

Dicho esto, ya podemos entender cómo se obtuvo la base de datos, paso por paso.

B.2 Segmentación

El proceso de segmentar consiste en resaltar elementos en primer plano. En este caso, se buscaba identificar y delimitar la forma de la próstata en las imágenes. Para ello, se llevó a cabo una segmentación automática de toda la glándula prostática en las secuencias de imágenes T2W.

Después de generar las máscaras, que se utilizan para identificar y aislar una región de interés, estas se procesaron. Primero se seleccionó la parte más grande de la máscara para, luego, suavizar bordes y, así, mejorar la precisión.

B.3 Co-registro

Cuando no se tienen estas máscaras disponibles, se requiere realizar un proceso adicional para crearlas a partir de imágenes originales. Debido a la falta de máscaras de segmentación para las secuencias de difusión, se realizó una co-registración de las secuencias T2W (imagen

en movimiento) al espacio de las secuencias DWI (imagen fija). Esto implica ajustar o alinear dos imágenes de diferentes secuencias para que estén en el mismo espacio. De esta forma, las imágenes pueden superponerse y coincidir de manera precisa, lo que facilita el análisis y la comparación.

La máscara segmentada previamente se transformó según esta co-registración y se utilizó para extraer características radiómicas de las secuencias de difusión. Además, se aplicó un recorte central a las secuencias DWI de campo de visión amplio para facilitar la co-registración.

B.4 Evaluación de la calidad de la segmentación

Se quería analizar cómo de precisa y confiable era la delimitación. Así, se le solicitó a un radiólogo que evaluara la calidad de la segmentación de 125 volúmenes T2W y sus respectivas DWI. Se asignó un grado de tres niveles de la siguiente manera: 1 se otorgó a segmentaciones buenas / aceptables; 2 a máscaras que necesitaban correcciones menores; y 3 a aquellas que requerían correcciones mayores (menos del 50 % de la glándula cubierta). El proceso era el siguiente:

- El radiólogo califica la calidad de cada máscara.
- El radiólogo aporta observaciones que permiten una base de datos con variables que describen aspectos de las imágenes y máscaras.
- Se utilizan estos datos para entrenar modelos de regresión logística para predecir la calidad de las máscaras, y los coeficientes entrenados se analizaron para obtener información sobre las posibles causas de baja calidad.

B.5 Extracción de características radiómicas

Para ello, se llevaron a cabo varios pasos previos:

- Corrección del campo de sesgo en las secuencias T2W: consiste en eliminar irregularidades en la intensidad de las imágenes para mejorar la calidad.
- Verificación del espaciado x, y, z de cada imagen: dado que existían discrepancias, la extracción de características se llevó a cabo en 2D.
- Re-muestreo de las imágenes: implica ajustar las imágenes para que tengan una separación específica y uniforme entre los píxeles. si los píxeles varían entre imágenes, podría haber errores en las medidas.
- Normalización de intensidades: garantiza que los valores de píxeles estén en una escala coherente y comparable.
- Discretización de imágenes: simplifica la información contenida en la imagen y extrae características significativas. Agrupa los valores de intensidad en bins, de manera que los valores similares se asignen al mismo. Cada intervalo se convierte en una característica radiómica.
- Extracción de características: se extraen de la segmentación de toda la glándula. Se extraen 1223 características por cada secuencia de imágenes. Cada característica es una medida cuantitativa de alguna propiedad o aspecto de la imagen que es relevante para el análisis.

En resumen, este proceso implica ajustar y preparar las imágenes para la extracción eficaz de características radiómicas, que posteriormente se utilizarán en el análisis de datos.

B.6 Extracción de características profundas

Las características profundas son atributos específicos extraídos de las imágenes médicas mediante técnicas de aprendizaje automático. Son utilizadas para mejorar la detección, diagnóstico o análisis de enfermedades. Proporcionan información detallada y valiosa sobre las imágenes, que no sería fácilmente accesible mediante métodos convencionales.

B.7 Características clínicas

Se recopilaron datos clínicos para cada caso de uso que se utilizarán en conjunto con el resto de características para realizar predicciones y análisis más completos.

B.8 Construcción del *dataset*

Se dividió el conjunto en entrenamiento y prueba para entrenar y evaluar los modelos. Esta división se hizo a nivel de paciente, lo que quiere decir que se organizó y se separó el conjunto de datos basado en los pacientes en lugar de a nivel de las imágenes. Todas las imágenes de un paciente cuentan como una unidad y se mantienen juntas.

Además, los datos se estratificaron para asegurarse que tanto entrenamiento como prueba tengan una distribución equitativa de pacientes con diferentes características o condiciones médicas. Esto evita que se concentren demasiados pacientes con características en un conjunto y pocos en el otro, lo que podría sesgar los resultados.

También se probaron diferentes subconjuntos de datos para su capacidad de entrenamiento. Las diferentes combinaciones resultaron en 96 modelos entrenados para UC2 y 5.

B.9 Pipeline de preprocesamiento

Por pipeline nos referimos al flujo de pasos que se realizan de manera ordenada para procesar los datos antes de utilizarlos en un modelo. En este caso, los pasos se llevaron a cabo en el conjunto de entrenamiento y solo en las variables numéricas:

- Escalamiento de características: media de cero y desviación típico. Permite la comparación y facilita el análisis.
- Eliminación de características con baja varianza: se utilizó un umbral de 0.01. Se eliminaron las características cuya varianza estaba por debajo de él, pues las características con baja varianza tienen valores que no cambian mucho entre las observaciones y, por tanto, no aportan información significativa.
- Evaluación de correlación: si dos características están altamente correlacionadas, pueden contener información redundante. Dos características son consideradas altamente correlacionadas si su correlación de Spearman es mayor que 0.8. De estas dos, se eliminó la

que tenía una correlación promedio más alta con el resto de características. Esto evita redundancia y reduce la complejidad del modelo.

B.10 Entrenamiento

- Para los modelos de radiómica y raddeep, se utilizó un *light gradient boosting machine* para entrenar los modelos. El LGBM es un algoritmo utilizado para problemas de clasificación y regresión. También se entrenó un clasificaor SVM. Los modelos raddeep son los que combinan radiomicas con caracteristicas profundas.
- Para modelos radclin o híbridos, dado que pueden incluir datos categóricos, se utilizó un algoritmo CatBoost, que es adecuado para manejar eso datos. En el caso de los modelos radclin se combinan radiomicas con clínicas.

Además, se realizó una sintonización de hiperparámetros para cada algoritmo y cada combinación de parámetros se evaluó mediante validación cruzada de 5 pliegues. Para UC2 y 5, esto se realizó con una búsqueda exhaustiva en cuadrícula, mientras que para UC 7b se seleccionó un enfoque de búsqueda aleatoria, ya que hay menos datos disponibles y se prefiere una optimización menos sesgada.

B.11 Postprocesamiento

Todos los modelos finales se analizaron en dos áreas principales:

- Explicabilidad: con ella se trata de comprender cómo el modelo toma decisiones y por qué hace ciertas predicciones. Para ello, se utilizó un análisis de 'SHapley Additive exPlanations (SHAP)' para identificar las variables más relevantes para la predicción en el conjunto de prueba. Se mostraron las 20 variables más relevantes para la salida de cada modelo.
- Equidad: se refiere a la igualdad y justicia en el trato de diferentes grupos o subgrupos de datos por parte del modelo. Por eso se probó el rendimiento del modelo para diferentes subgrupos de datos. Se informa sobre ROC-AUC, puntuación f2, precisión y *recall* para cada subgrupo, así como el tamaño del subgrupo en los conjuntos de entrenamiento y prueba y la distribución de etiquetas en el conjunto de prueba. Para los subgrupos donde solo está presente una etiqueta de destino, la métrica ROC-AUC se reemplaza con la precisión.

B.12 Resultados

Así pues, el conjunto de datos final se compone de 5474 pacientes que cumplen con los requisitos para el caso de uso de agresividad de la enfermedad, UC2. De estos, 814 pacientes también son adecuados para el UC5, el caso de uso de recurrencia bioquímica, y 272 pacientes para el UC 7b, en relación con la calidad de vida después de la prostatectomía.

Sin embargo, dado que la segmentación se hace manualmente por los radiólogos, en nuestro caso solo contamos con 419 observaciones radiómicas, ya que solo se han segmentado las lesiones en 419 casos. A esto le añadimos 4646 observaciones clínicas.

En la explicación adicional, se concluye con 5474 pacientes que cumplen los requisitos para ser parte de UC2. Sin embargo, en las bases de datos proporcionadas, contamos con 4646 observaciones clínicas y 420 observaciones radiómicas. Es importante señalar que en el proyecto solo tenemos 420 observaciones radiómicas, ya que solo se han segmentado las lesiones (los tumores) en 420 casos de resonancia magnética (la segmentación se realiza manualmente en todos los casos por radiólogos de diversos centros clínicos). Estos casos, nuevamente, pertenecen a diferentes centros y hospitales europeos, no exclusivamente a La Fe.

APÉNDICE C
Análisis exploratorio de datos

C.1 Edad (age)

En primer lugar, estudiamos brevemente la edad. Si nos fijamos en el gráfico y las estadísticas descriptivas, podemos apreciar que la distribución de las edades tiene una tendencia central alrededor de la media y la mediana, con una ligera asimetría negativa, lo que significa que hay una tendencia hacia las edades más jóvenes. Además, aunque el promedio de edades en el estudio es 65 años, la desviación estándar sugiere que las edades tienden a desviarse, en promedio, alrededor de 7.88 años de la media. Esto indica una cierta variabilidad en las edades de los pacientes en nuestro estudio. Los cuartiles también destacan esta variabilidad.

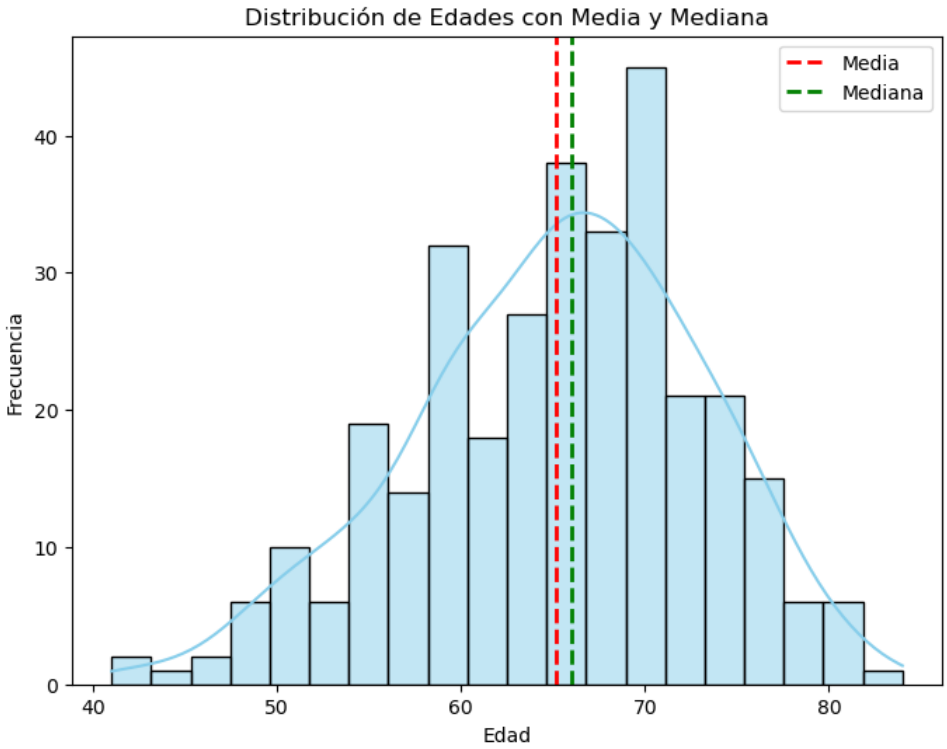

Figura C.1: Distribución de Edades

Tabla C.1: Estadísticas de la edad

Medida	**Valor**
Media	65.14
Mediana	66.0
Coeficiente de Asimetría	-0.37
Moda	65 (con frecuencia 24 veces)
Desviación Típica	7.88
Varianza	62.14
Cuartiles (25 %, 50 %, 75 %)	[60. 66. 71.]

C.2 PSA (psas_0_total)

Esta variable mide el nivel de PSA en sangre, un PSA alto puede indicar problemas en la próstata. Para ello, usamos un gráfico de caja y bigotes. En un primer intento, vemos que tiene muchos valores atípicos, lo que dificulta la visualización. Así pues, vamos a usar un percentil 95 para truncar estos valores atípicos y poder hacer una interpretación más sencilla. En primer lugar, nos fijamos en la mediana, que está alrededor de 7. Dado que está un poco posicionada hacia la izquierda, podemos decir que la distribución es asimétrica positiva. La longitud de la caja informa sobre la dispersión intercuartílica, que va desde 5 hasta 12, esto quiere decir que el 50 % de los pacientes tienen un nivel de PSA entre esos valores.En cuanto a los valores mínimo y máximo, vemos que estos están alrededor de 0 y 21. Esto quiere decir que el bigote derecho se extiende hasta 21, lo que implica una mayor dispersión. Finalmente, apreciamos bastantes valores a partir del 21, estos son considerados atípicos, pues son considerablemente mayores que la mayoría de los datos.

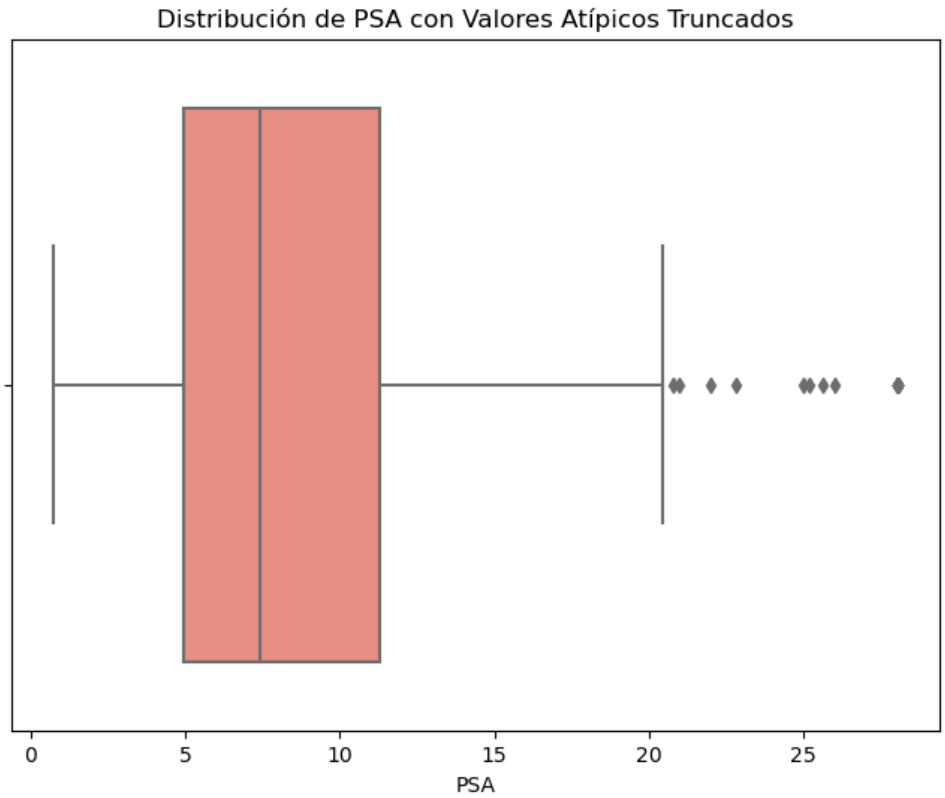

Figura C.2: Distribución de PSA

Tabla C.2: Estadísticas del PSA

Medida	**Valor**
Count	323
Mean	11.269524
Standard Deviation	17.911108
Minimum	0.73
25th Percentile	4.965
50th Percentile (Median)	7.4
75th Percentile	11.269524
Maximum	226

C.3 PI-RADS (lesions_0_pi_rads)

Esta variable es un sitema de puntuación que evalua la probabilidad de que una lesión en la próstata sea cancerosa. Así pues, se ha querido mostrar la frecuencia de cada una de las categorías. De esta forma, podemos apreciar que la mayoría de pacientes tienen una alta probabilidad de que la lesión sea cancerosa, pues la mayoría de valores comprenden entre 4-5.

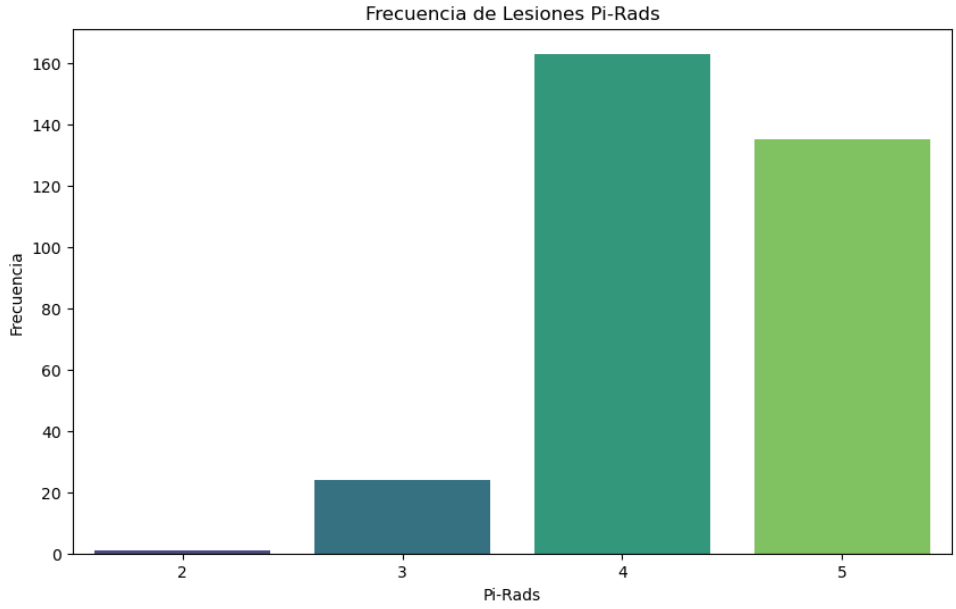

Figura C.3: Frecuencia de la variable PI-RADS

C.4 Gleason (lesions_0_gleason1.1 , lesions_0_gleason2.1)

El *Gleason score* es una puntuación que se asigna después de examinar las muestras de tejido de la próstata en una biopsia. Evalua la agresividad del cáncer y consiste en dos valores. Estas puntuaciones reflejan el grado de anormalidad del tejido observado en la biopsia. Al representar las frecuencias de cada uno de los valores, nos damos cuenta que la gran mayoría de los pacientes presentan tejidos anormales.

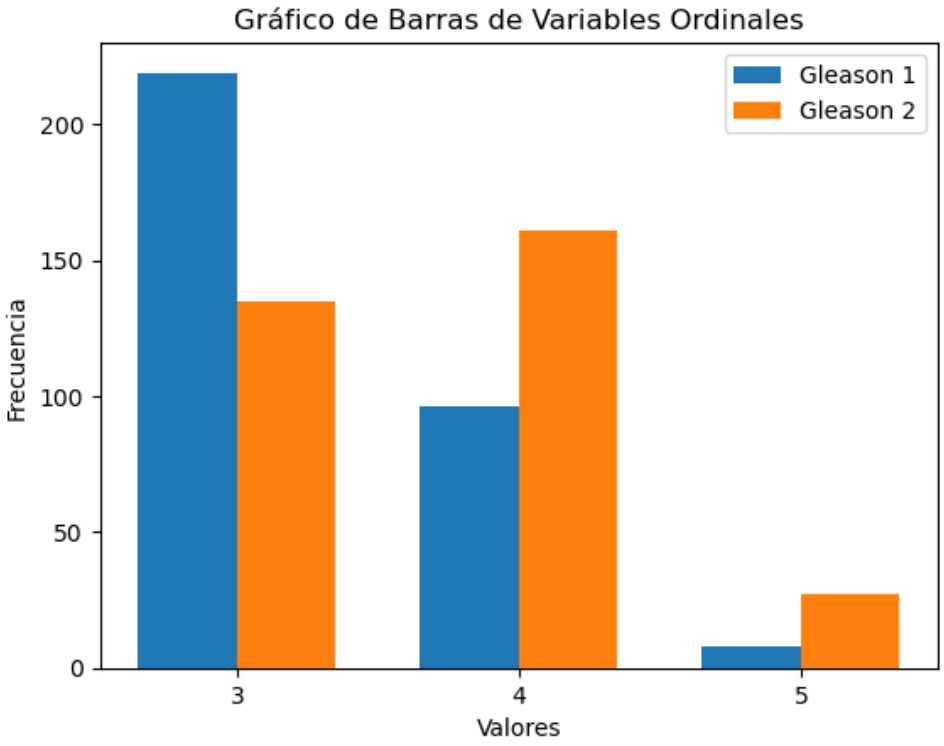

Figura C.4: Frecuencia de los Gleason

Tabla C.3: Frecuencia de lesiones para gleason1.1 y gleason2.1

	lesions_0_gleason1.1	lesions_0_gleason2.1
3	219	135
4	96	161
5	8	27

C.5 ISUP

Esta variable clasifica el grado de severidad del cáncer. Se calcula a partir de la suma de Gleason primario + Gleason secundario. Tal y como podemos ver en el gráfico, la mayor frecuencia se presenta en el 1 y 2, con lo que la mayoría de pacientes no tienen un cáncer muy severo.

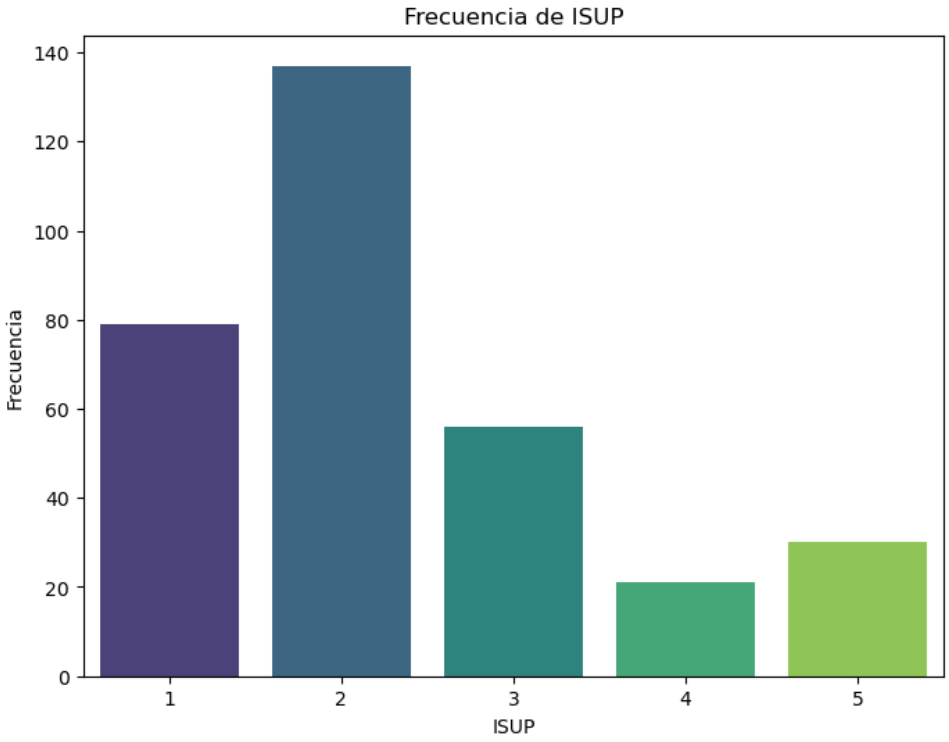

Figura C.5: Frecuencia de ISUP

C.6 TZ, PZ, CZ, AZ

Estas cuatro variables indican si hay infiltración en diferentes zonas de la próstata, concretamente en la zona de transición, en la zona periférica, la central y la anterior. El gráfico muestra que la zona más afectada es la periférica.

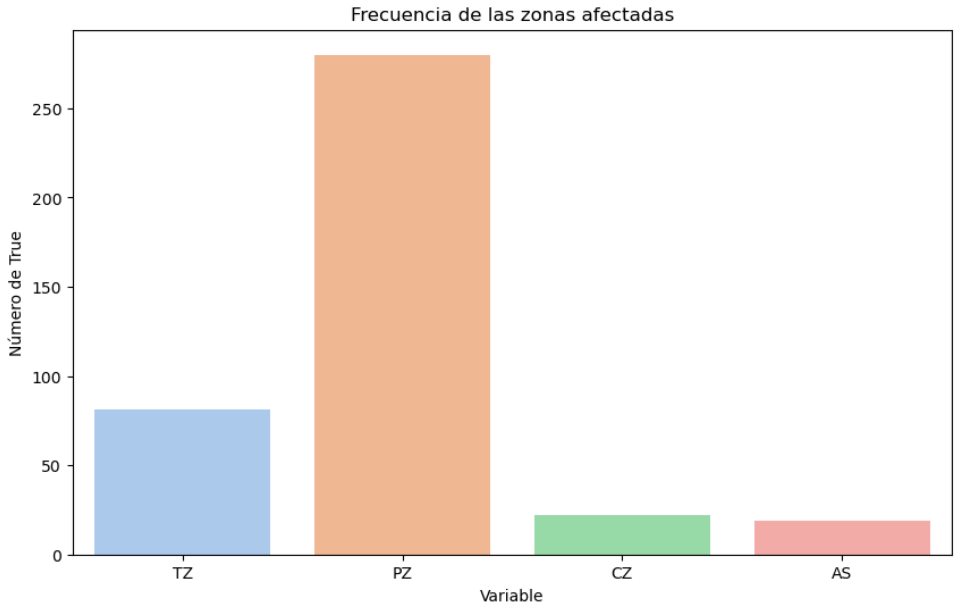

Figura C.6: Frecuencia de las zonas afectadas

C.7 Correlaciones

Como era de esperar, las variable ISUP está muy correlacionada con las variables lesions_0_gleason1.1 y lesions_0_gleason2.1. También podemos apreciar una correlación negativa entre las zonas de infiltración TZ y PZ.

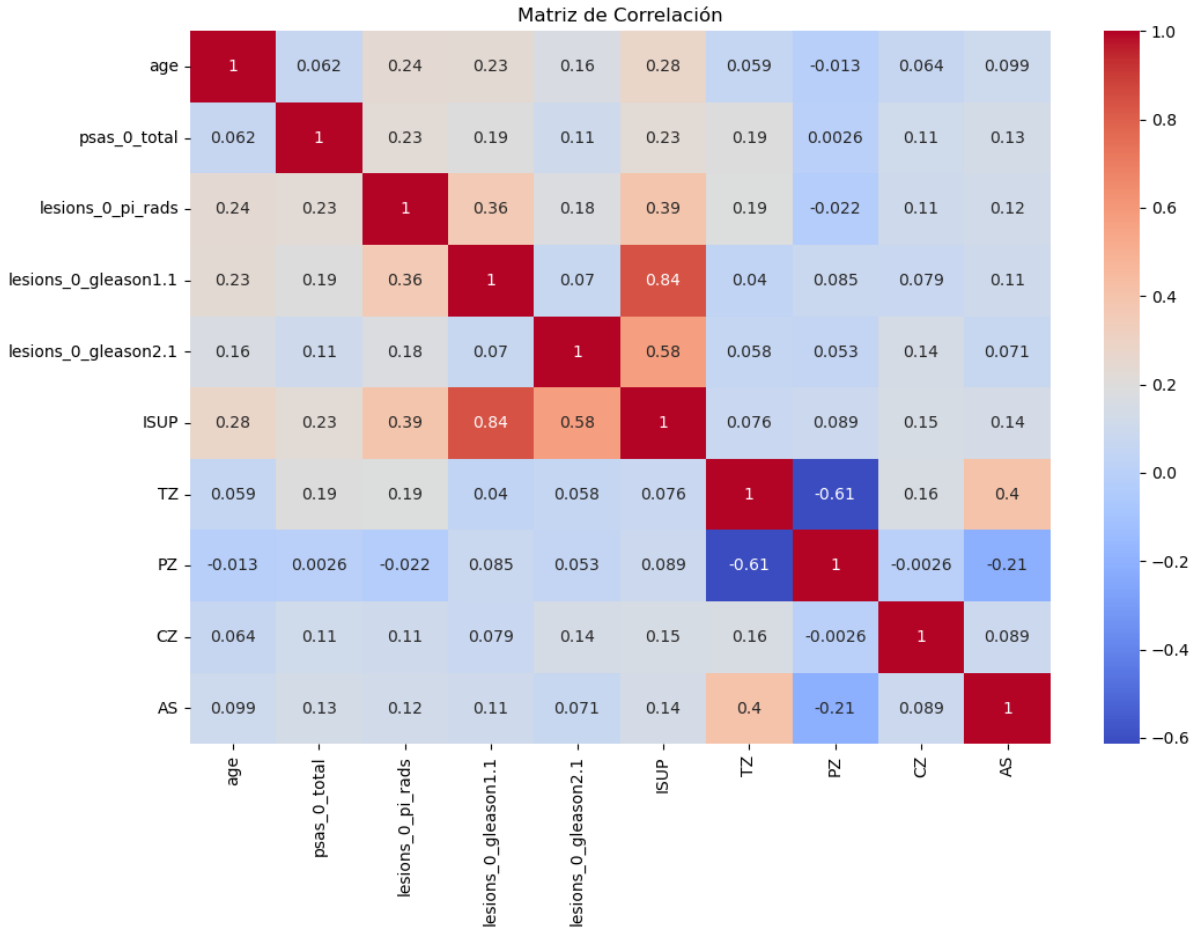

Figura C.7: Matriz de correlación

APÉNDICE D
Clasificación

D.1 Clasificación 1

D.1.1. ADASYN

Logistic Regression

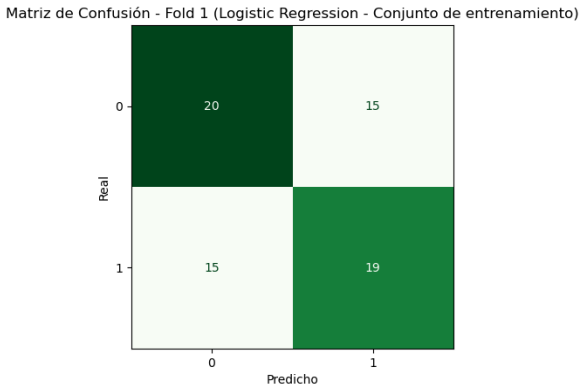

Figura D.1: Logistic Regression, Fold 1, C1, ADASYN

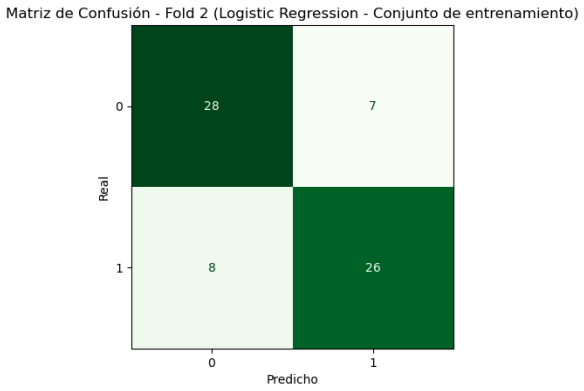

Figura D.2: Logistic Regression, Fold 2, C1, ADASYN

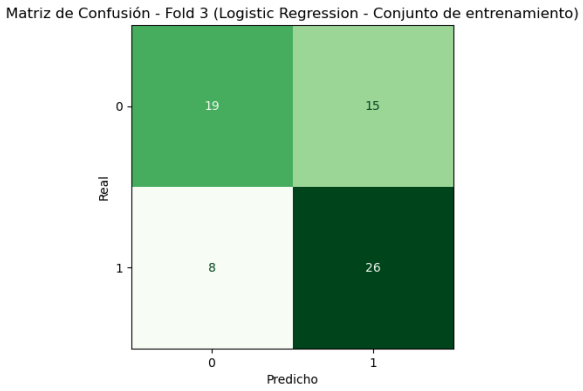

Figura D.2: Logistic Regression, Fold 3, C1, ADASYN

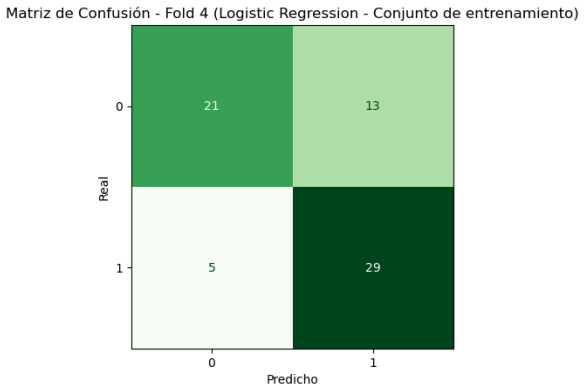

Figura D.3: Logistic Regression, Fold 4, C1, ADASYN

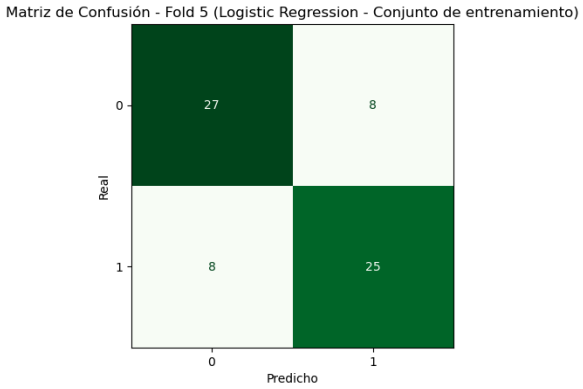

Figura D.3: Logistic Regression, Fold 5, C1, ADASYN

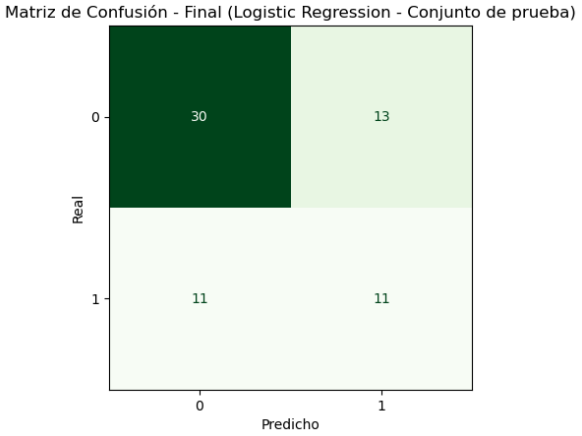

Figura D.4: Logistic Regression, Final, C1, ADASYN

Logistic Regression Lasso

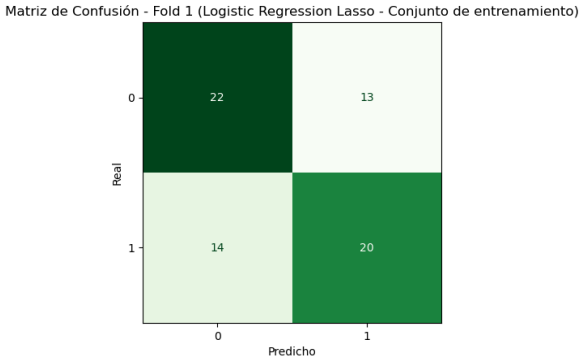

Figura D.5: Logistic Regression Lasso, Fold 1, C1, ADASYN

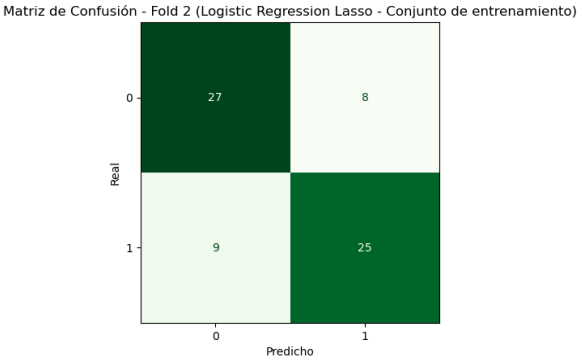

Figura D.6: Logistic Regression Lasso, Fold 2, C1, ADASYN

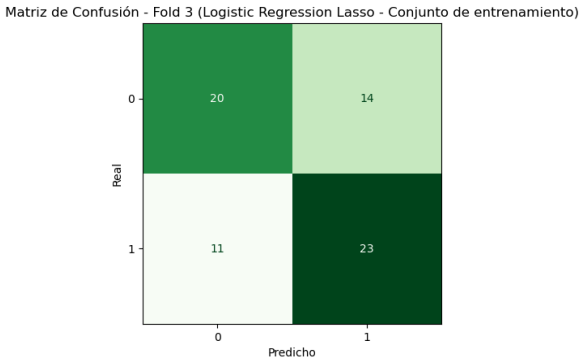

Figura D.6: Logistic Regression Lasso, Fold 3, C1, ADASYN

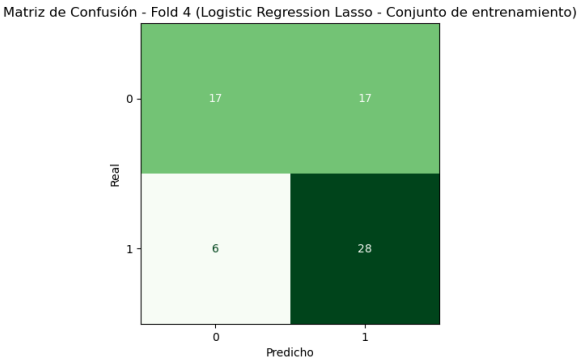

Figura D.7: Logistic Regression Lasso, Fold 4, C1, ADASYN

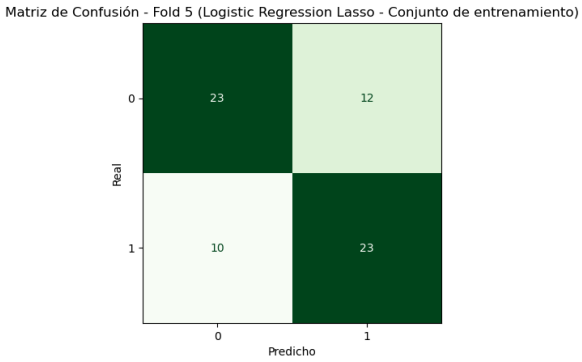

Figura D.7: Logistic Regression Lasso, Fold 5, C1, ADASYN

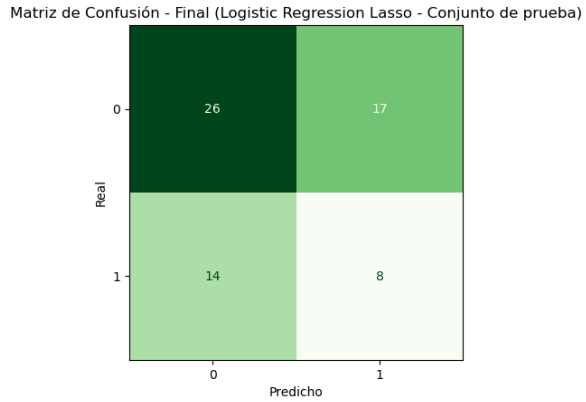

Figura D.8: Logistic Regression Lasso, Final, C1, ADASYN

Ridge

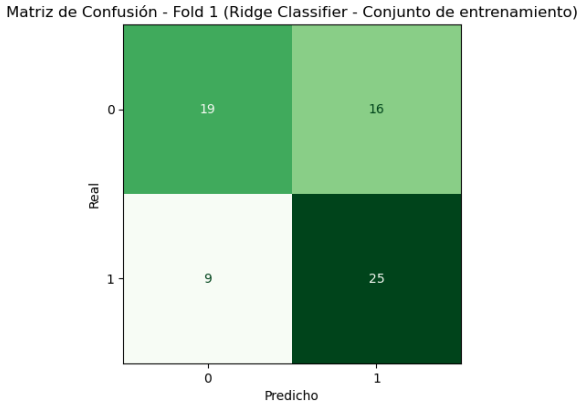

Figura D.9: Ridge, Fold 1, C1, ADASYN

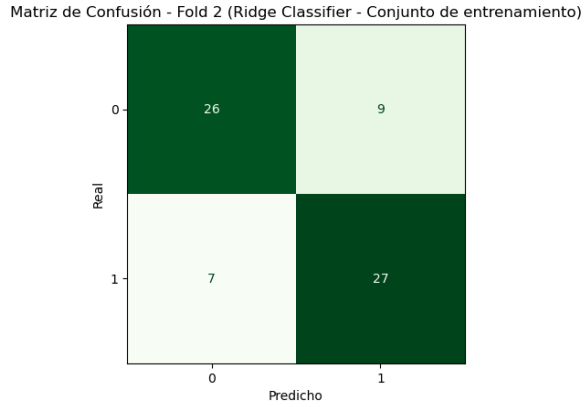

Figura D.10: Ridge, Fold 2, C1, ADASYN

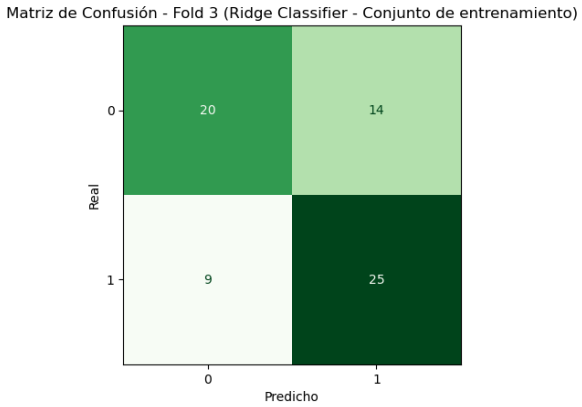

Figura D.10: Ridge, Fold 3, C1, ADASYN

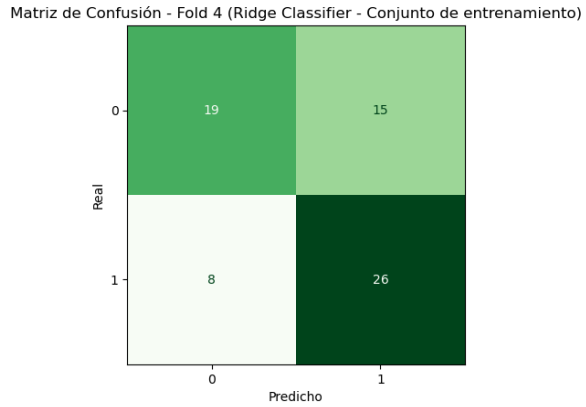

Figura D.11: Ridge, Fold 4, C1, ADASYN

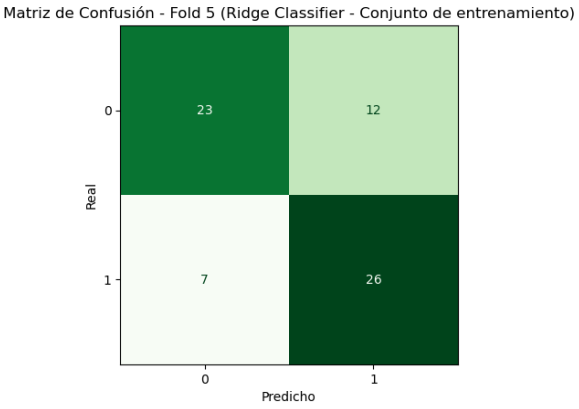

Figura D.11: Ridge, Fold 5, C1, ADASYN

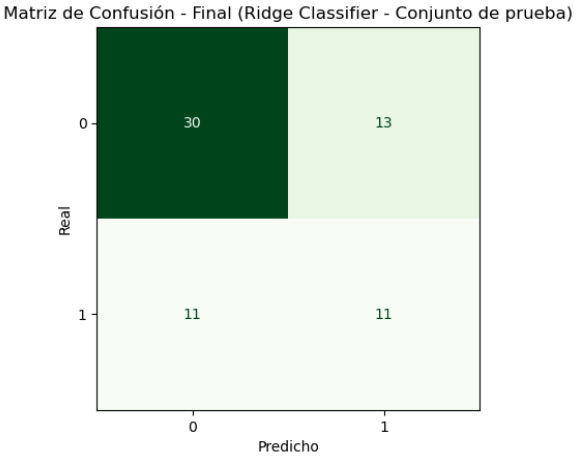

Figura D.12: Ridge, Final, C1, ADASYN

SDG

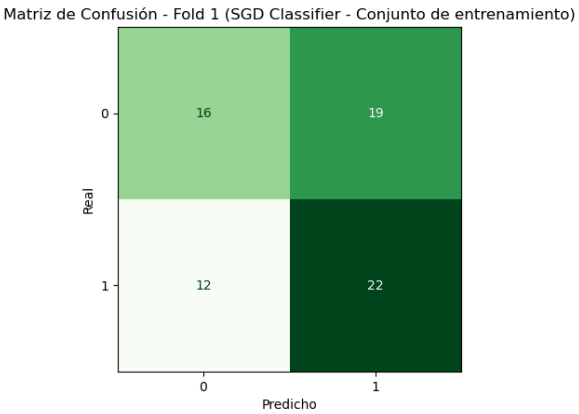

Figura D.13: SDG, Fold 1, C1, ADASYN

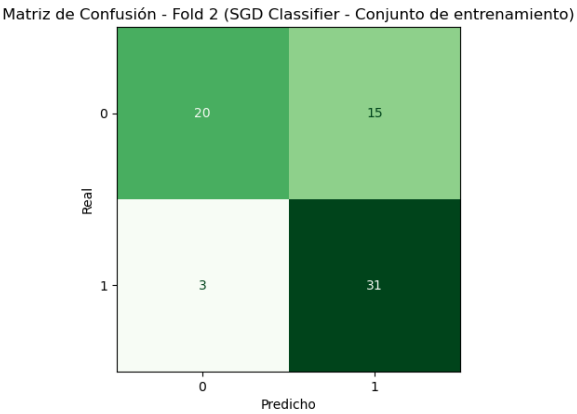

Figura D.14: SDG, Fold 2, C1, ADASYN

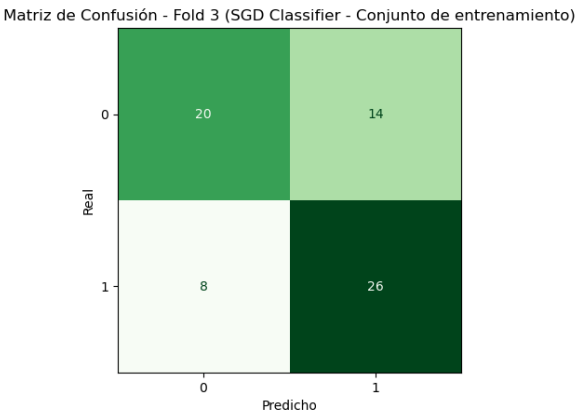

Figura D.14: SDG, Fold 3, C1, ADASYN

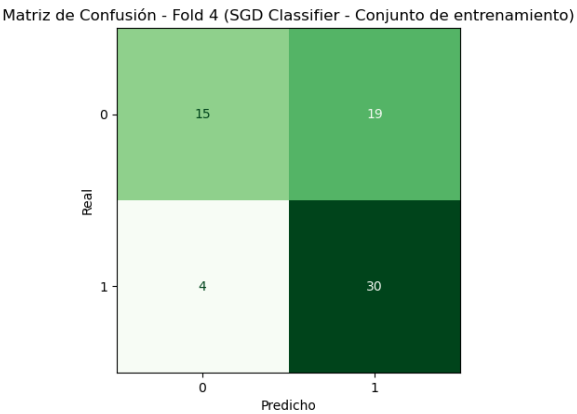

Figura D.15: SDG, Fold 4, C1, ADASYN

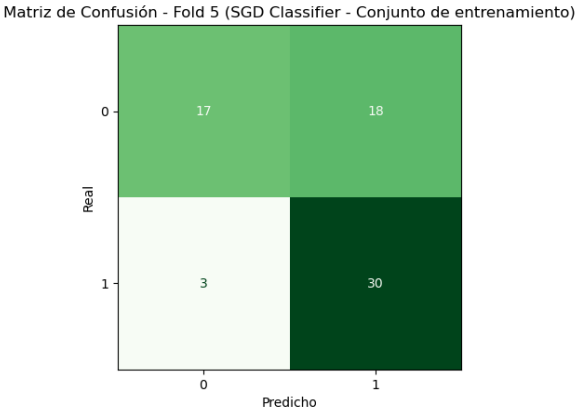

Figura D.15: SDG, Fold 5, C1, ADASYN

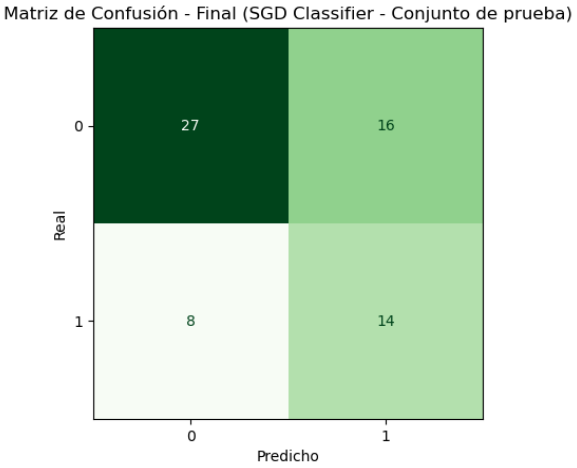

Figura D.16: SDG, Final, C1, ADASYN

Decision Tree

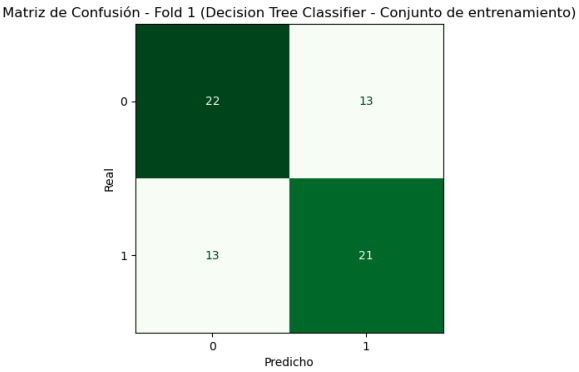

Figura D.17: Decision Tree, Fold 1, C1, ADASYN

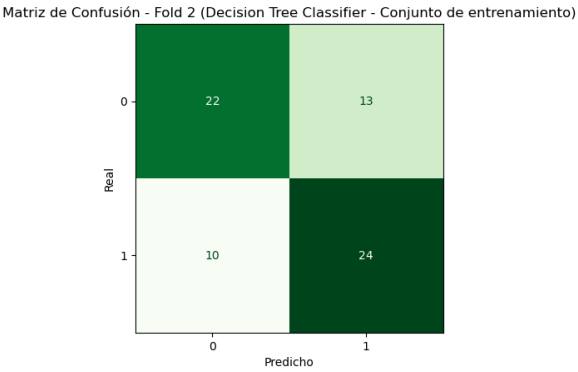

Figura D.18: Decision Tree, Fold 2, C1, ADASYN

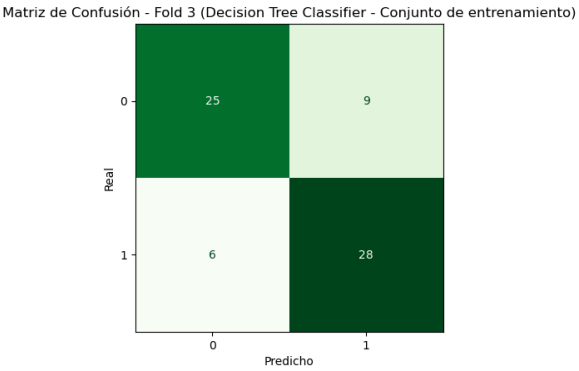

Figura D.18: Decision Tree, Fold 3, C1, ADASYN

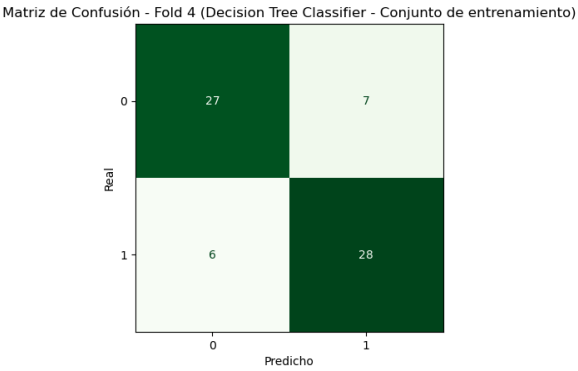

Figura D.19: Decision Tree, Fold 4, C1, ADASYN

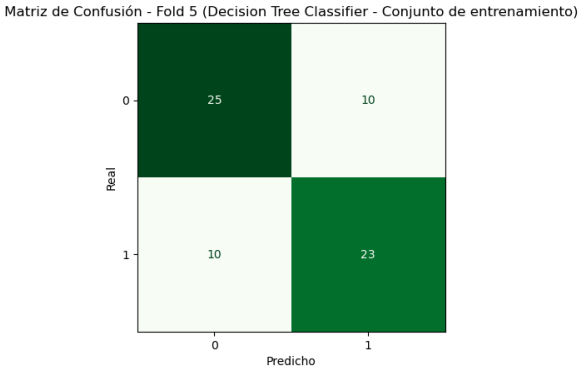

Figura D.19: Decision Tree, Fold 5, C1, ADASYN

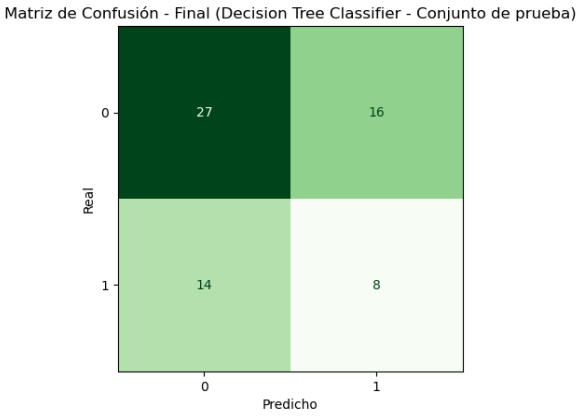

Figura D.20: Decision Tree, Final, C1, ADASYN

Random Forest

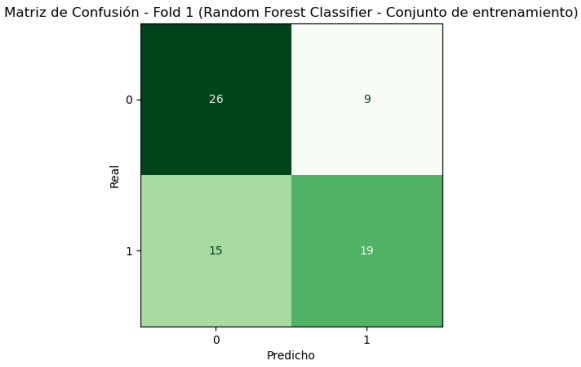

Figura D.21: Random Forest, Fold 1, C1, ADASYN

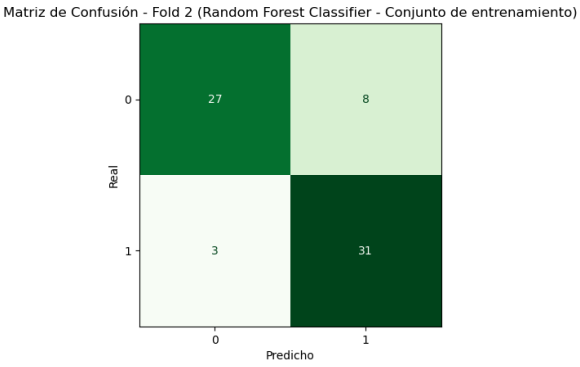

Figura D.22: Random Forest, Fold 2, C1, ADASYN

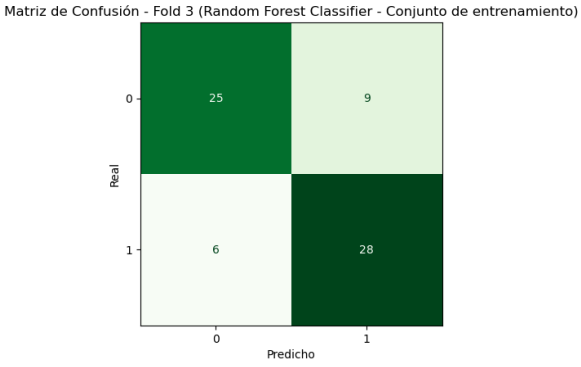

Figura D.22: Random Forest, Fold 3, C1, ADASYN

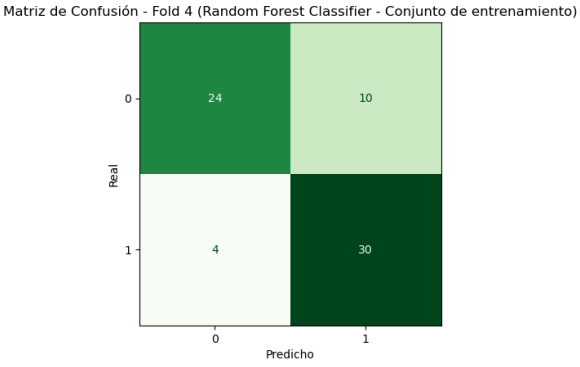

Figura D.23: Random Forest, Fold 4, C1, ADASYN

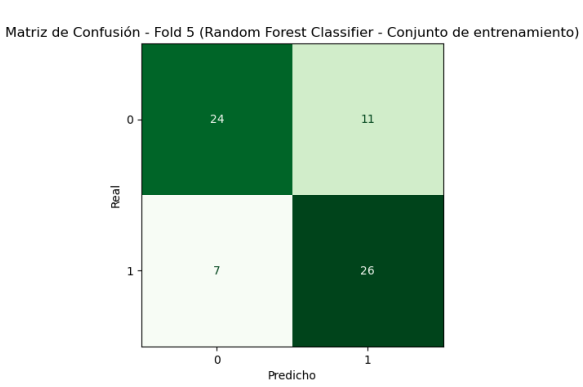

Figura D.23: Random Forest, Fold 5, C1, ADASYN

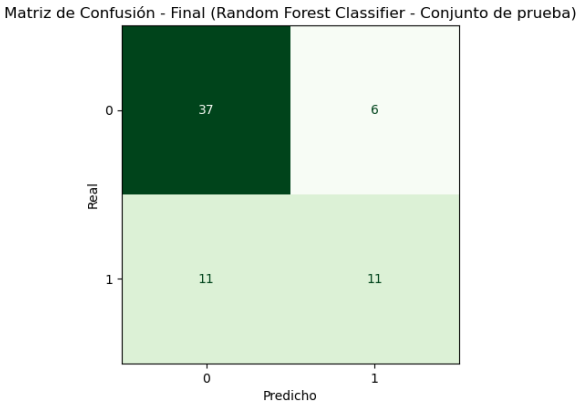

Figura D.24: Random Forest, Final, C1, ADASYN

AdaBoost

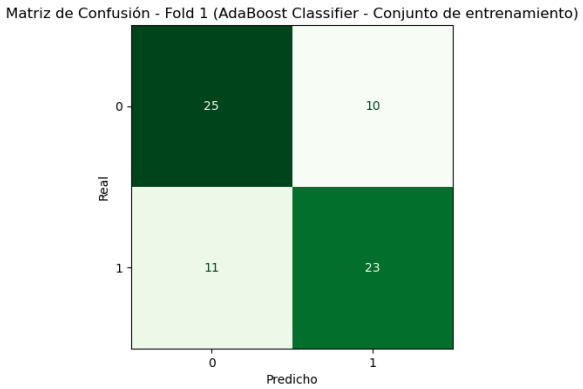

Figura D.25: AdaBoost, Fold 1, C1, ADASYN

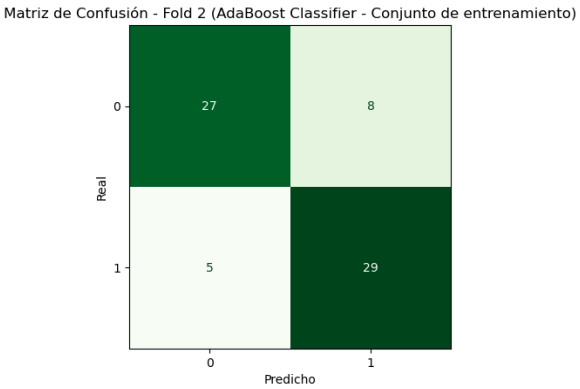

Figura D.26: AdaBoost, Fold 2, C1, ADASYN

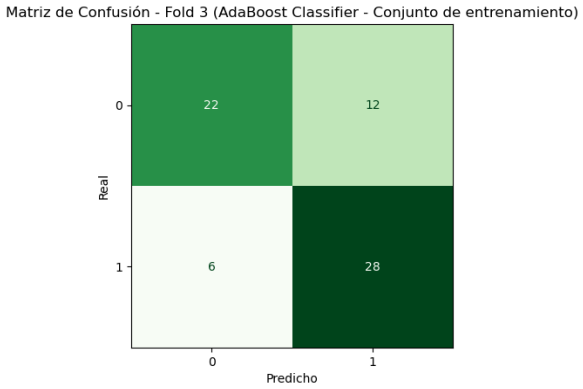

Figura D.26: AdaBoost, Fold 3, C1, ADASYN

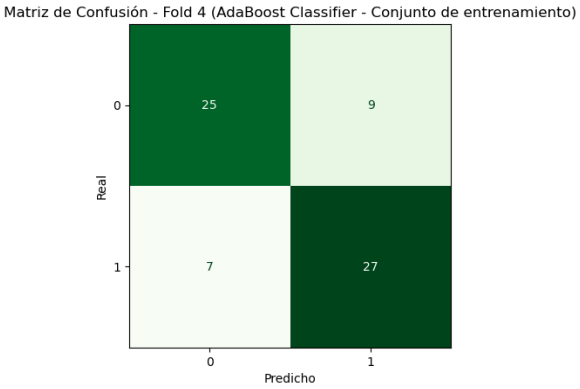

Figura D.27: AdaBoost, Fold 4, C1, ADASYN

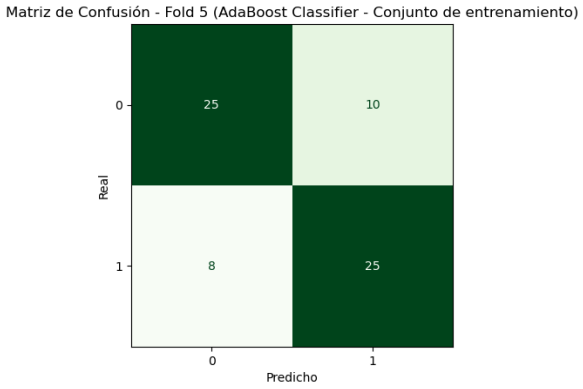

Figura D.27: AdaBoost, Fold 5, C1, ADASYN

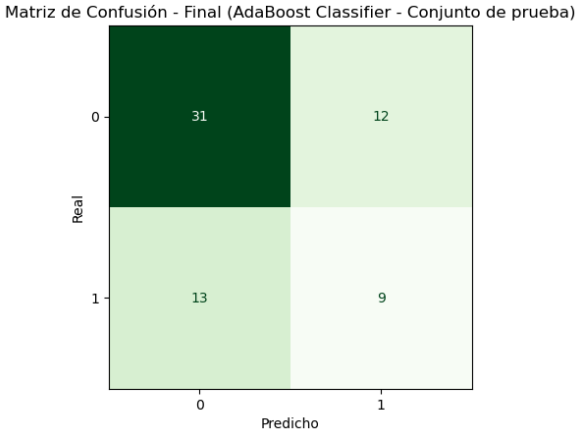

Figura D.28: AdaBoost, Final, C1, ADASYN

Gradient Boosting

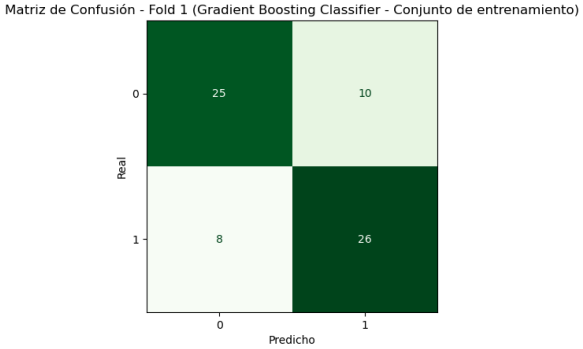

Figura D.29: Gradient Boosting, Fold 1, C1, ADASYN

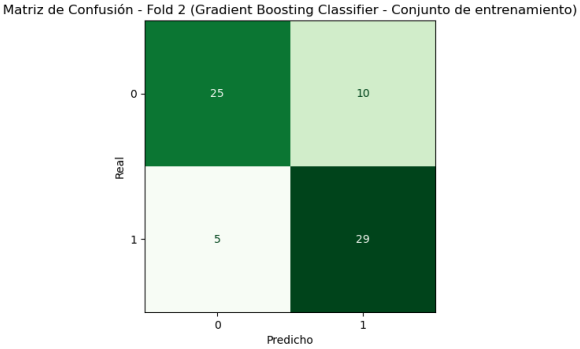

Figura D.30: Gradient Boosting, Fold 2, C1, ADASYN

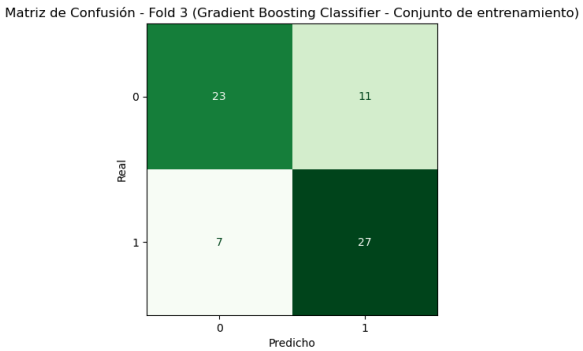

Figura D.30: Gradient Boosting, Fold 3, C1, ADASYN

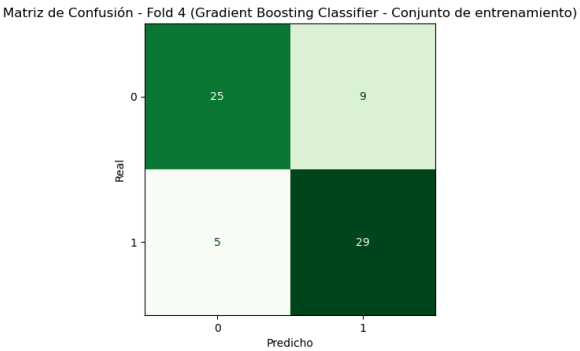

Figura D.31: Gradient Boosting, Fold 4, C1, ADASYN

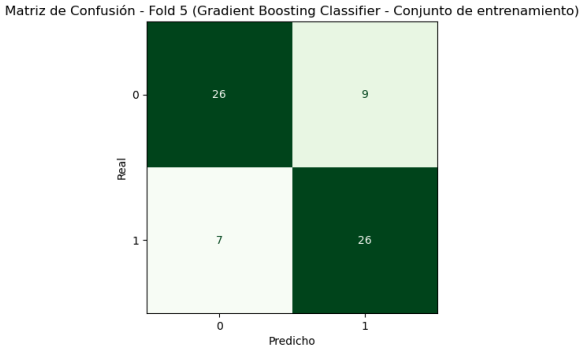

Figura D.31: Gradient Boosting, Fold 5, C1, ADASYN

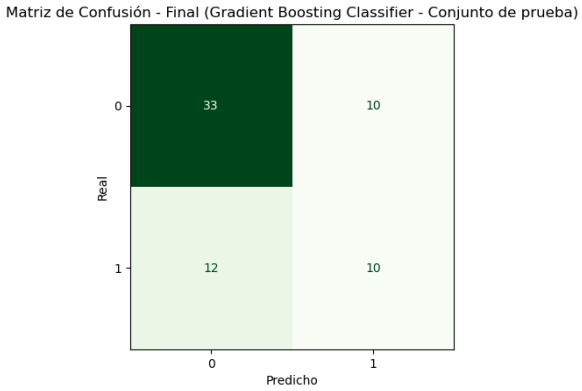

Figura D.32: Gradient Boosting, Final, C1, ADASYN

SVC

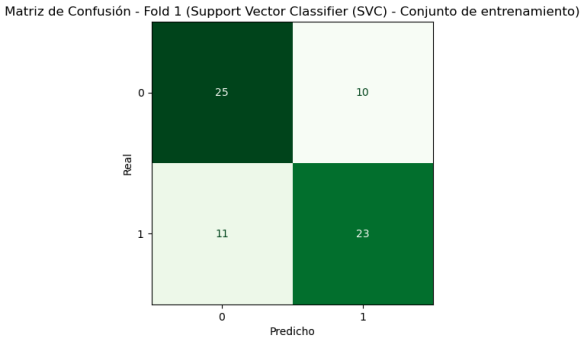

Figura D.33: SVC, Fold 1, C1, ADASYN

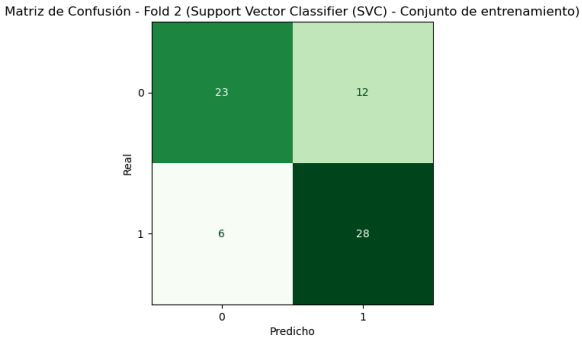

Figura D.34: SVC, Fold 2, C1, ADASYN

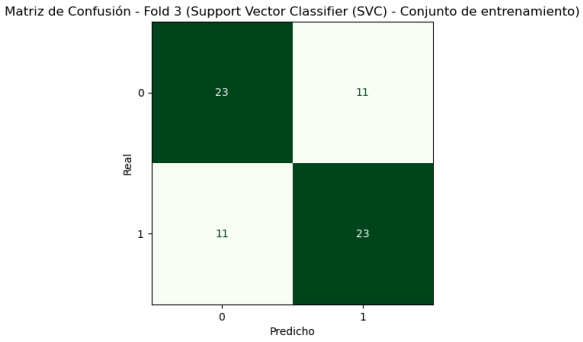

Figura D.34: SVC, Fold 3, C1, ADASYN

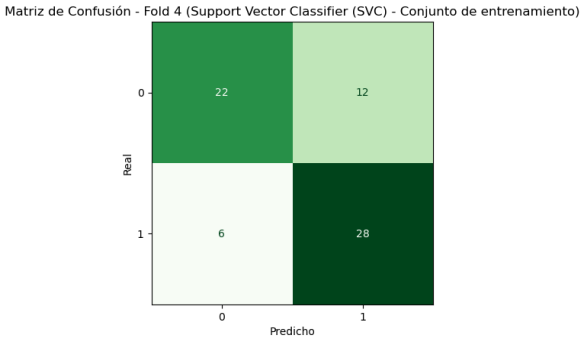

Figura D.35: SVC, Fold 4, C1, ADASYN

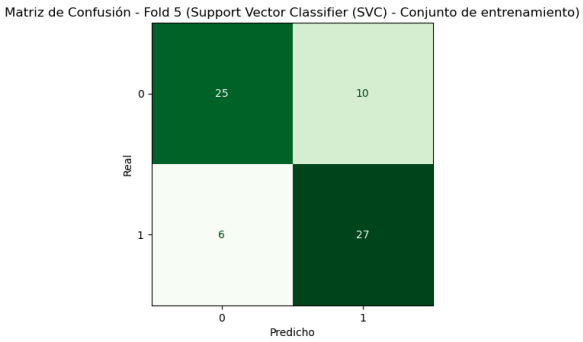

Figura D.35: SVC, Fold 5, C1, ADASYN

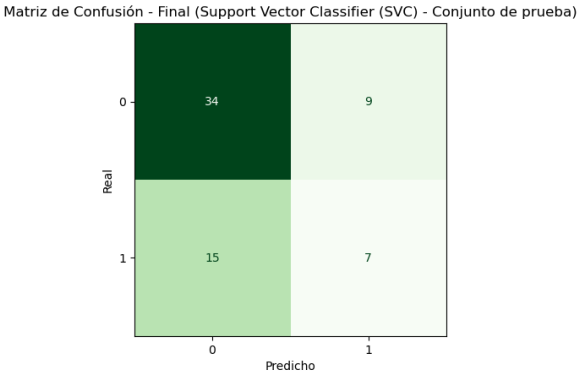

Figura D.36: SVC, Final, C1, ADASYN

k-Nearest Neighbors

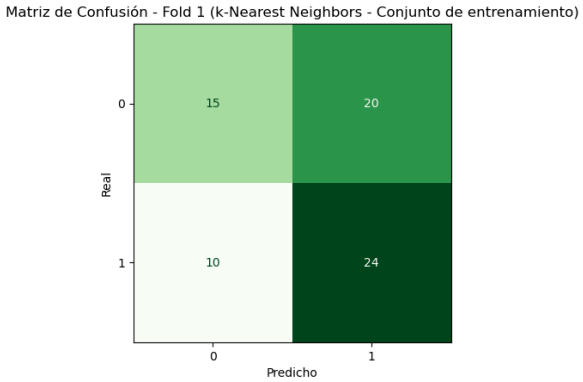

Figura D.37: k-Nearest Neighbors, Fold 1, C1, ADASYN

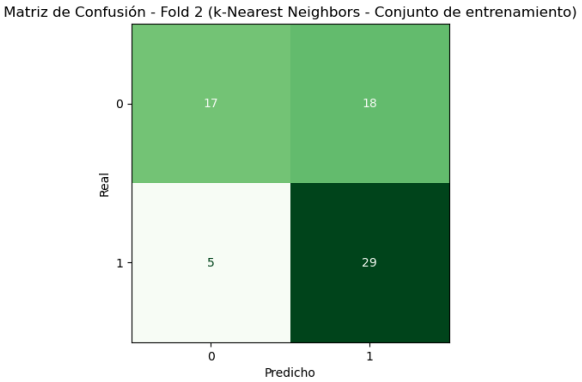

Figura D.38: k-Nearest Neighbors, Fold 2, C1, ADASYN

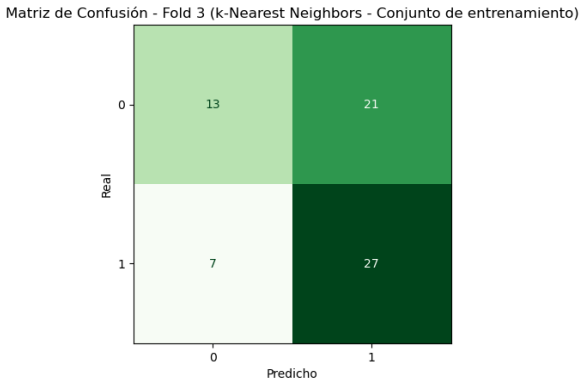

Figura D.38: k-Nearest Neighbors, Fold 3, C1, ADASYN

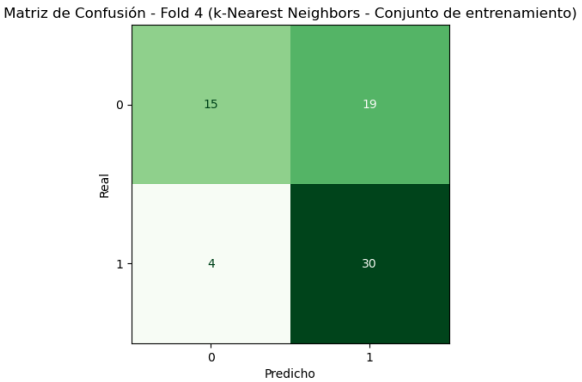

Figura D.39: k-Nearest Neighbors, Fold 4, C1, ADASYN

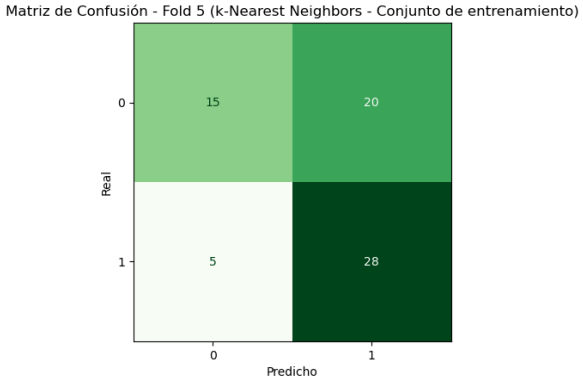

Figura D.39: k-Nearest Neighbors, Fold 5, C1, ADASYN

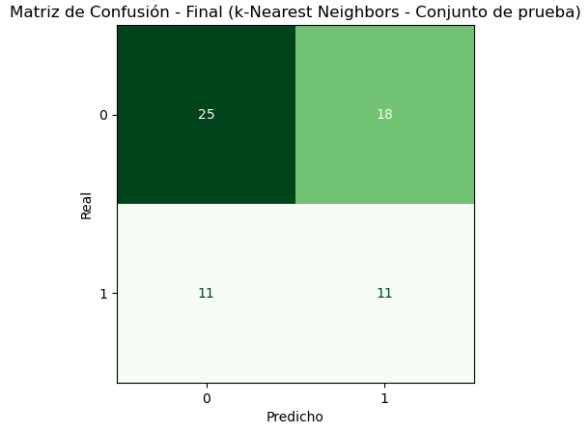

Figura D.40: k-Nearest Neighbors, Final, C1, ADASYN

Gaussian Naive Bayes

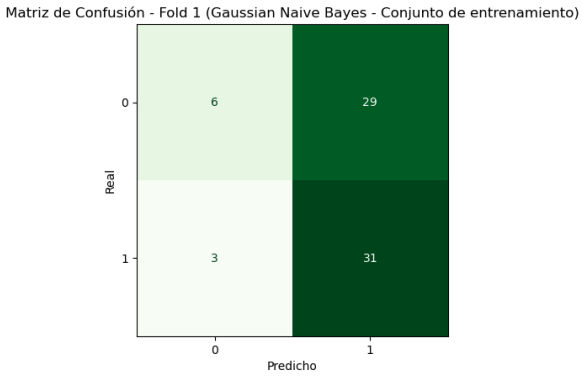

Figura D.41: Gaussian Naive Bayes, Fold 1, C1, ADASYN

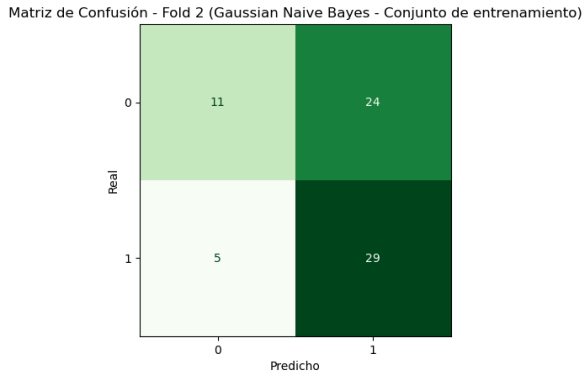

Figura D.42: Gaussian Naive Bayes, Fold 2, C1, ADASYN

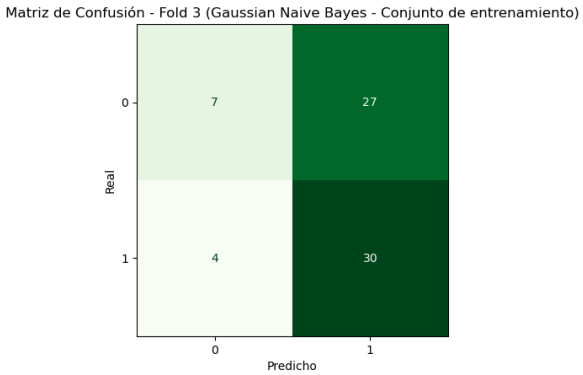

Figura D.42: Gaussian Naive Bayes, Fold 3, C1, ADASYN

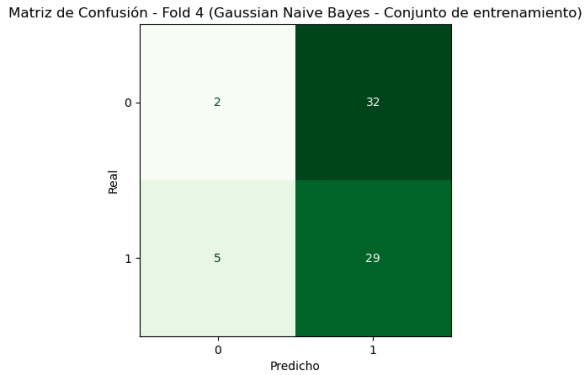

Figura D.43: Gaussian Naive Bayes, Fold 4, C1, ADASYN

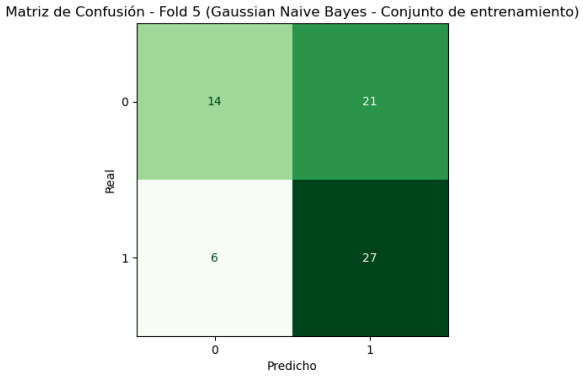

Figura D.43: Gaussian Naive Bayes, Fold 5, C1, ADASYN

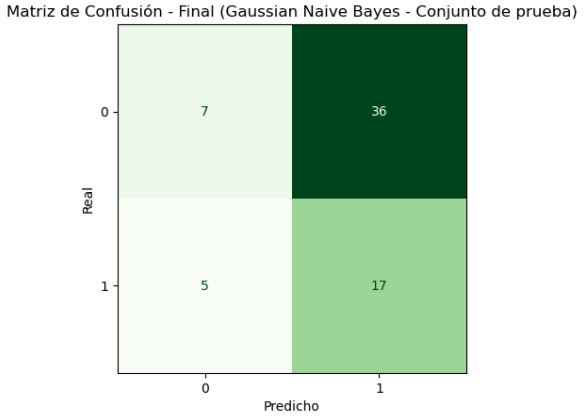

Figura D.44: Gaussian Naive Bayes, Final, C1, ADASYN

MLP

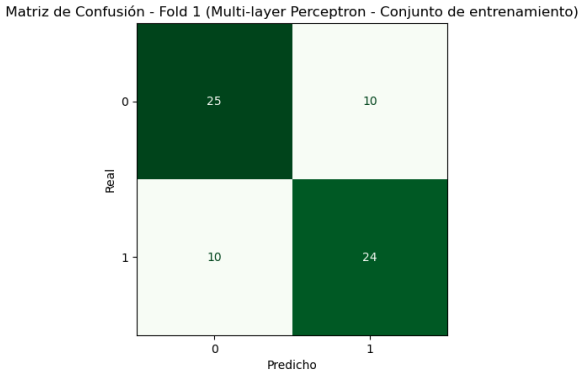

Figura D.45: MLP, Fold 1, C1, ADASYN

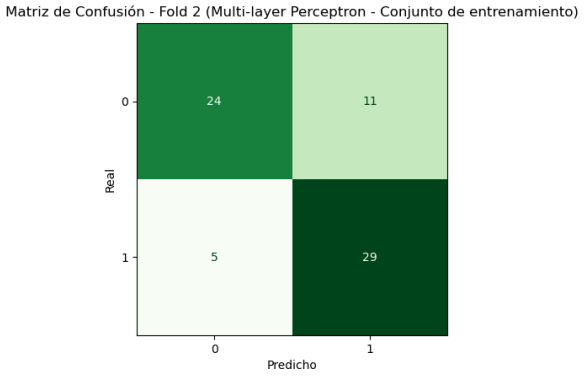

Figura D.46: MLP, Fold 2, C1, ADASYN

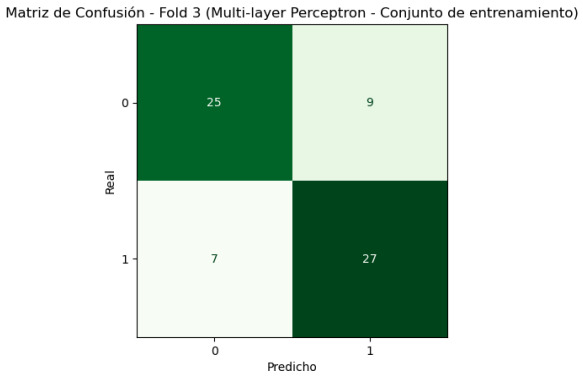

Figura D.46: MLP, Fold 3, C1, ADASYN

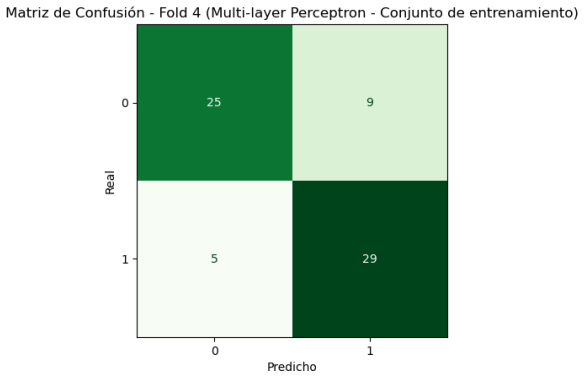

Figura D.47: MLP, Fold 4, C1, ADASYN

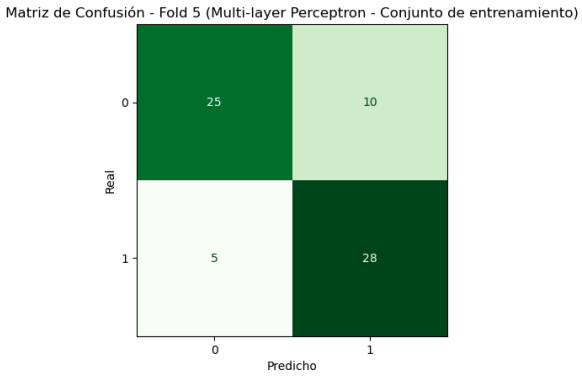

Figura D.47: MLP, Fold 5, C1, ADASYN

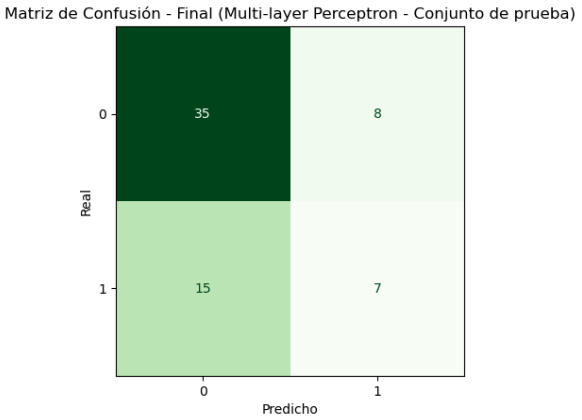

Figura D.48: MLP, Final, C1, ADASYN

D.1.2. SMOTE

Logistic Regression

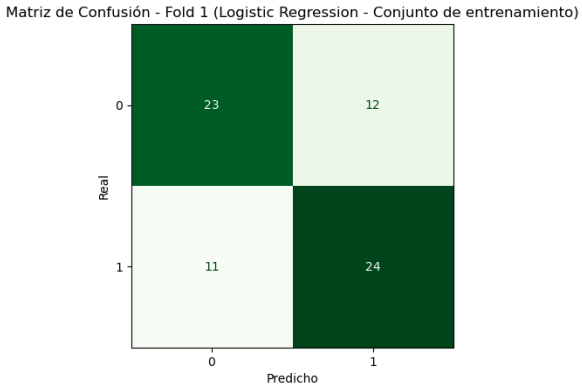

Figura D.49: Logistic Regression, Fold 1, C1, SMOTE

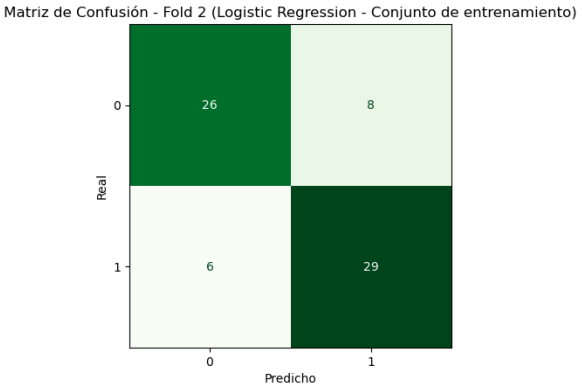

Figura D.50: Logistic Regression, Fold 2, C1, SMOTE

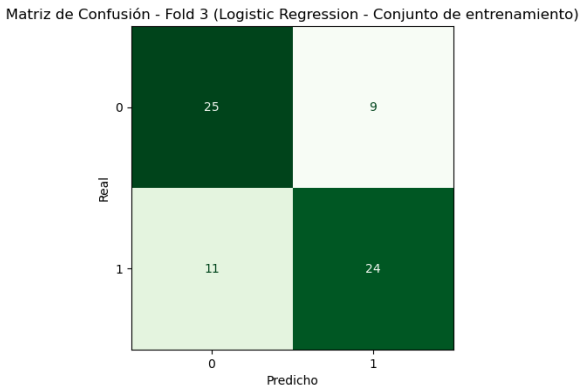

Figura D.50: Logistic Regression, Fold 3, C1, SMOTE

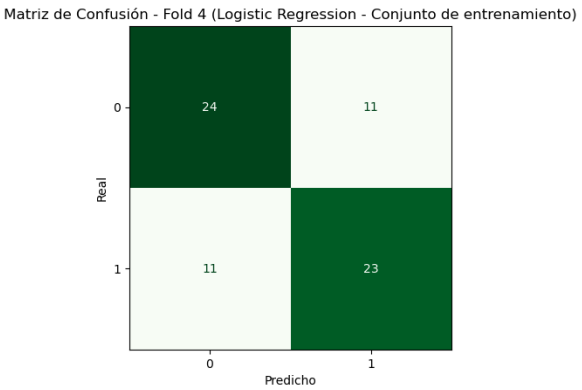

Figura D.51: Logistic Regression, Fold 4, C1, SMOTE

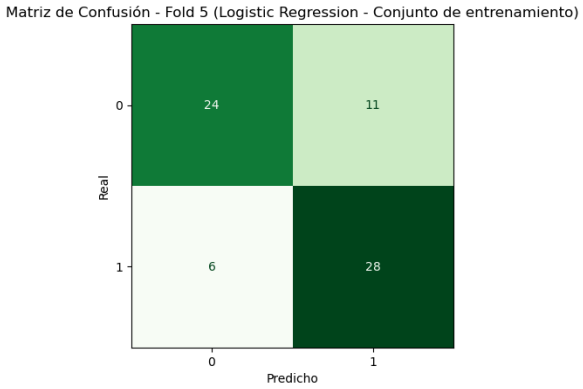

Figura D.51: Logistic Regression, Fold 5, C1, SMOTE

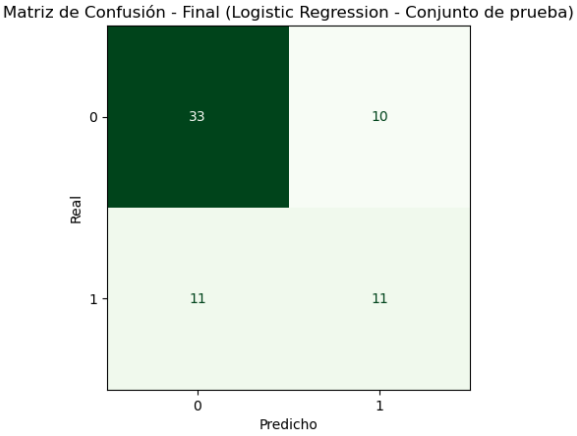

Figura D.52: Logistic Regression, Final, C1, SMOTE

Logistic Regression Lasso

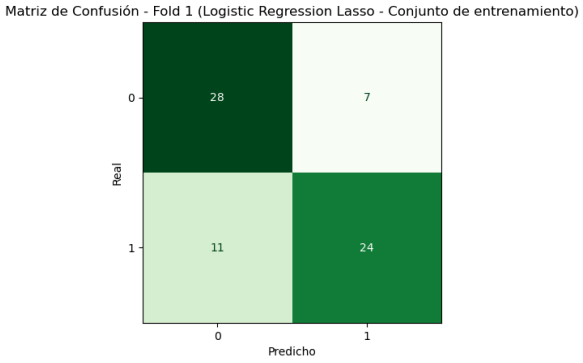

Figura D.53: Logistic Regression Lasso, Fold 1, C1, SMOTE

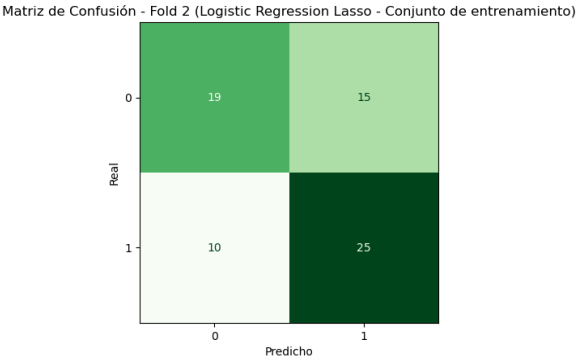

Figura D.54: Logistic Regression Lasso, Fold 2, C1, SMOTE

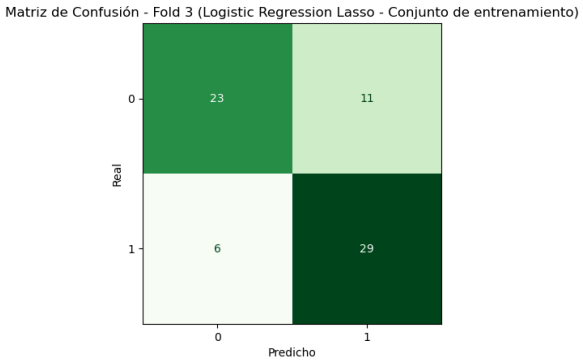

Figura D.54: Logistic Regression Lasso, Fold 3, C1, SMOTE

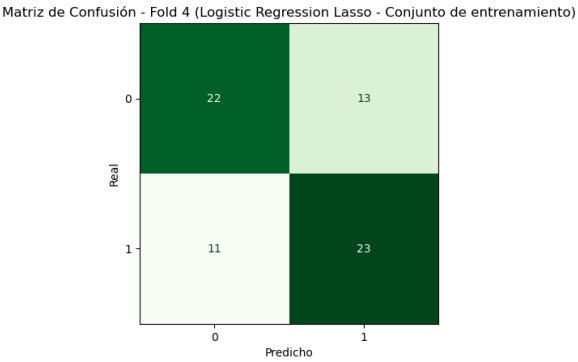

Figura D.55: Logistic Regression Lasso, Fold 4, C1, SMOTE

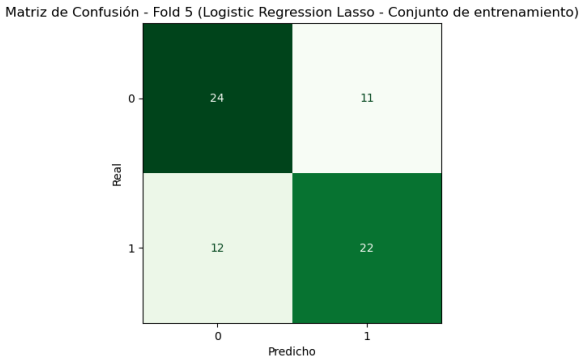

Figura D.55: Logistic Regression Lasso, Fold 5, C1, SMOTE

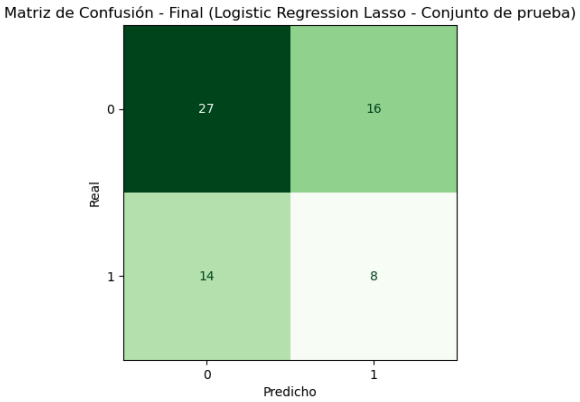

Figura D.56: Logistic Regression Lasso, Final, C1, SMOTE

Ridge

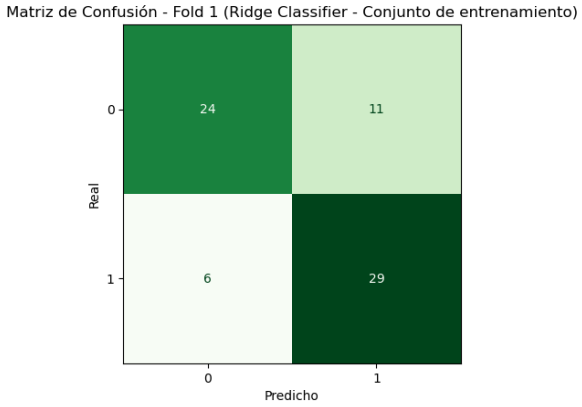

Figura D.57: Ridge, Fold 1, C1, SMOTE

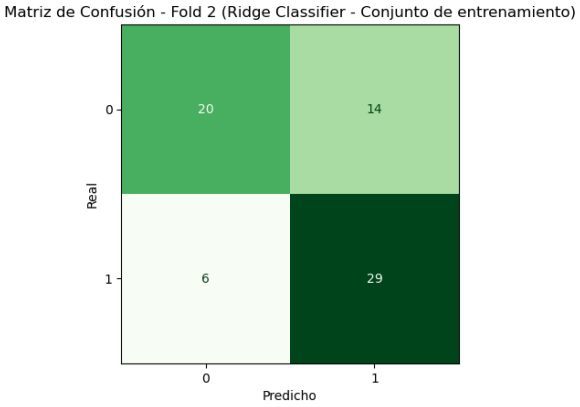

Figura D.58: Ridge, Fold 2, C1, SMOTE

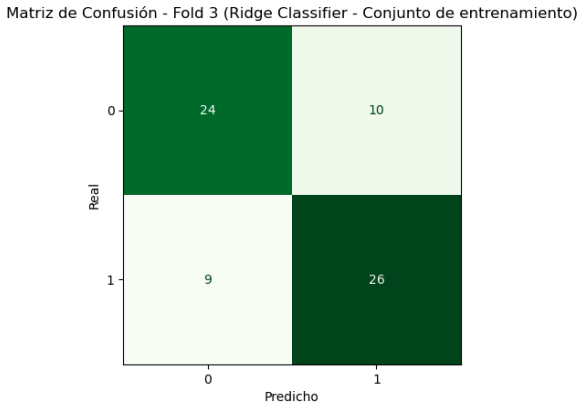

Figura D.58: Ridge, Fold 3, C1, SMOTE

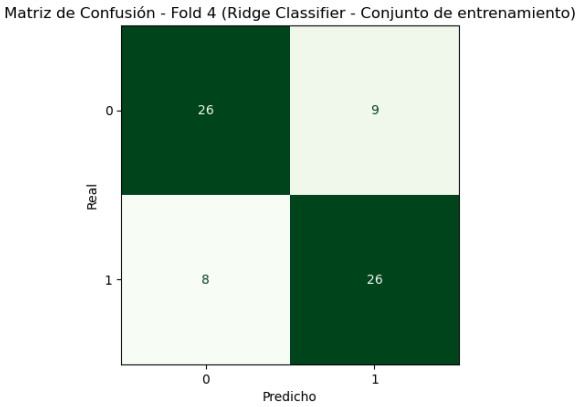

Figura D.59: Ridge, Fold 4, C1, SMOTE

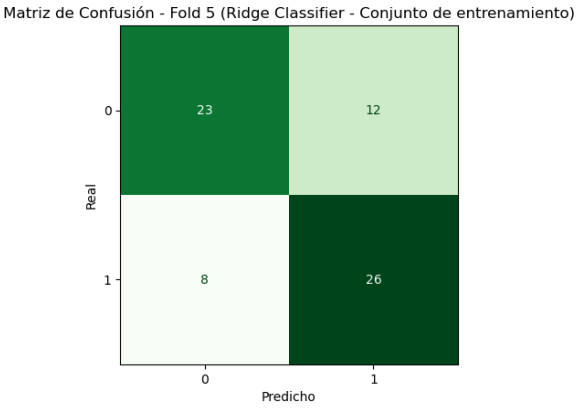

Figura D.59: Ridge, Fold 5, C1, SMOTE

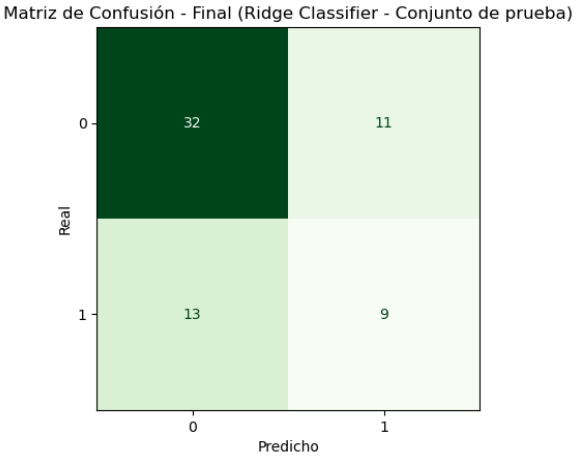

Figura D.60: Ridge, Final, C1, SMOTE

SDG

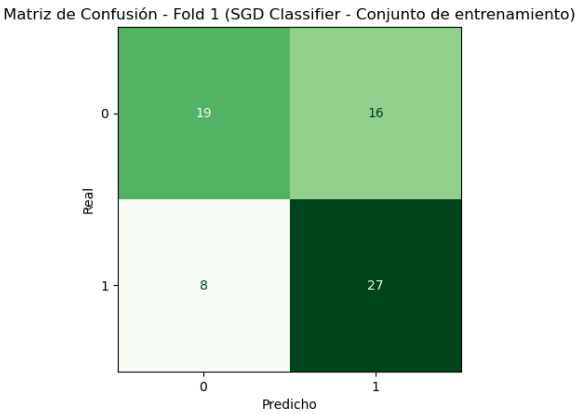

Figura D.61: SDG, Fold 1, C1, SMOTE

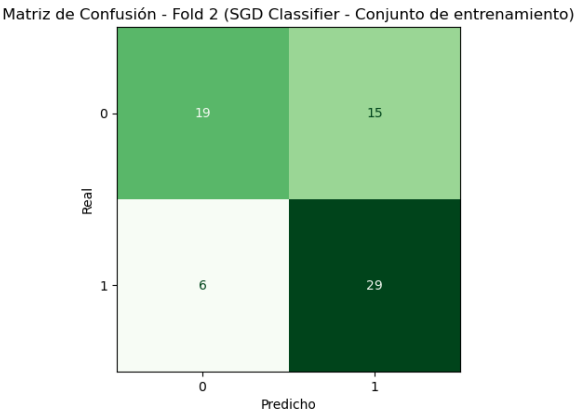

Figura D.62: SDG, Fold 2, C1, SMOTE

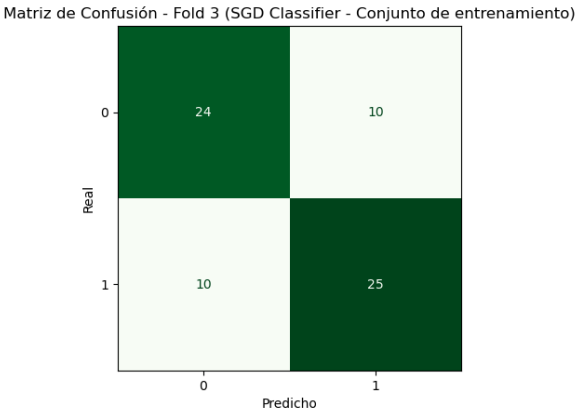

Figura D.62: SDG, Fold 3, C1, SMOTE

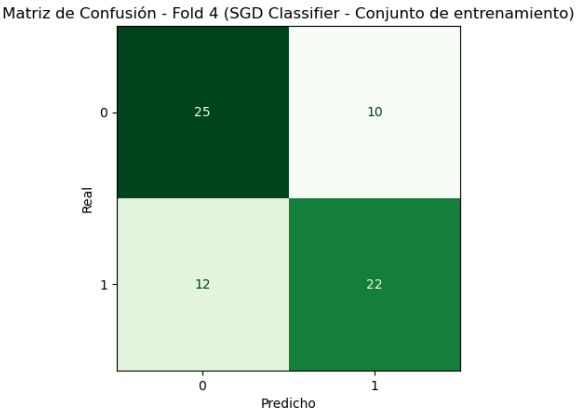

Figura D.63: SDG, Fold 4, C1, SMOTE

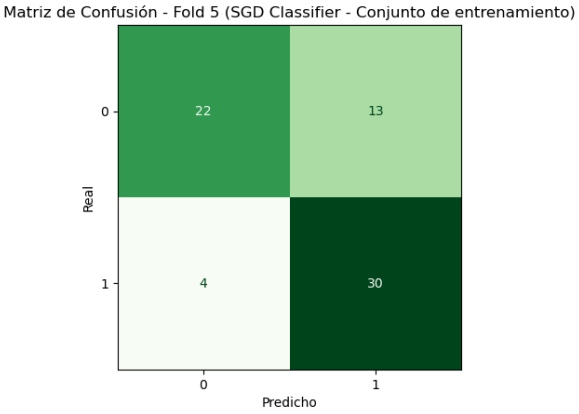

Figura D.63: SDG, Fold 5, C1, SMOTE

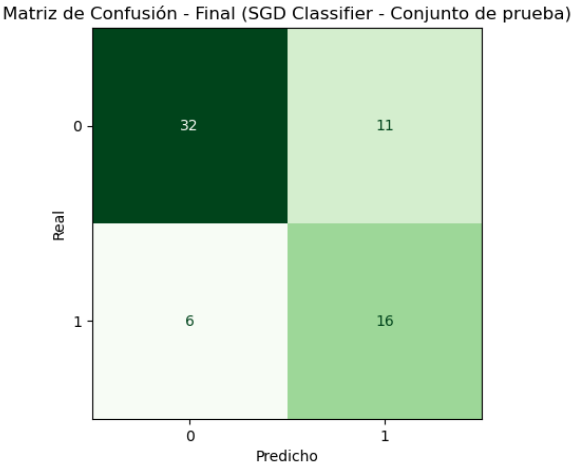

Figura D.64: SDG, Final, C1, SMOTE

Decision Tree

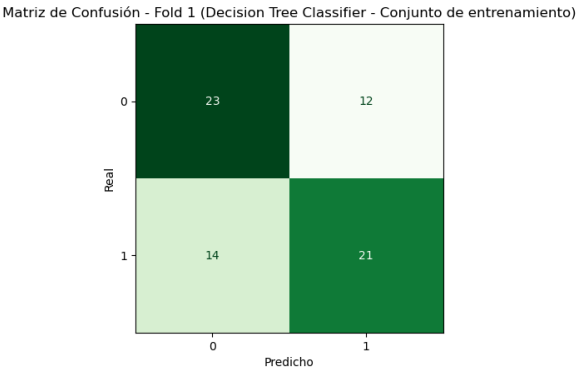

Figura D.65: Decision Tree, Fold 1, C1, SMOTE

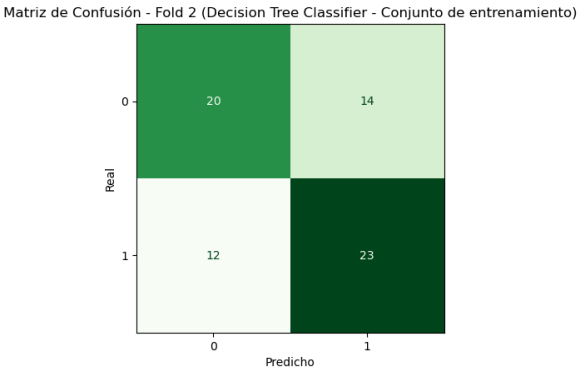

Figura D.66: Decision Tree, Fold 2, C1, SMOTE

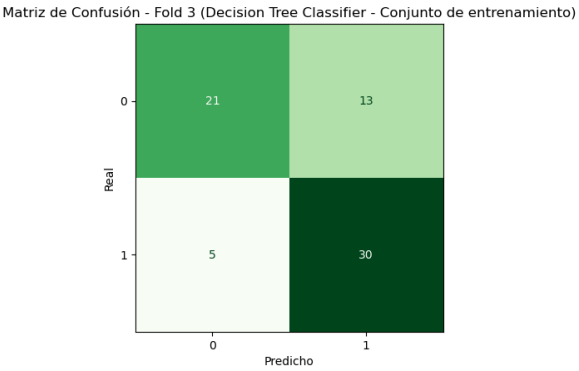

Figura D.66: Decision Tree, Fold 3, C1, SMOTE

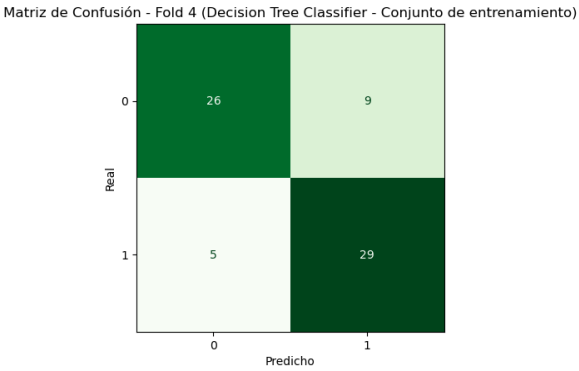

Figura D.67: Decision Tree, Fold 4, C1, SMOTE

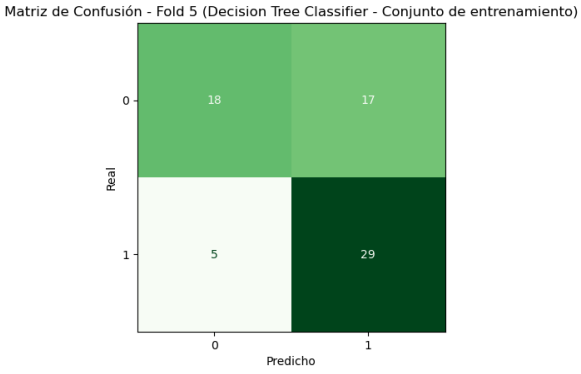

Figura D.67: Decision Tree, Fold 5, C1, SMOTE

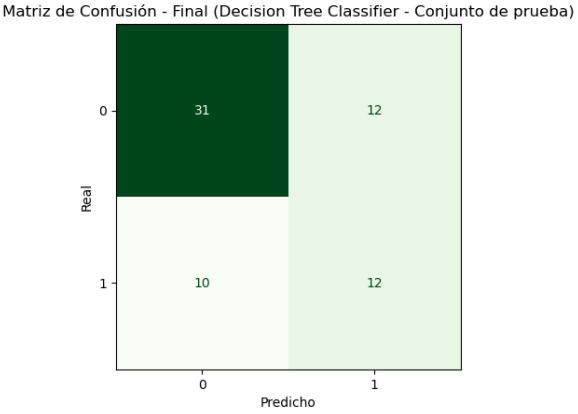

Figura D.68: Decision Tree, Final, C1, SMOTE

Random Forest

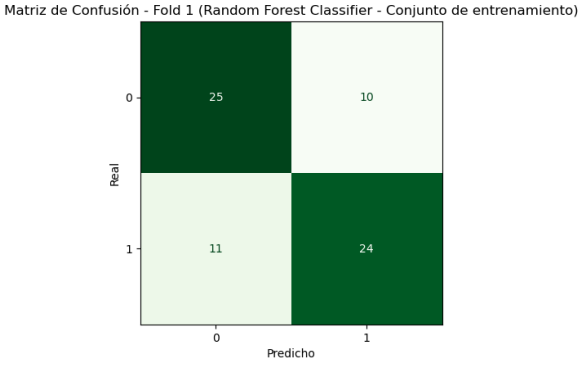

Figura D.69: Random Forest, Fold 1, C1, SMOTE

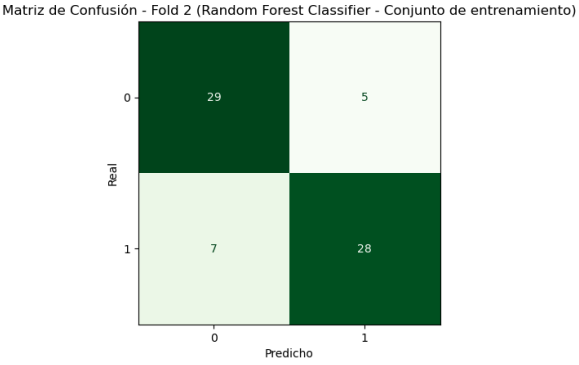

Figura D.70: Random Forest, Fold 2, C1, SMOTE

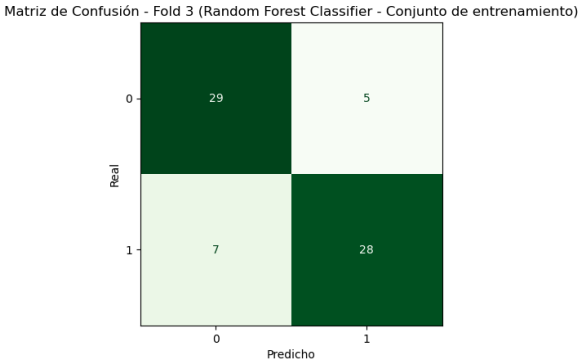

Figura D.70: Random Forest, Fold 3, C1, SMOTE

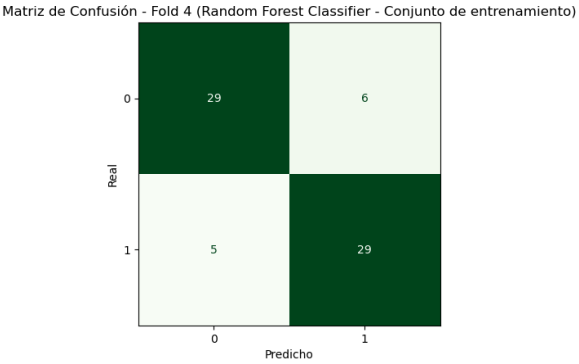

Figura D.71: Random Forest, Fold 4, C1, SMOTE

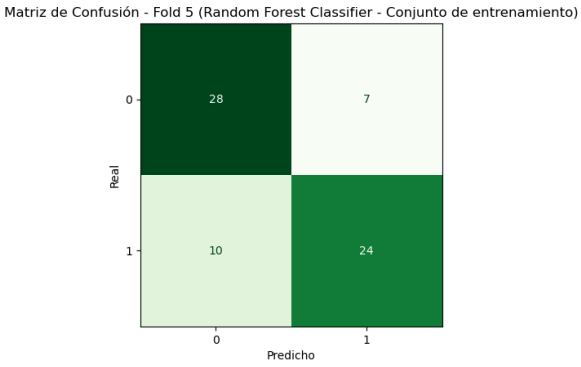

Figura D.71: Random Forest, Fold 5, C1, SMOTE

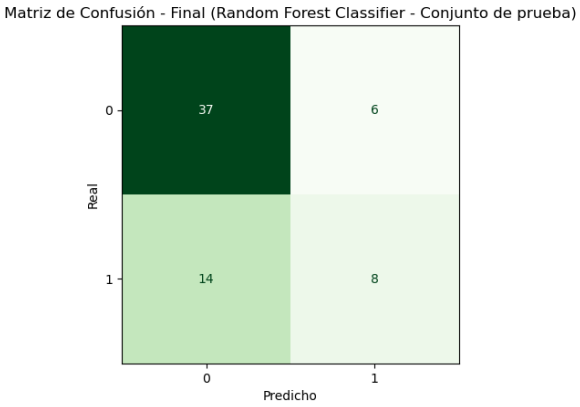

Figura D.72: Random Forest, Final, C1, SMOTE

AdaBoost

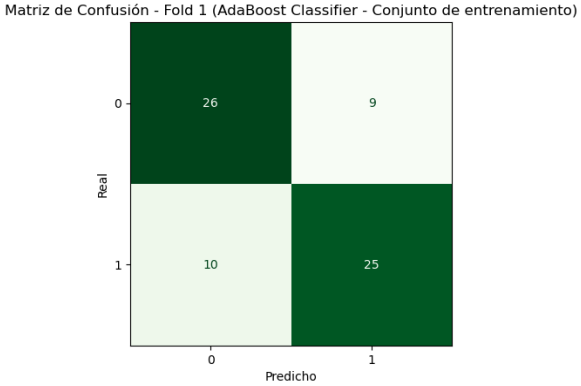

Figura D.73: AdaBoost, Fold 1, C1, SMOTE

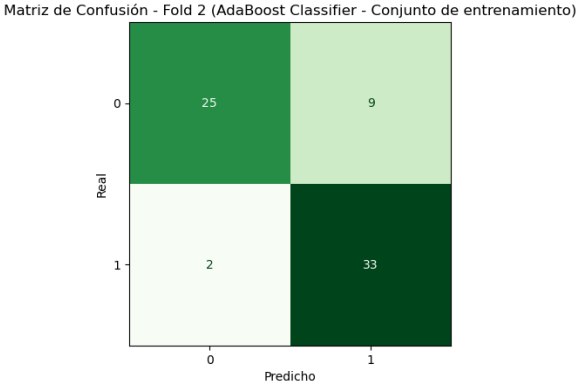

Figura D.74: AdaBoost, Fold 2, C1, SMOTE

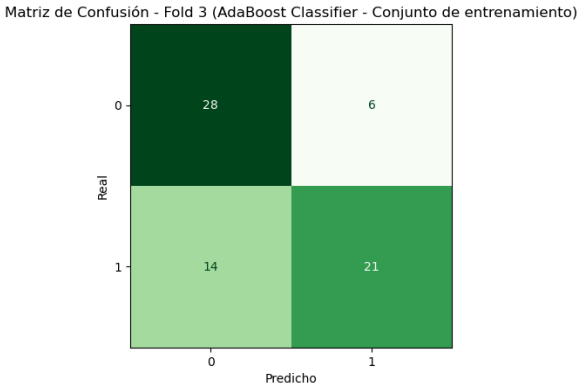

Figura D.74: AdaBoost, Fold 3, C1, SMOTE

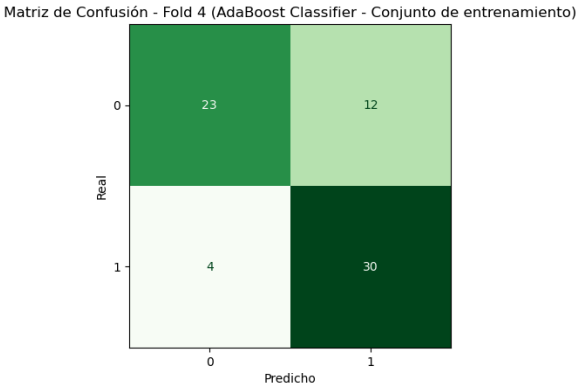

Figura D.75: AdaBoost, Fold 4, C1, SMOTE

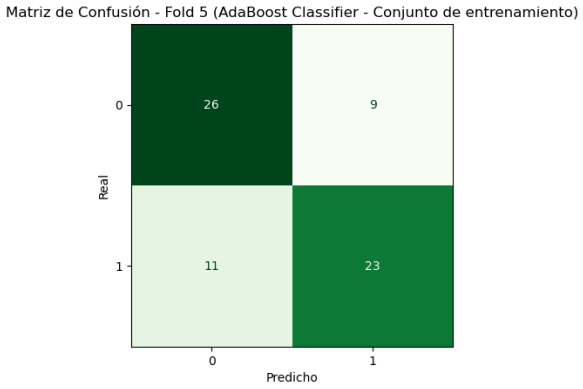

Figura D.75: AdaBoost, Fold 5, C1, SMOTE

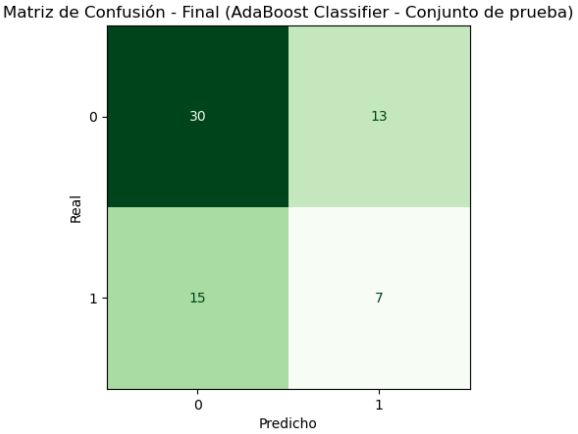

Figura D.76: AdaBoost, Final, C1, SMOTE

Gradient Boosting

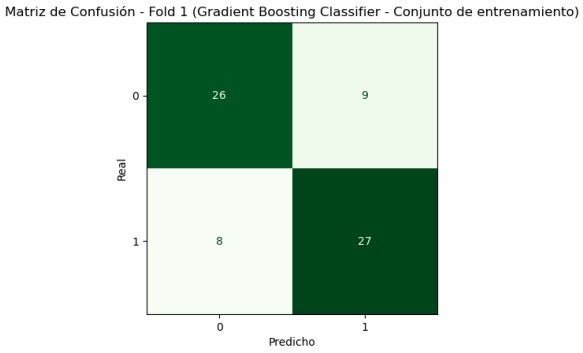

Figura D.77: Gradient Boosting, Fold 1, C1, SMOTE

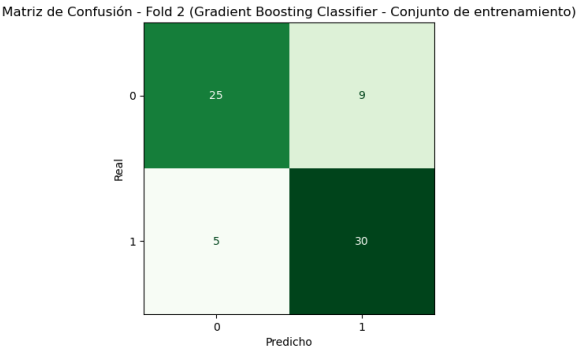

Figura D.78: Gradient Boosting, Fold 2, C1, SMOTE

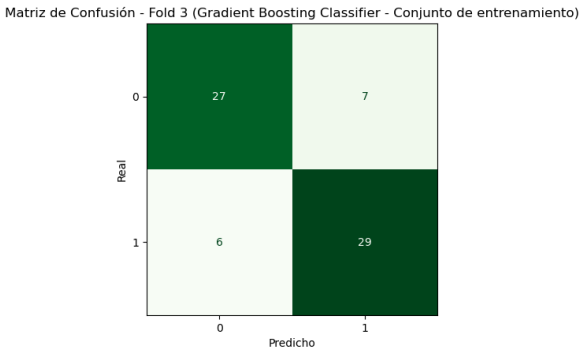

Figura D.78: Gradient Boosting, Fold 3, C1, SMOTE

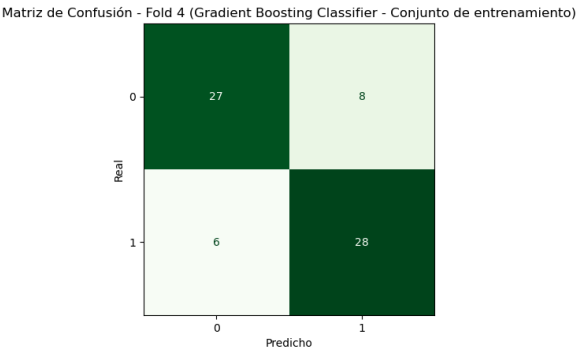

Figura D.79: Gradient Boosting, Fold 4, C1, SMOTE

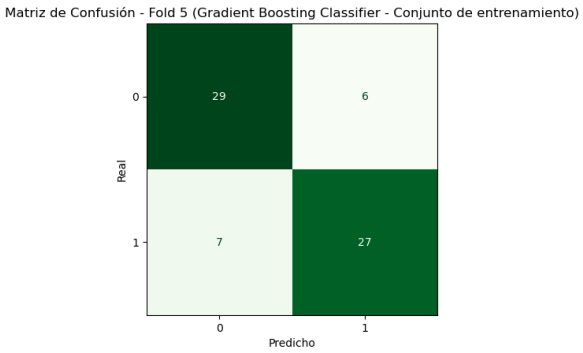

Figura D.79: Gradient Boosting, Fold 5, C1, SMOTE

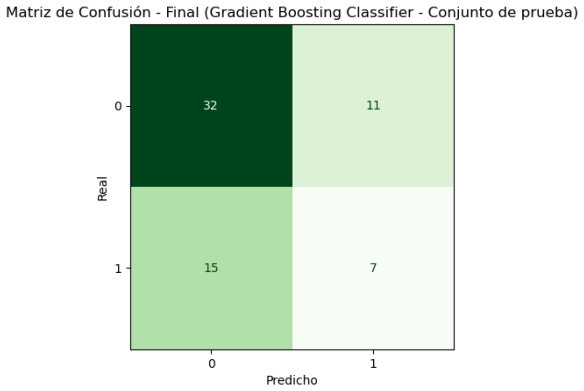

Figura D.80: Gradient Boosting, Final, C1, SMOTE

SVC

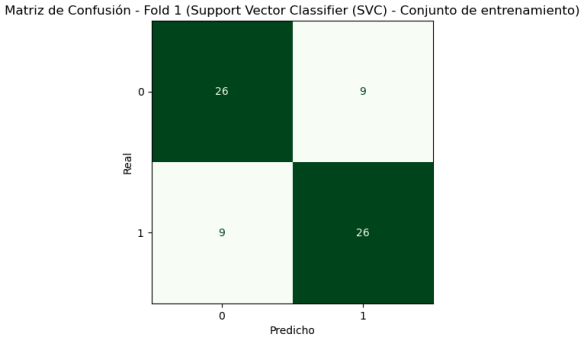

Figura D.81: SVC, Fold 1, C1, SMOTE

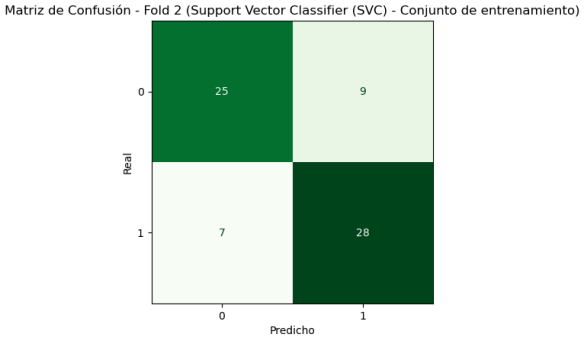

Figura D.82: SVC, Fold 2, C1, SMOTE

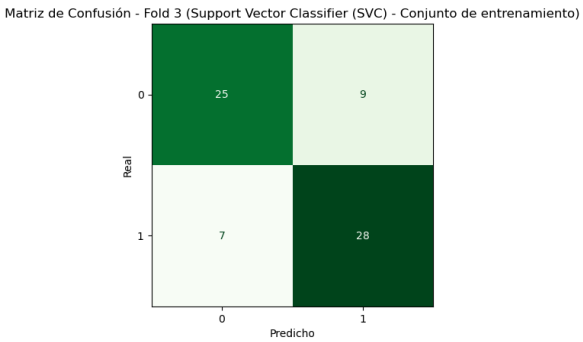

Figura D.82: SVC, Fold 3, C1, SMOTE

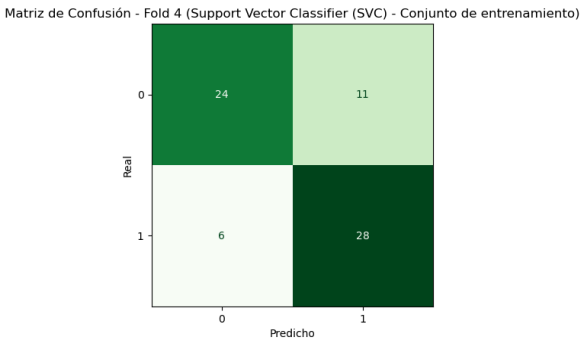

Figura D.83: SVC, Fold 4, C1, SMOTE

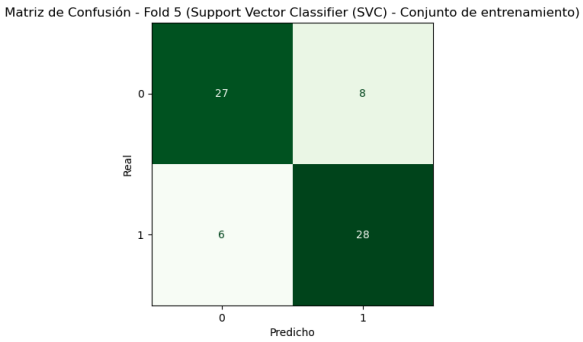

Figura D.83: SVC, Fold 5, C1, SMOTE

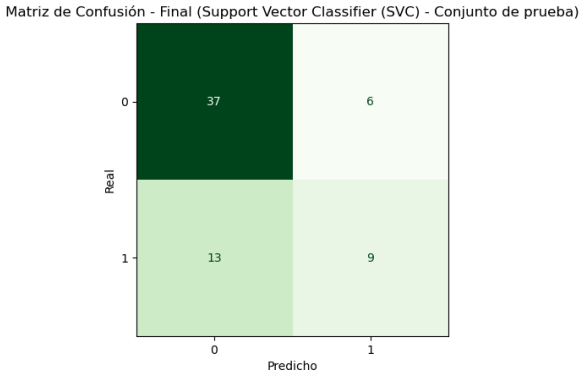

Figura D.84: SVC, Final, C1, SMOTE

k-Nearest Neighbors

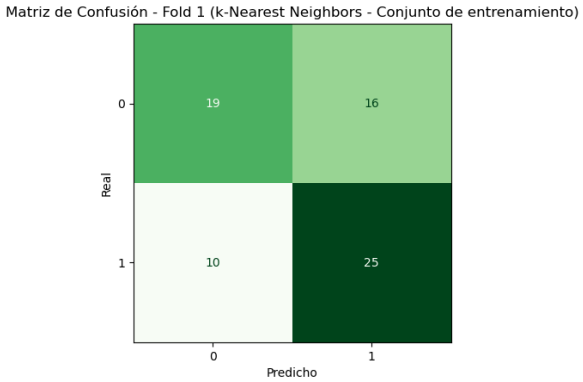

Figura D.85: k-Nearest Neighbors, Fold 1, C1, SMOTE

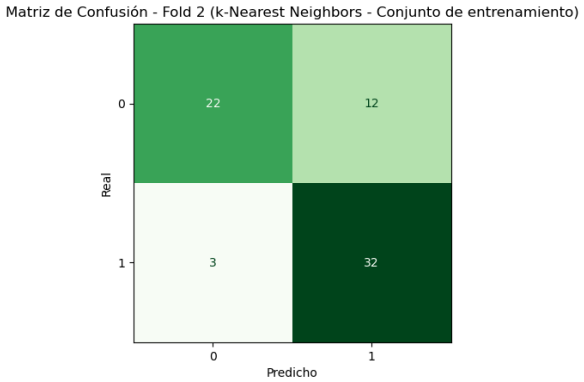

Figura D.86: k-Nearest Neighbors, Fold 2, C1, SMOTE

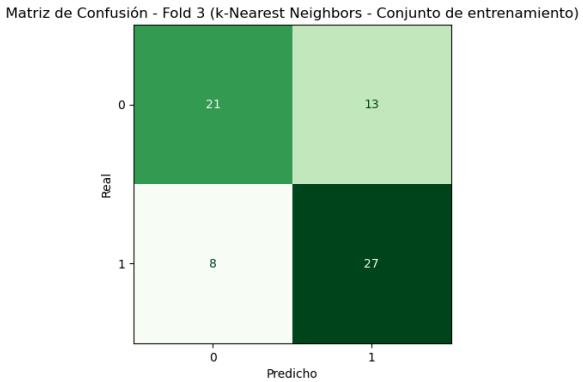

Figura D.86: k-Nearest Neighbors, Fold 3, C1, SMOTE

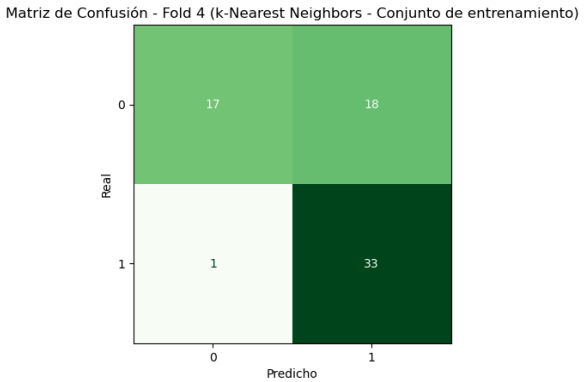

Figura D.87: k-Nearest Neighbors, Fold 4, C1, SMOTE

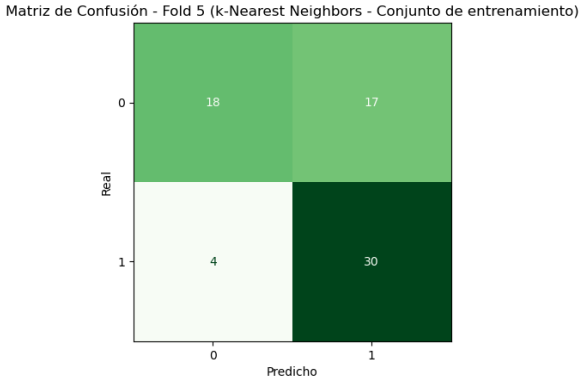

Figura D.87: k-Nearest Neighbors, Fold 5, C1, SMOTE

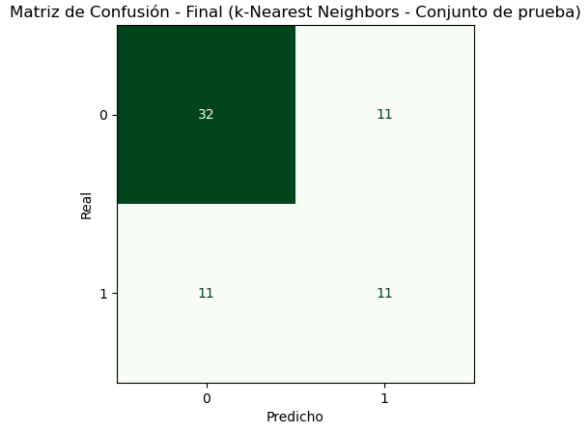

Figura D.88: k-Nearest Neighbors, Final, C1, SMOTE

Gaussian Naive Bayes

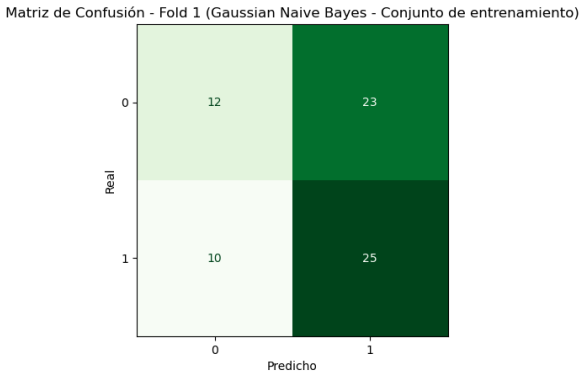

Figura D.89: Gaussian Naive Bayes, Fold 1, C1, SMOTE

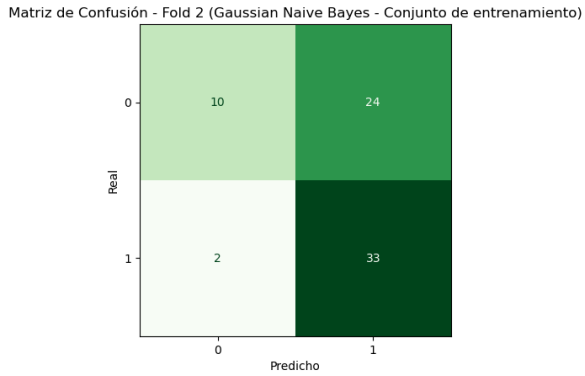

Figura D.90: Gaussian Naive Bayes, Fold 2, C1, SMOTE

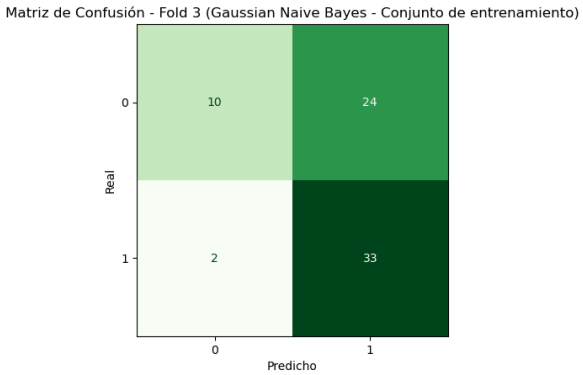

Figura D.90: Gaussian Naive Bayes, Fold 3, C1, SMOTE

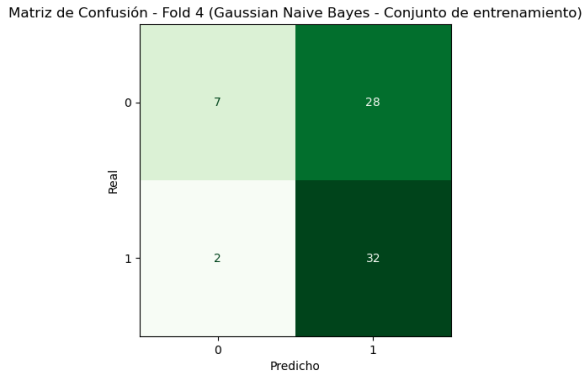

Figura D.91: Gaussian Naive Bayes, Fold 4, C1, SMOTE

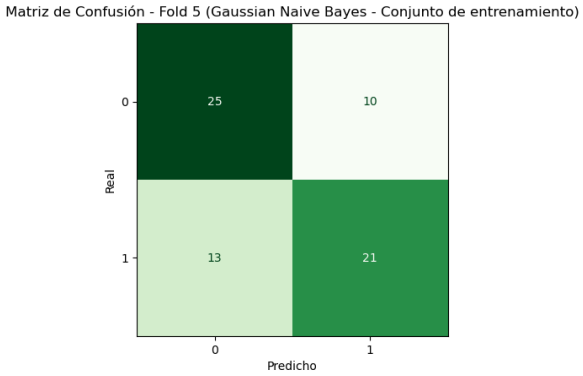

Figura D.91: Gaussian Naive Bayes, Fold 5, C1, SMOTE

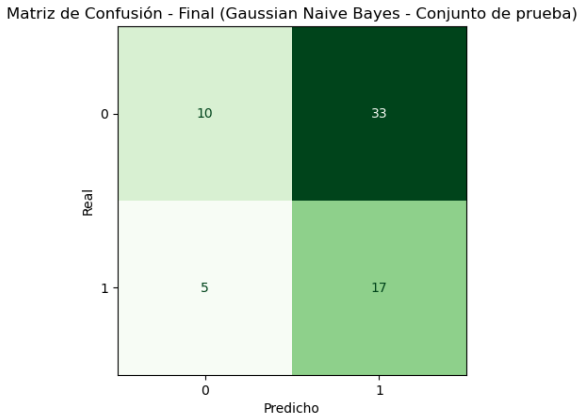

Figura D.92: Gaussian Naive Bayes, Final, C1, SMOTE

MLP

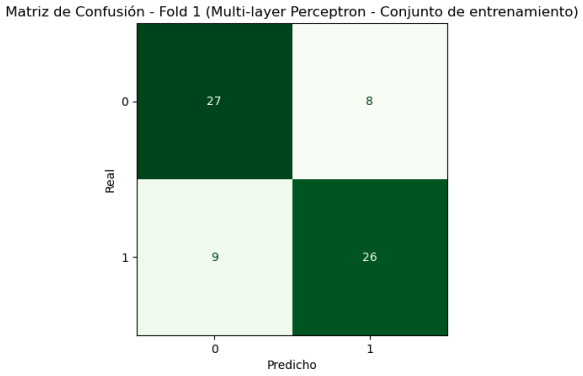

Figura D.93: MLP, Fold 1, C1, SMOTE

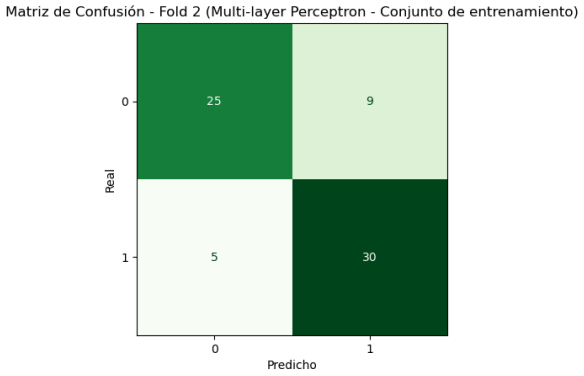

Figura D.94: MLP, Fold 2, C1, SMOTE

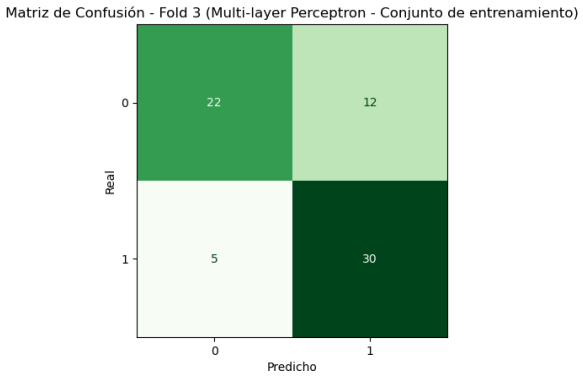

Figura D.94: MLP, Fold 3, C1, SMOTE

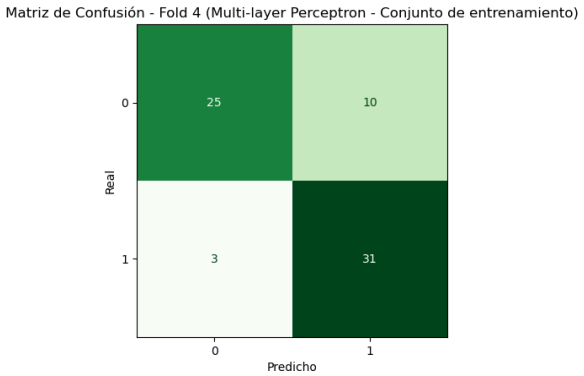

Figura D.95: MLP, Fold 4, C1, SMOTE

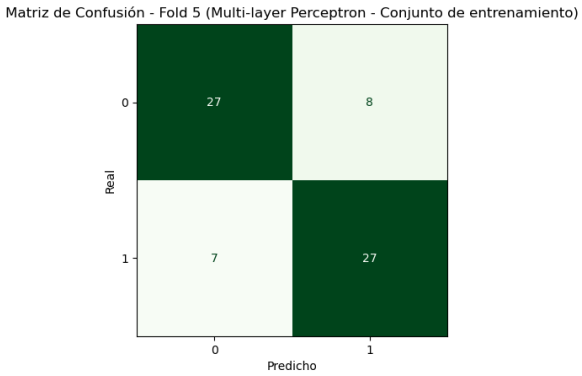

Figura D.95: MLP, Fold 5, C1, SMOTE

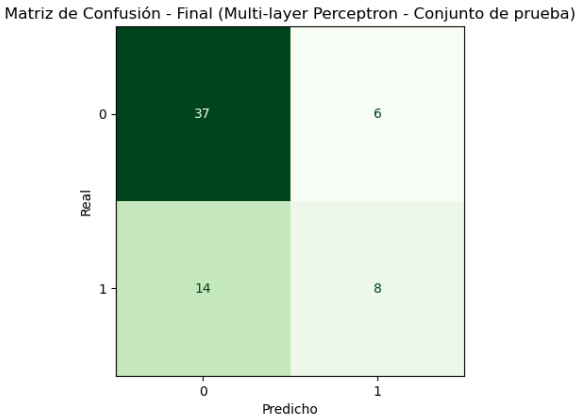

Figura D.96: MLP, Final, C1, SMOTE

D.2 Clasificación 2

D.2.1. ADASYN

Logistic Regression

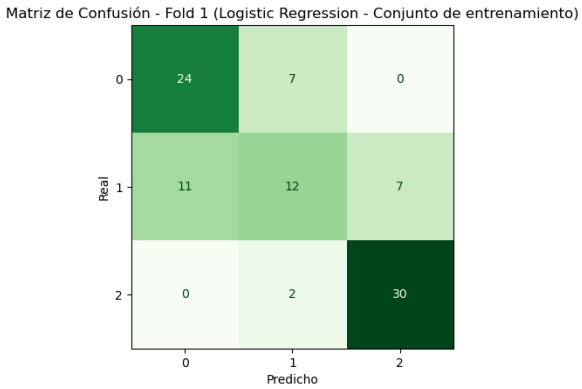

Figura D.97: Logistic Regression, Fold 1, C2, ADASYN

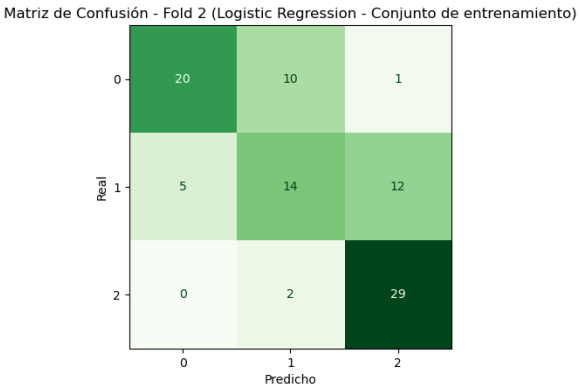

Figura D.98: Logistic Regression, Fold 2, C2, ADASYN

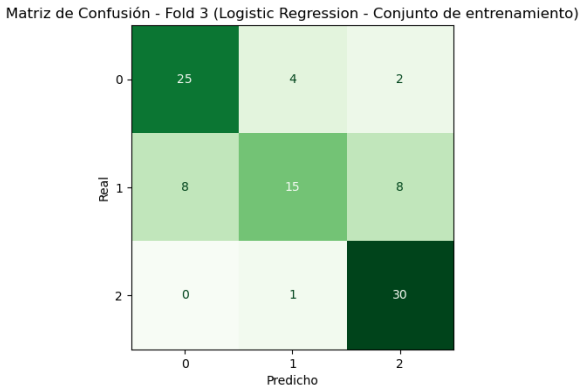

Figura D.98: Logistic Regression, Fold 3, C2, ADASYN

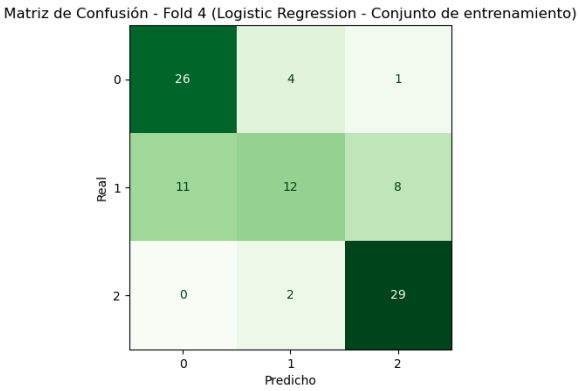

Figura D.99: Logistic Regression, Fold 4, C2, ADASYN

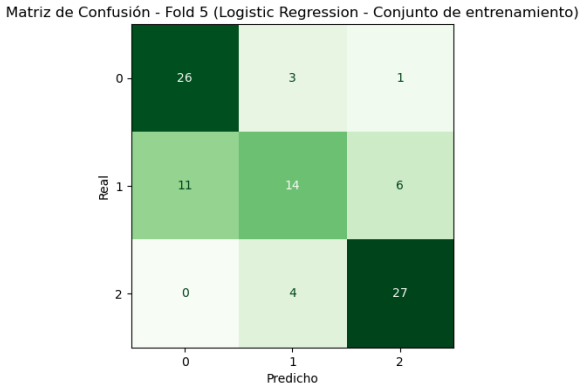

Figura D.99: Logistic Regression, Fold 5, C2, ADASYN

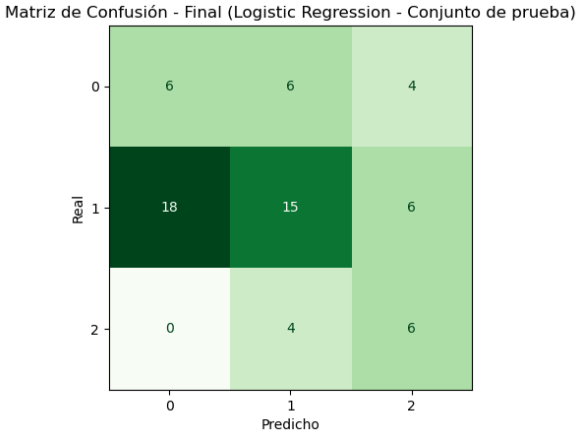

Figura D.100: Logistic Regression, Final, C2, ADASYN

Logistic Regression Lasso

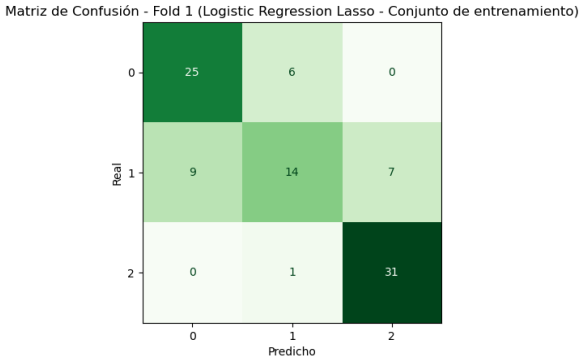

Figura D.101: Logistic Regression Lasso, Fold 1, C2, ADASYN

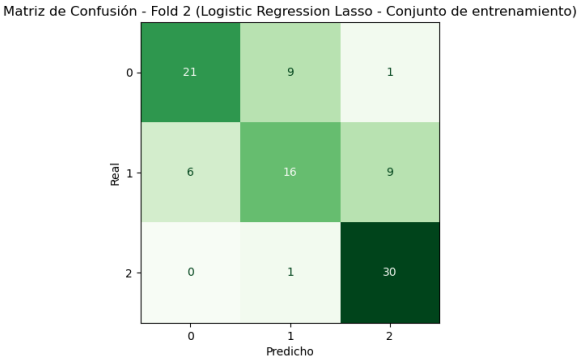

Figura D.102: Logistic Regression Lasso, Fold 2, C2, ADASYN

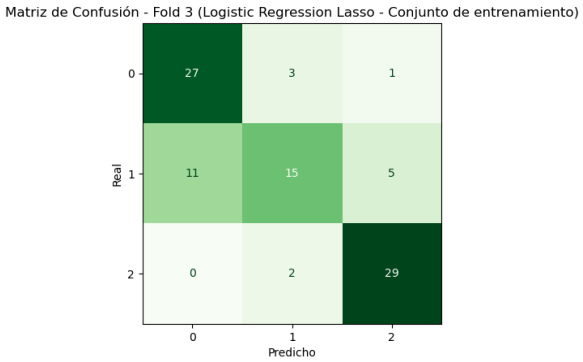

Figura D.102: Logistic Regression Lasso, Fold 3, C2, ADASYN

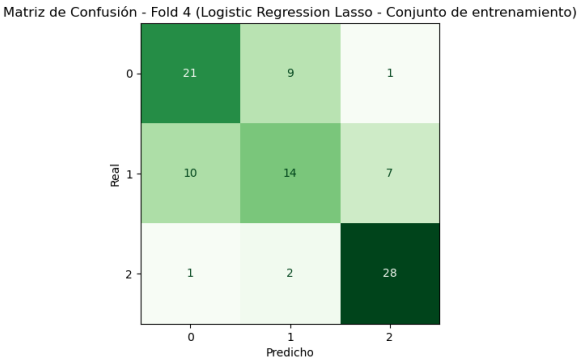

Figura D.103: Logistic Regression Lasso, Fold 4, C2, ADASYN

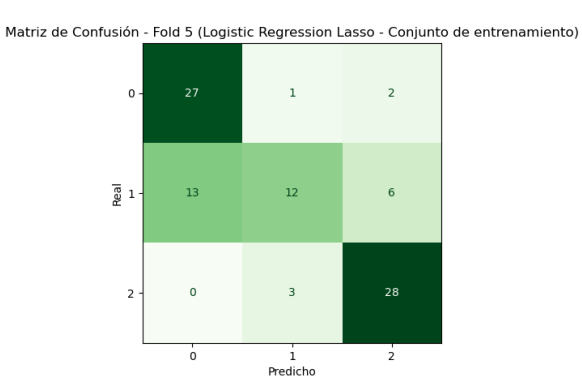

Figura D.103: Logistic Regression Lasso, Fold 5, C2, ADASYN

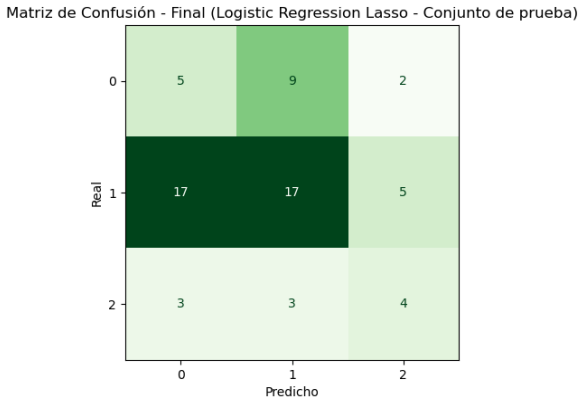

Figura D.104: Logistic Regression Lasso, Final, C2, ADASYN

Ridge

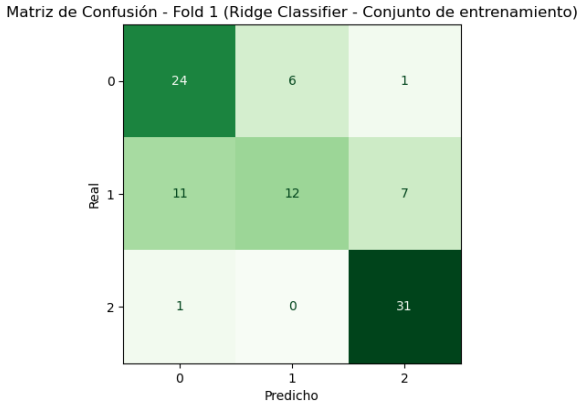

Figura D.105: Ridge, Fold 1, C2, ADASYN

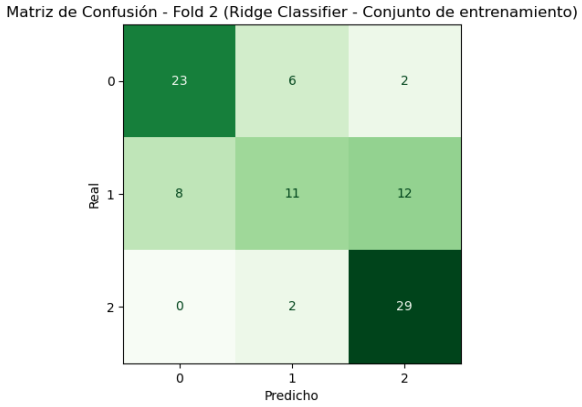

Figura D.106: Ridge, Fold 2, C2, ADASYN

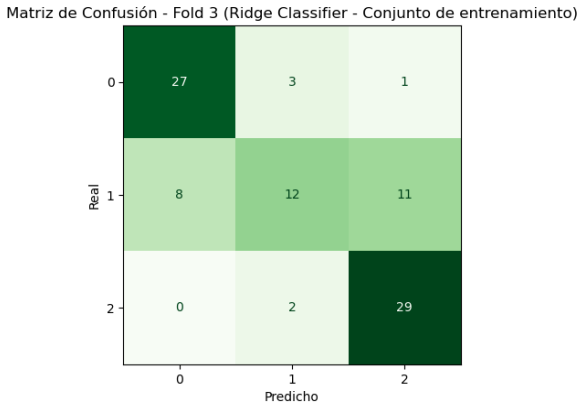

Figura D.106: Ridge, Fold 3, C2, ADASYN

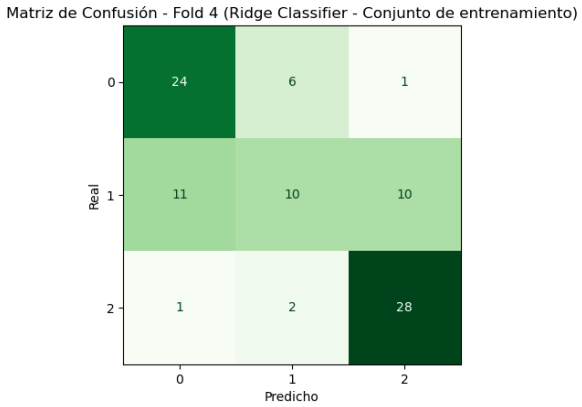

Figura D.107: Ridge, Fold 4, C2, ADASYN

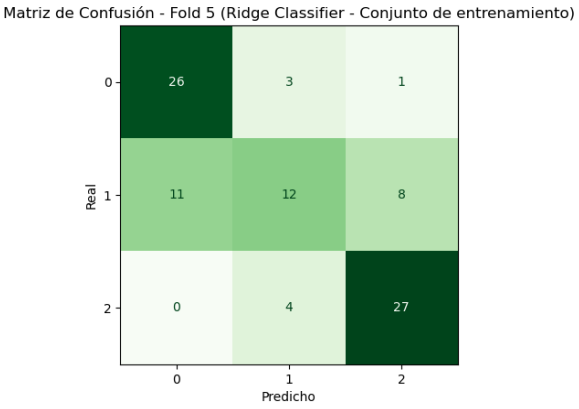

Figura D.107: Ridge, Fold 5, C2, ADASYN

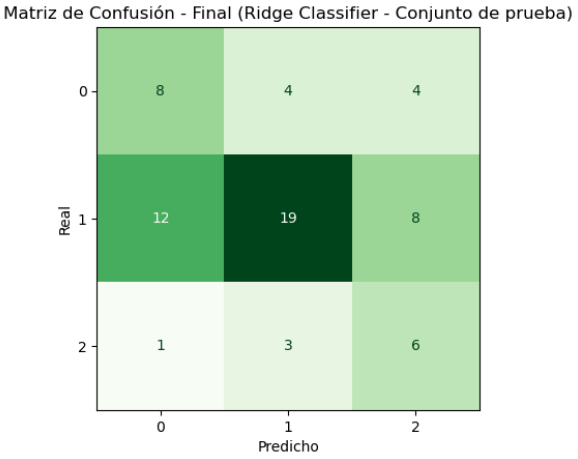

Figura D.108: Ridge, Final, C2, ADASYN

SDG

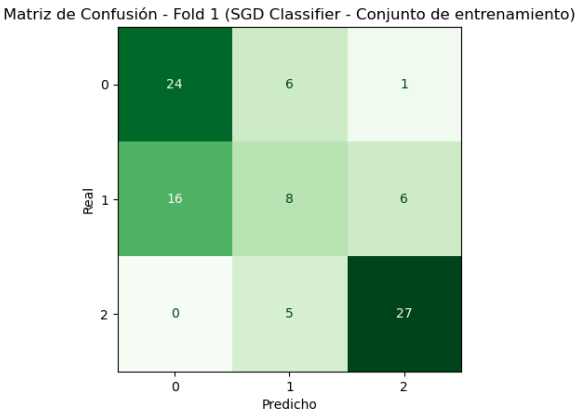

Figura D.109: SDG, Fold 1, C2, ADASYN

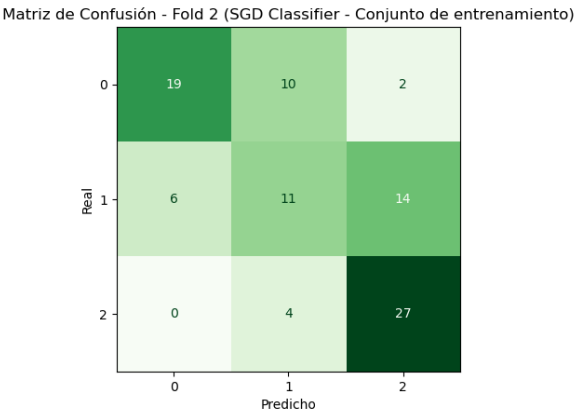

Figura D.110: SDG, Fold 2, C2, ADASYN

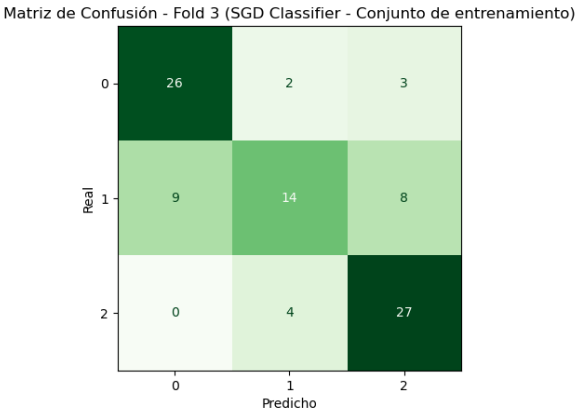

Figura D.110: SDG, Fold 3, C2, ADASYN

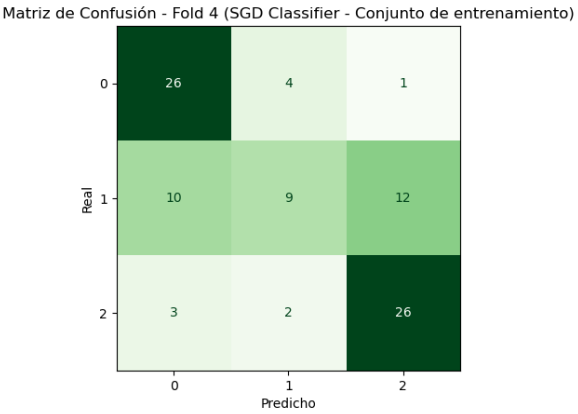

Figura D.111: SDG, Fold 4, C2, ADASYN

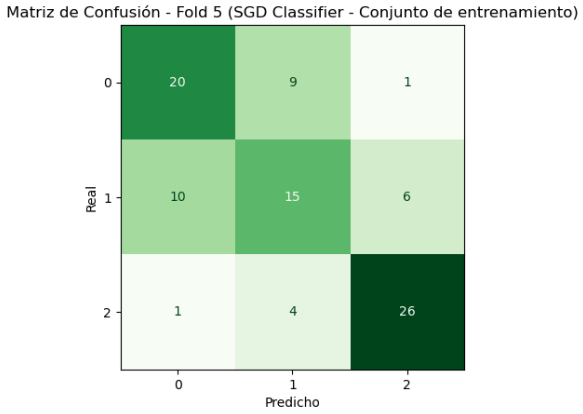

Figura D.111: SDG, Fold 5, C2, ADASYN

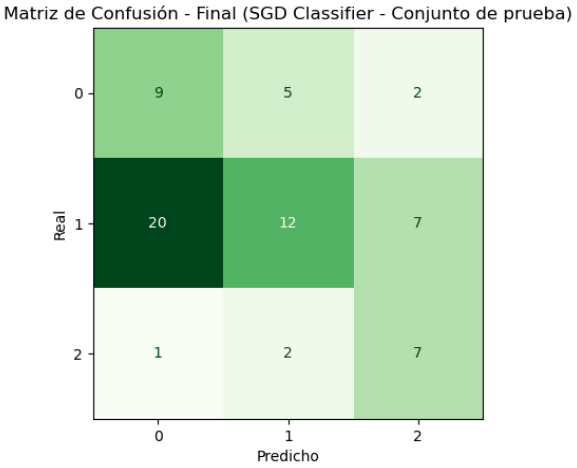

Figura D.112: SDG, Final, C2, ADASYN

Decision Tree

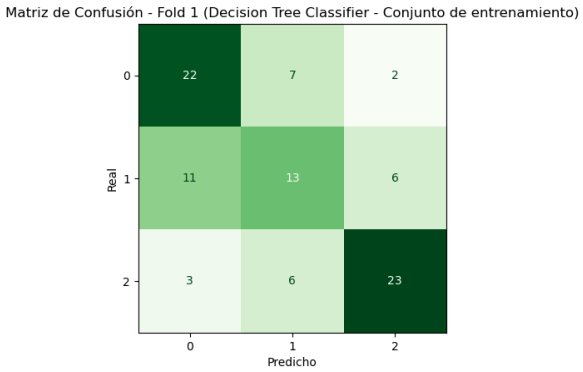

Figura D.113: Decision Tree, Fold 1, C2, ADASYN

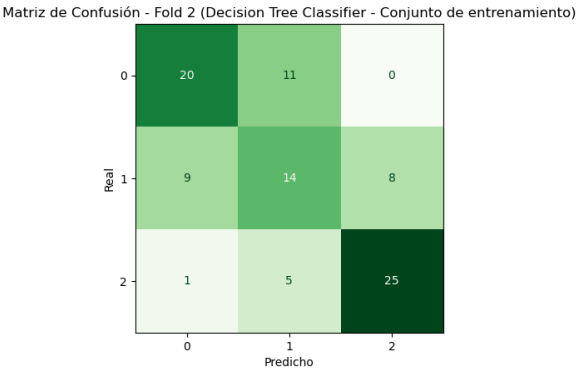

Figura D.114: Decision Tree, Fold 2, C2, ADASYN

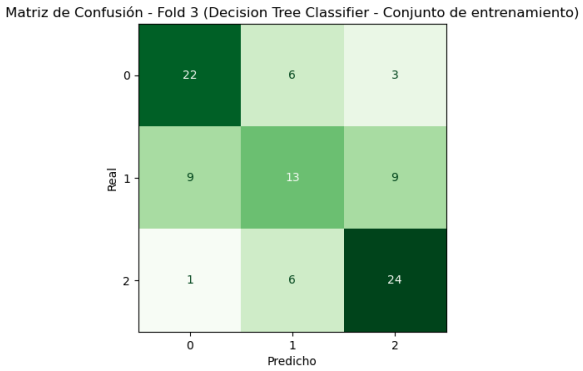

Figura D.114: Decision Tree, Fold 3, C2, ADASYN

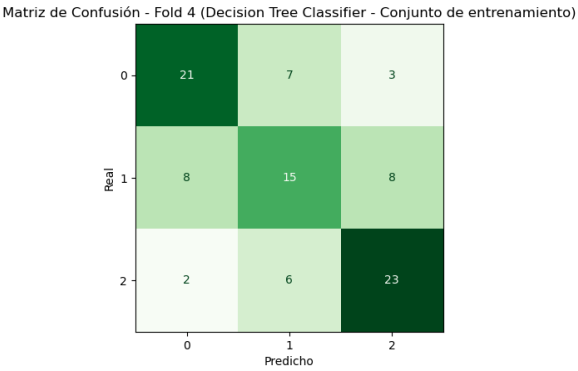

Figura D.115: Decision Tree, Fold 4, C2, ADASYN

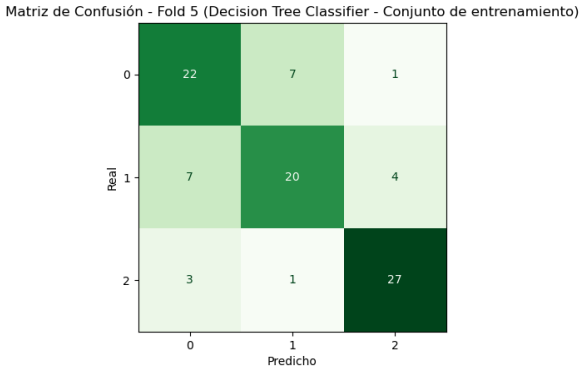

Figura D.115: Decision Tree, Fold 5, C2, ADASYN

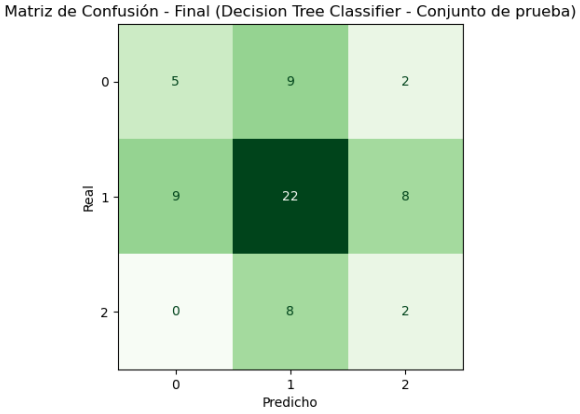

Figura D.116: Decision Tree, Final, C2, ADASYN

Random Forest

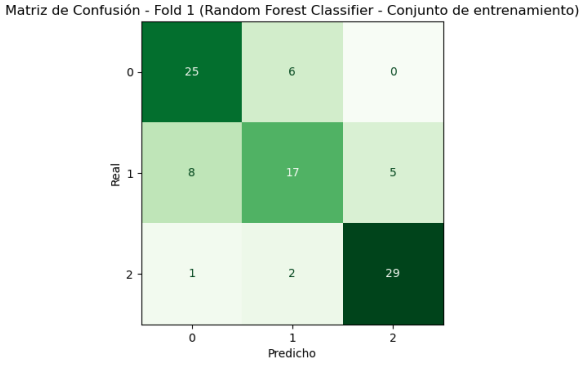

Figura D.117: Random Forest, Fold 1, C2, ADASYN

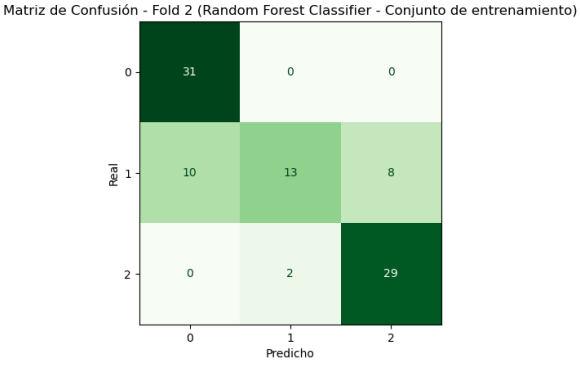

Figura D.118: Random Forest, Fold 2, C2, ADASYN

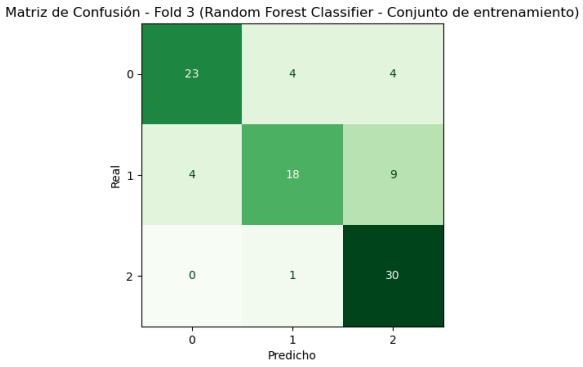

Figura D.118: Random Forest, Fold 3, C2, ADASYN

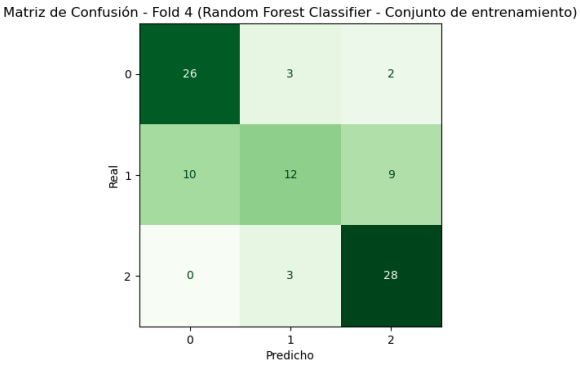

Figura D.119: Random Forest, Fold 4, C2, ADASYN

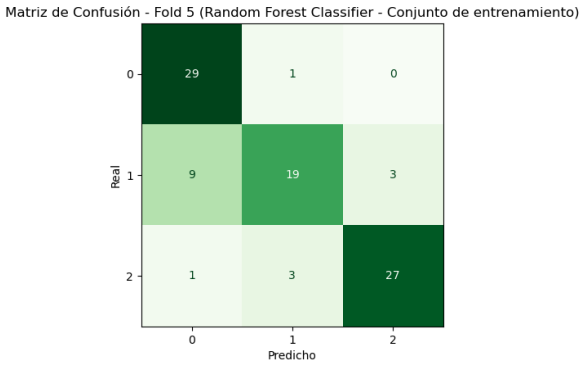

Figura D.119: Random Forest, Fold 5, C2, ADASYN

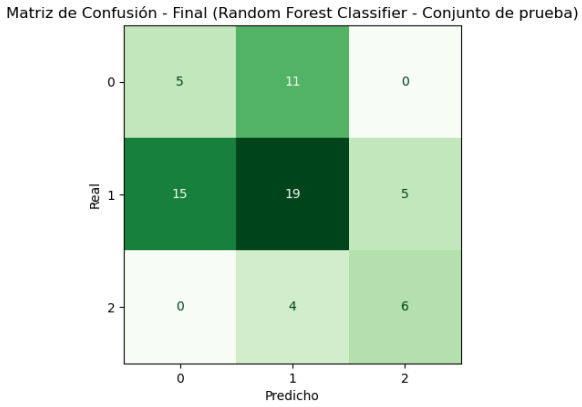

Figura D.120: Random Forest, Final, C2, ADASYN

AdaBoost

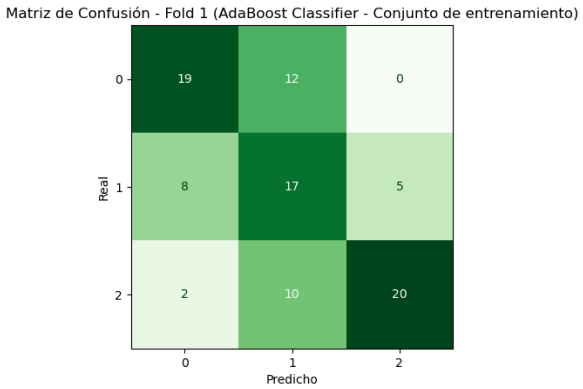

Figura D.121: AdaBoost, Fold 1, C2, ADASYN

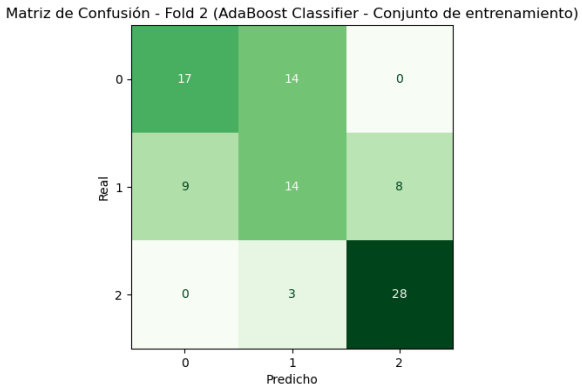

Figura D.122: AdaBoost, Fold 2, C2, ADASYN

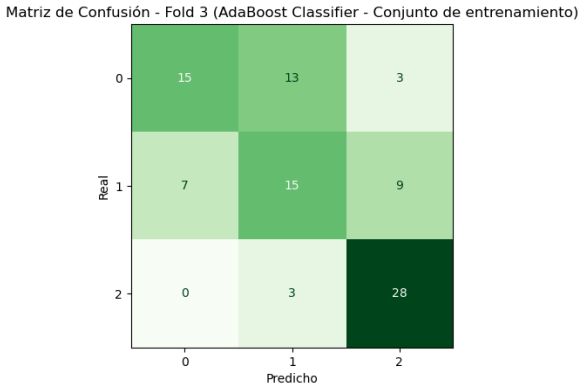

Figura D.122: AdaBoost, Fold 3, C2, ADASYN

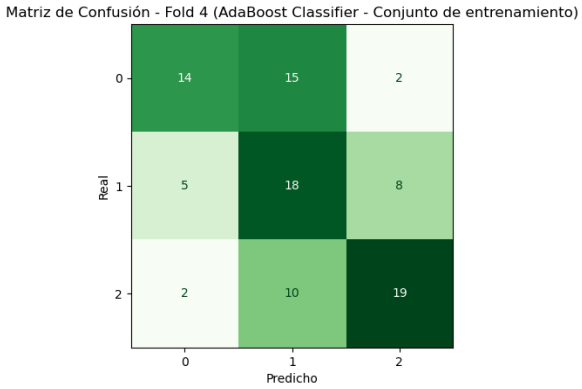

Figura D.123: AdaBoost, Fold 4, C2, ADASYN

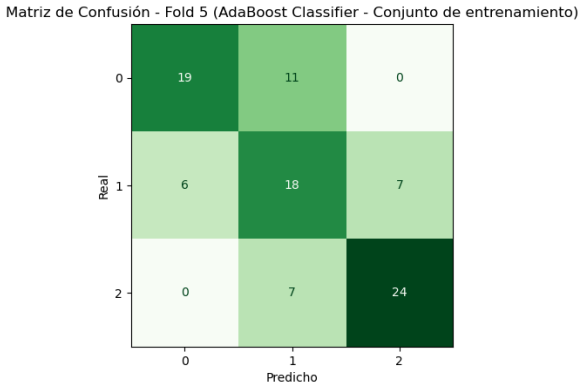

Figura D.123: AdaBoost, Fold 5, C2, ADASYN

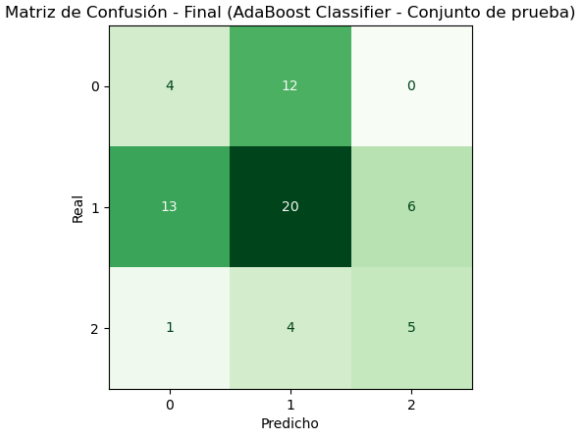

Figura D.124: AdaBoost, Final, C2, ADASYN

Gradient Boosting

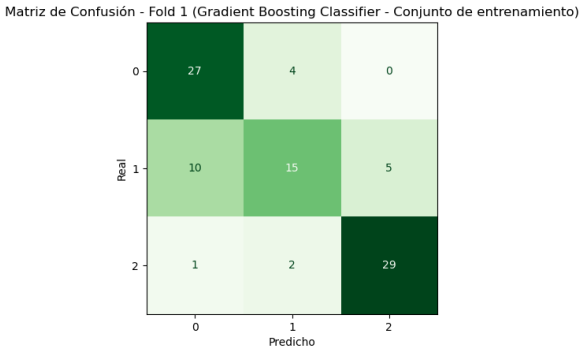

Figura D.125: Gradient Boosting, Fold 1, C2, ADASYN

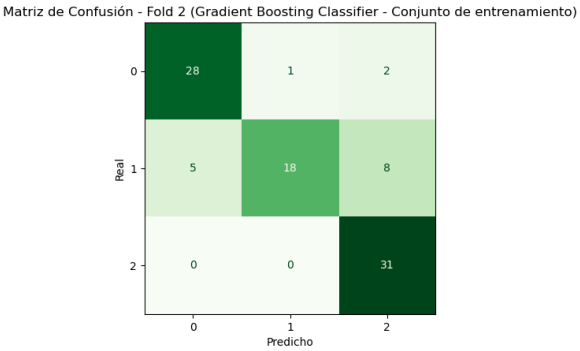

Figura D.126: Gradient Boosting, Fold 2, C2, ADASYN

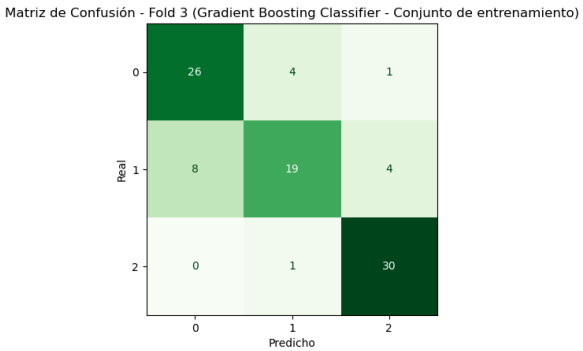

Figura D.126: Gradient Boosting, Fold 3, C2, ADASYN

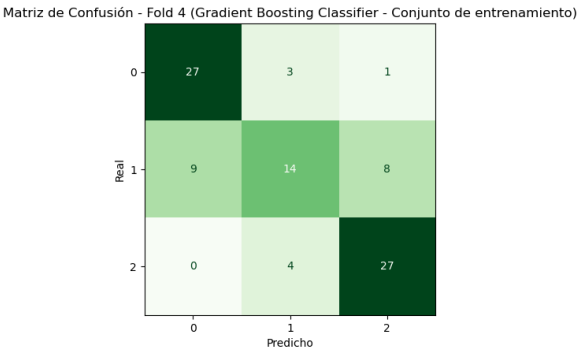

Figura D.127: Gradient Boosting, Fold 4, C2, ADASYN

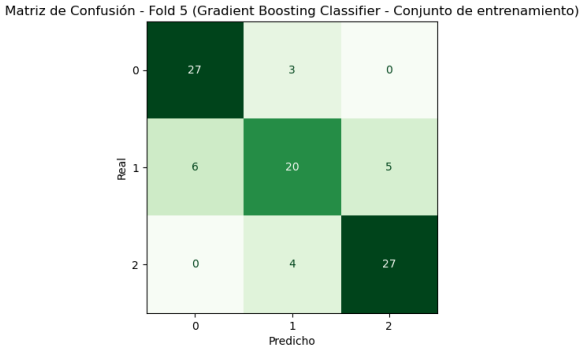

Figura D.127: Gradient Boosting, Fold 5, C2, ADASYN

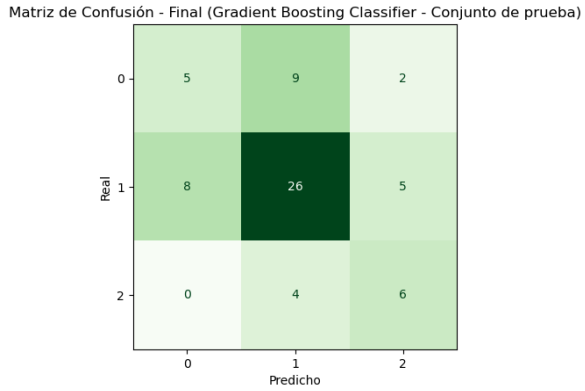

Figura D.128: Gradient Boosting, Final, C2, ADASYN

SVC

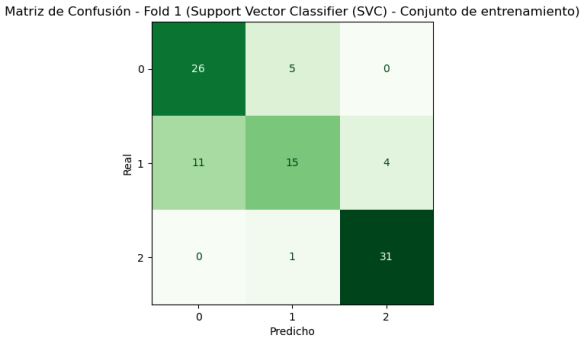

Figura D.129: SVC, Fold 1, C2, ADASYN

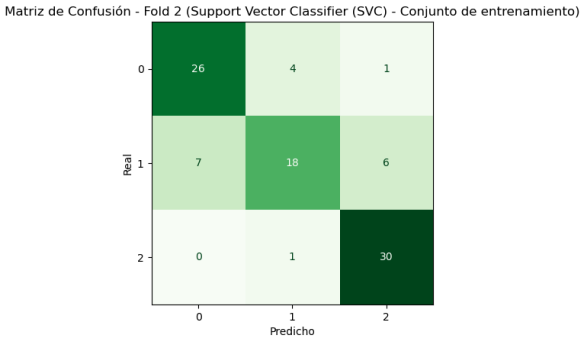

Figura D.130: SVC, Fold 2, C2, ADASYN

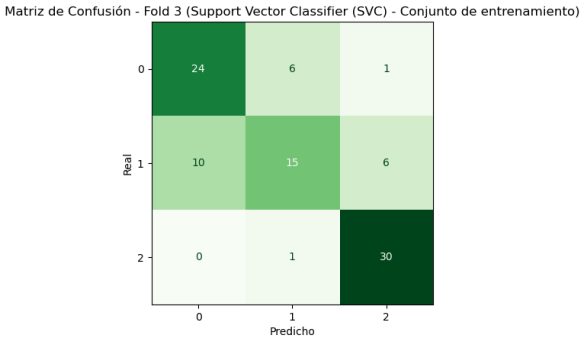

Figura D.130: SVC, Fold 3, C2, ADASYN

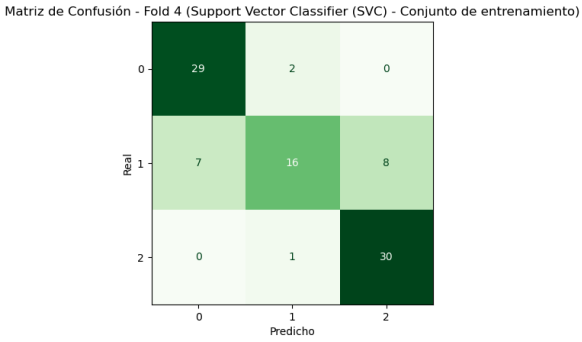

Figura D.131: SVC, Fold 4, C2, ADASYN

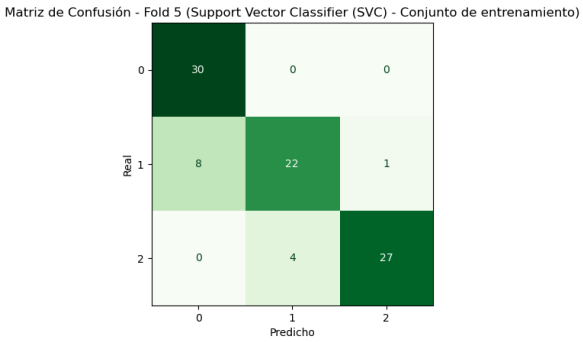

Figura D.131: SVC, Fold 5, C2, ADASYN

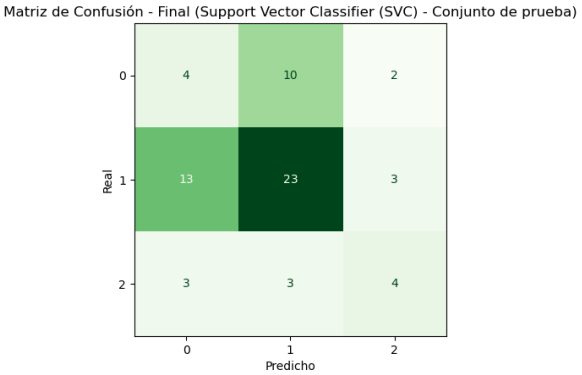

Figura D.132: SVC, Final, C2, ADASYN

k-Nearest Neighbors

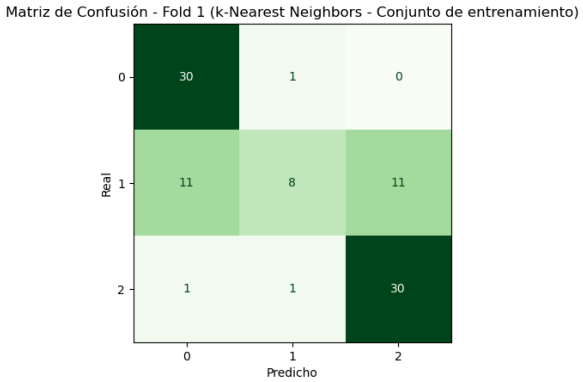

Figura D.133: k-Nearest Neighbors, Fold 1, C2, ADASYN

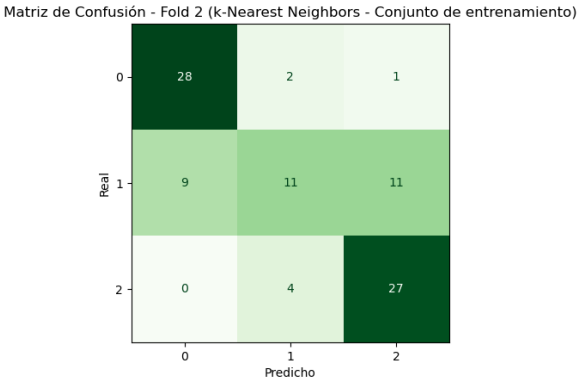

Figura D.134: k-Nearest Neighbors, Fold 2, C2, ADASYN

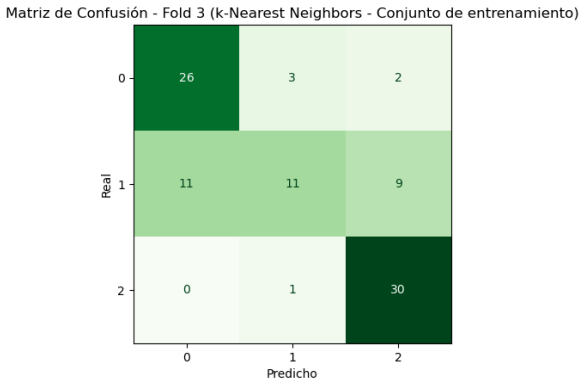

Figura D.134: k-Nearest Neighbors, Fold 3, C2, ADASYN

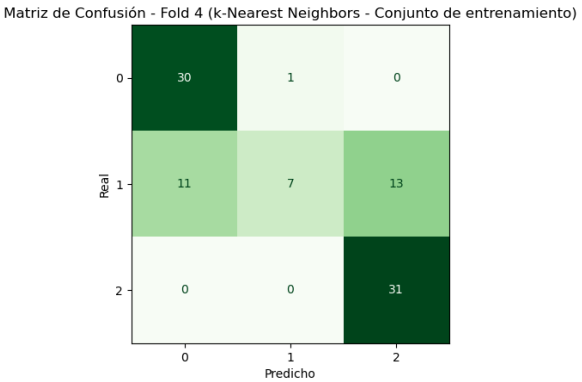

Figura D.135: k-Nearest Neighbors, Fold 4, C2, ADASYN

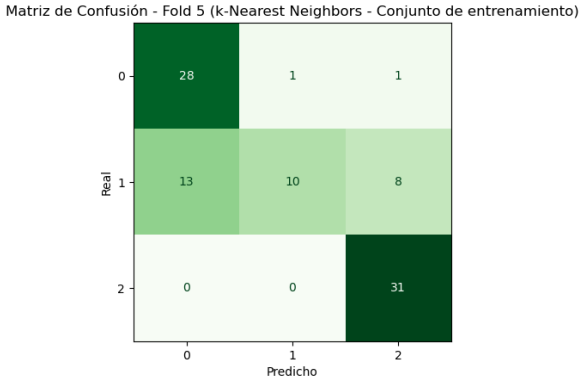

Figura D.135: k-Nearest Neighbors, Fold 5, C2, ADASYN

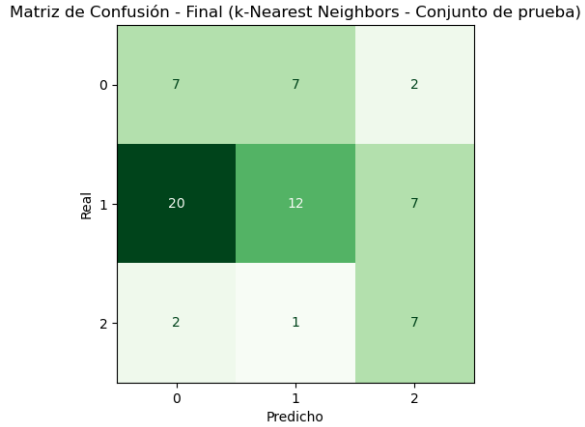

Figura D.136: k-Nearest Neighbors, Final, C2, ADASYN

Gaussian Naive Bayes

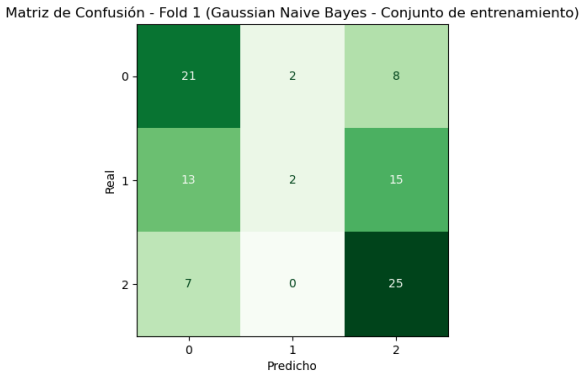

Figura D.137: Gaussian Naive Bayes, Fold 1, C2, ADASYN

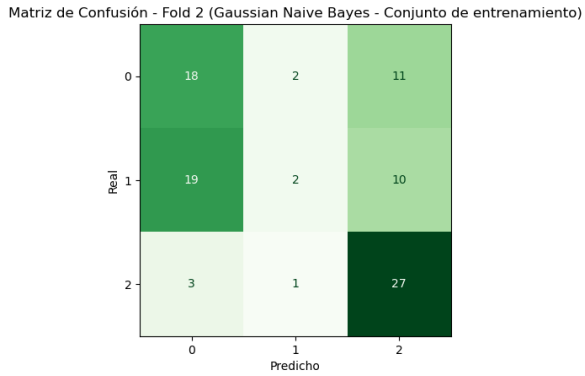

Figura D.138: Gaussian Naive Bayes, Fold 2, C2, ADASYN

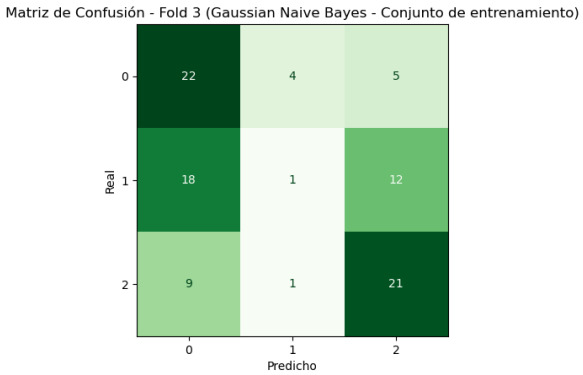

Figura D.138: Gaussian Naive Bayes, Fold 3, C2, ADASYN

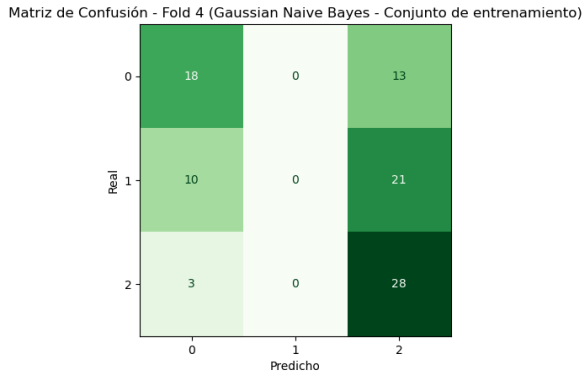

Figura D.139: Gaussian Naive Bayes, Fold 4, C2, ADASYN

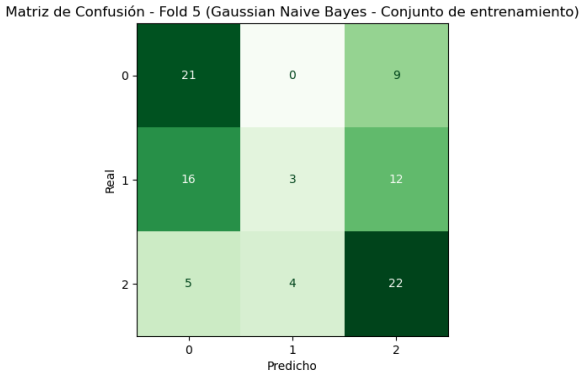

Figura D.139: Gaussian Naive Bayes, Fold 5, C2, ADASYN

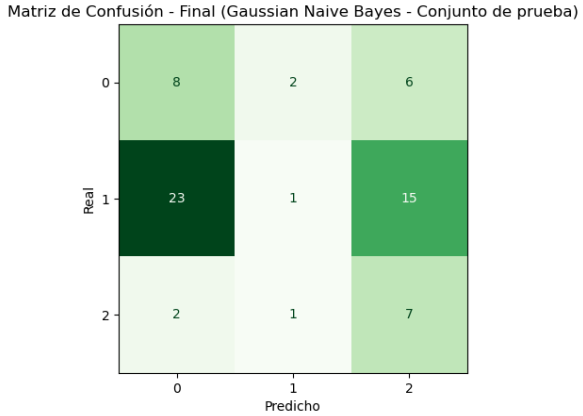

Figura D.140: Gaussian Naive Bayes, Final, C2, ADASYN

MLP

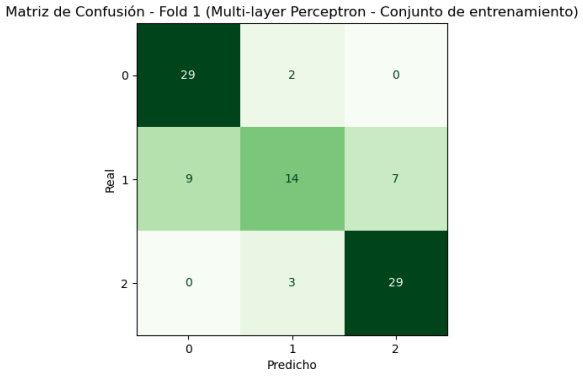

Figura D.141: MLP, Fold 1, C2, ADASYN

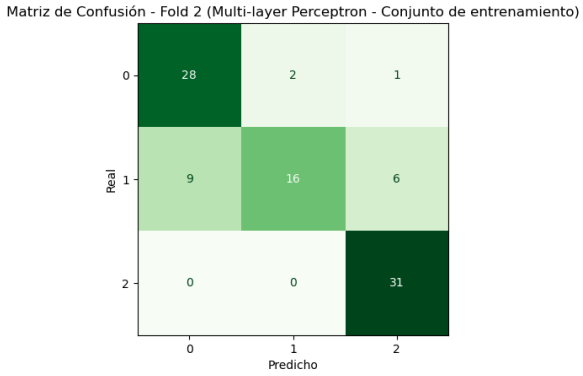

Figura D.142: MLP, Fold 2, C2, ADASYN

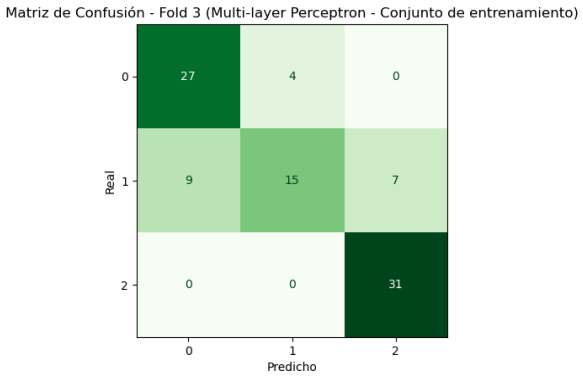

Figura D.142: MLP, Fold 3, C2, ADASYN

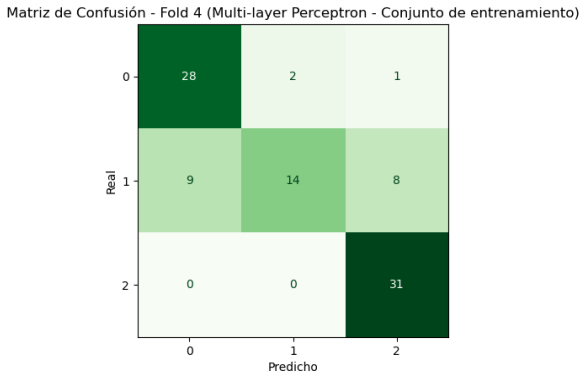

Figura D.143: MLP, Fold 4, C2, ADASYN

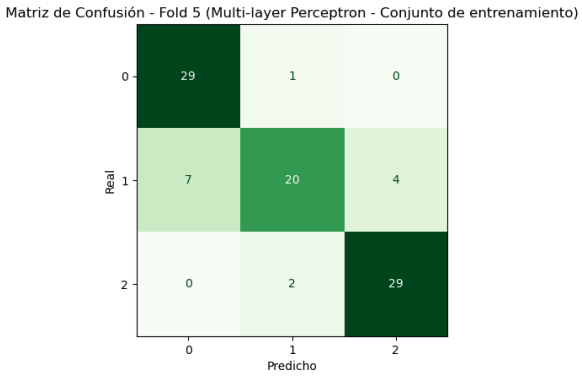

Figura D.143: MLP, Fold 5, C2, ADASYN

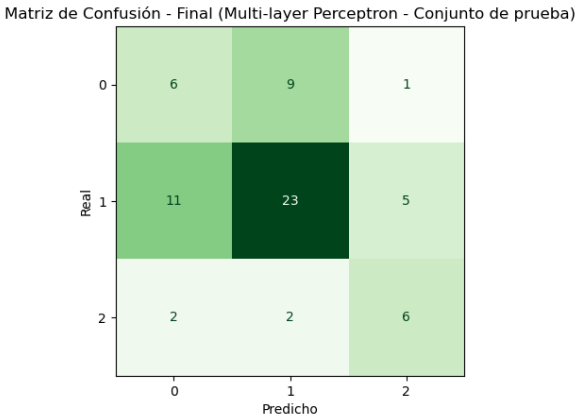

Figura D.144: MLP, Final, C2, ADASYN

D.2.2. SMOTE

Logistic Regression

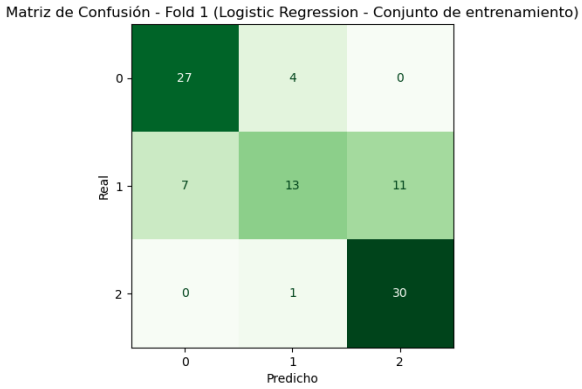

Figura D.145: Logistic Regression, Fold 1, C2, SMOTE

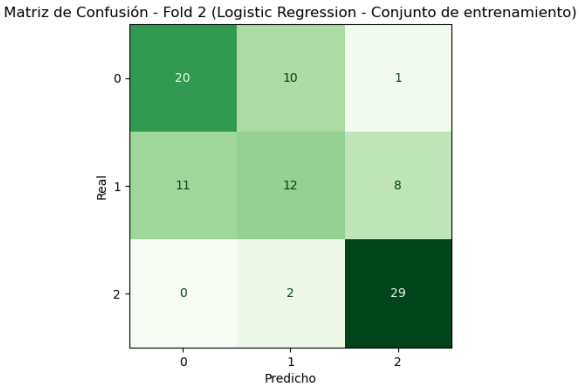

Figura D.146: Logistic Regression, Fold 2, C2, SMOTE

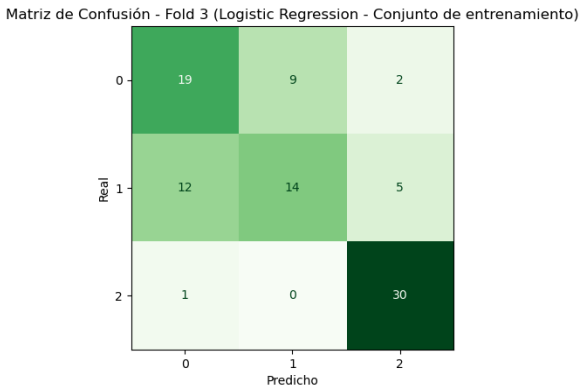

Figura D.146: Logistic Regression, Fold 3, C2, SMOTE

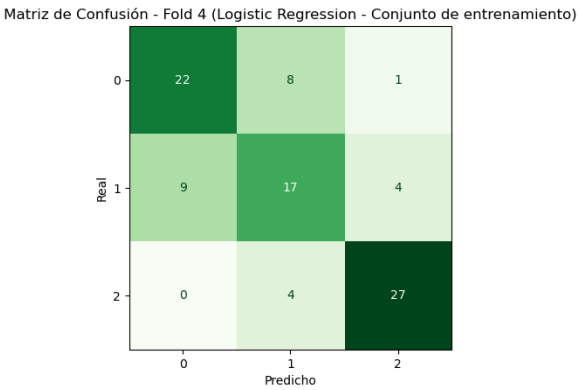

Figura D.147: Logistic Regression, Fold 4, C2, SMOTE

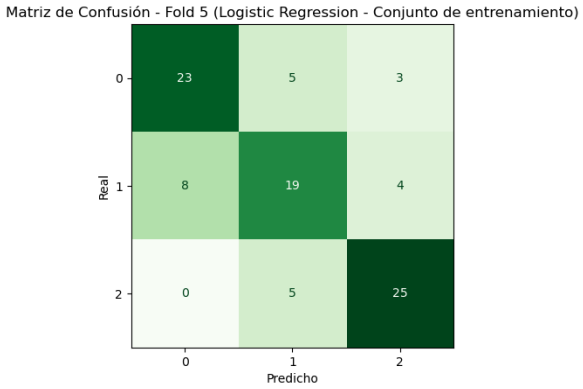

Figura D.147: Logistic Regression, Fold 5, C2, SMOTE

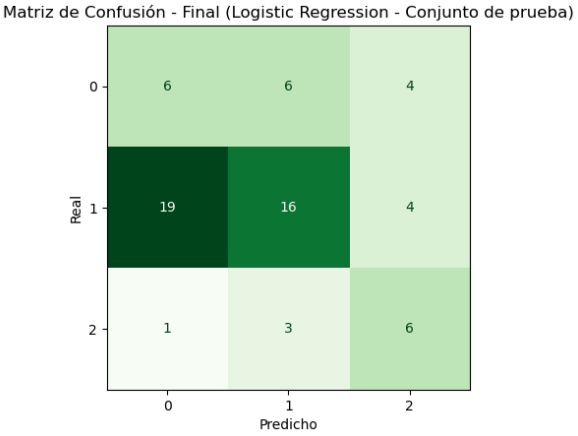

Figura D.148: Logistic Regression, Final, C2, SMOTE

Logistic Regression Lasso

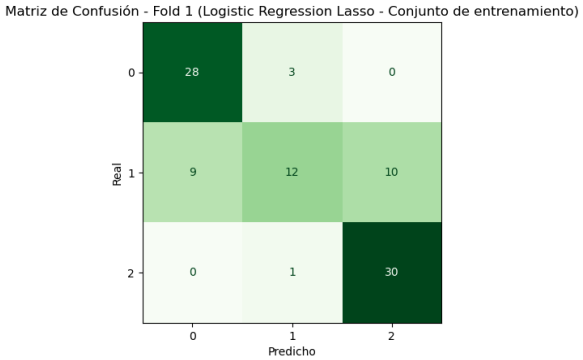

Figura D.149: Logistic Regression Lasso, Fold 1, C2, SMOTE

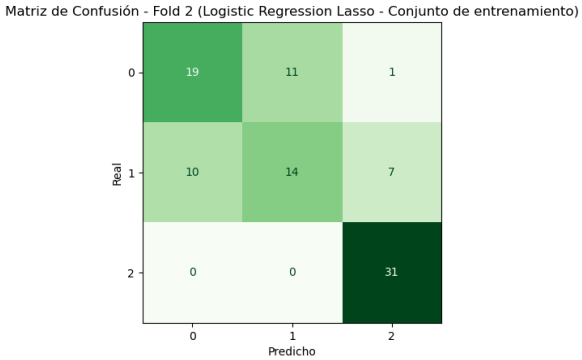

Figura D.150: Logistic Regression Lasso, Fold 2, C2, SMOTE

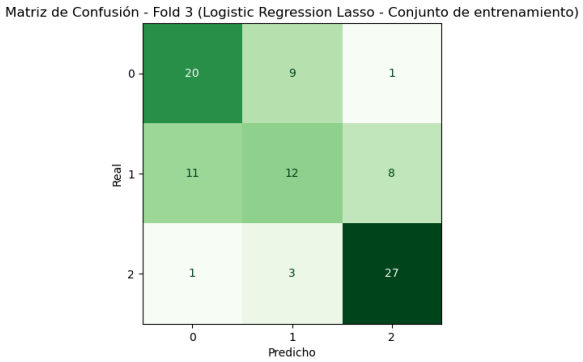

Figura D.150: Logistic Regression Lasso, Fold 3, C2, SMOTE

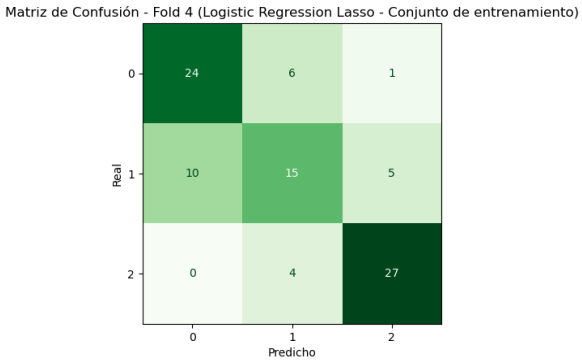

Figura D.151: Logistic Regression Lasso, Fold 4, C2, SMOTE

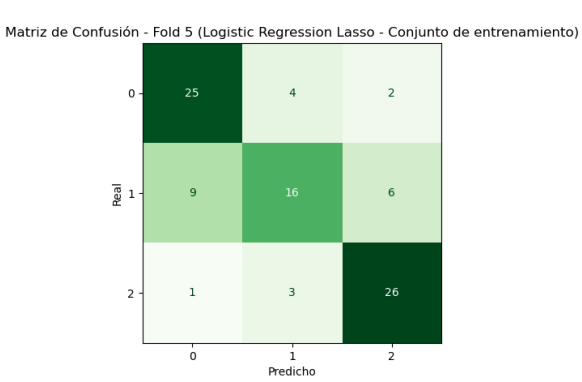

Figura D.151: Logistic Regression Lasso, Fold 5, C2, SMOTE

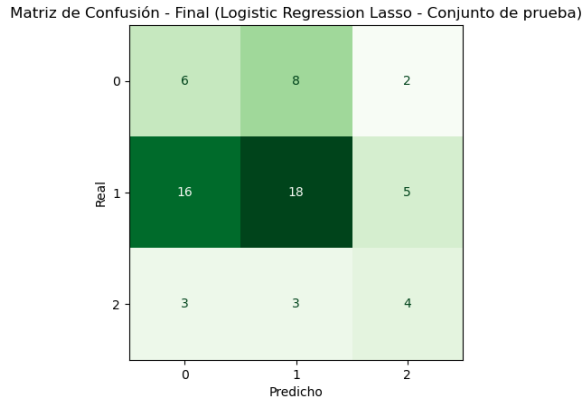

Figura D.152: Logistic Regression Lasso, Final, C2, SMOTE

Ridge

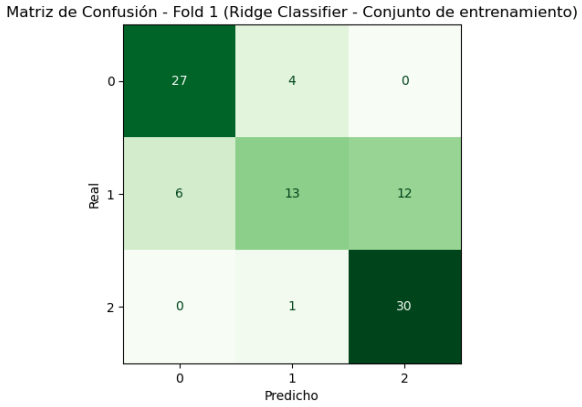

Figura D.153: Ridge, Fold 1, C2, SMOTE

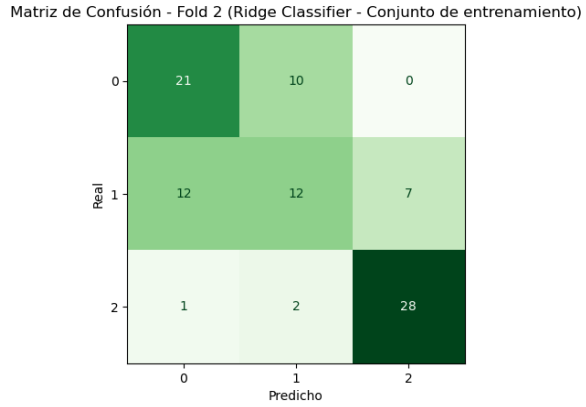

Figura D.154: Ridge, Fold 2, C2, SMOTE

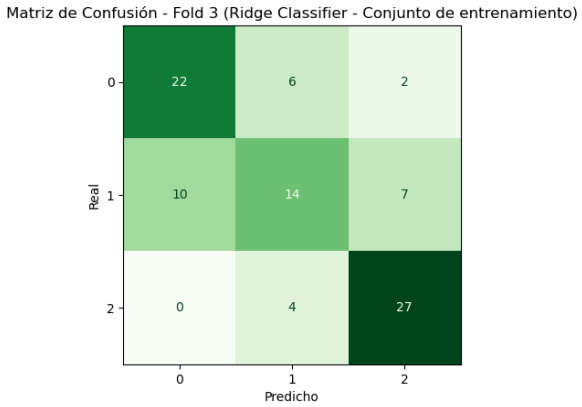

Figura D.154: Ridge, Fold 3, C2, SMOTE

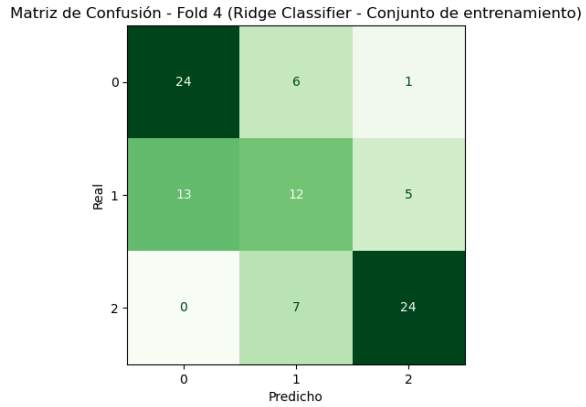

Figura D.155: Ridge, Fold 4, C2, SMOTE

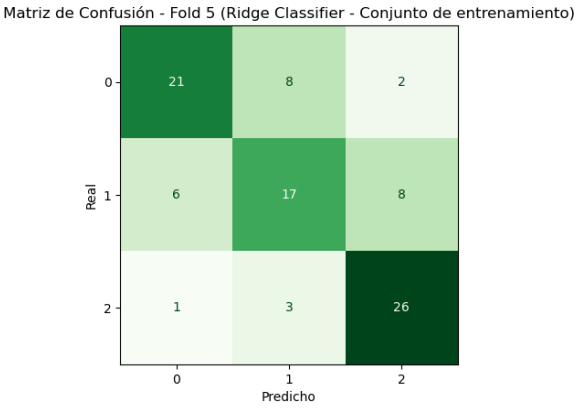

Figura D.155: Ridge, Fold 5, C2, SMOTE

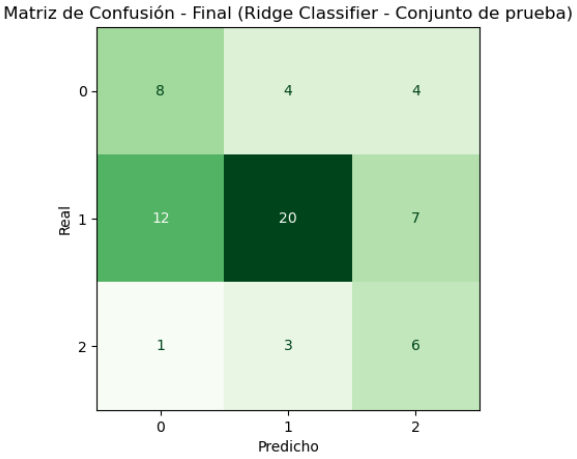

Figura D.156: Ridge, Final, C2, SMOTE

SDG

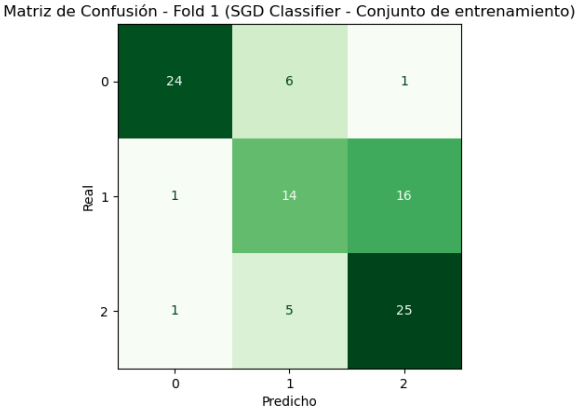

Figura D.157: SDG, Fold 1, C2, SMOTE

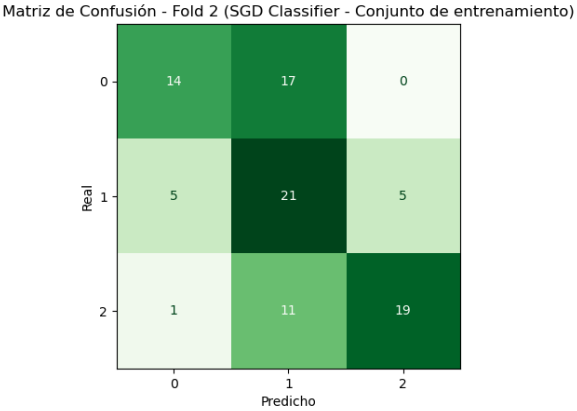

Figura D.158: SDG, Fold 2, C2, SMOTE

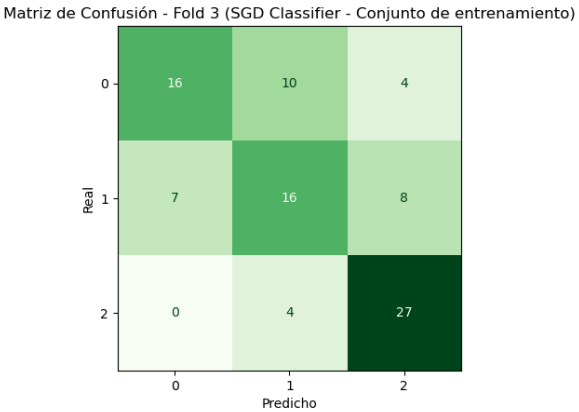

Figura D.158: SDG, Fold 3, C2, SMOTE

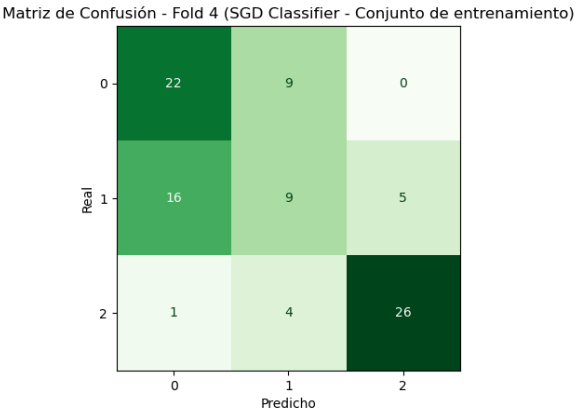

Figura D.159: SDG, Fold 4, C2, SMOTE

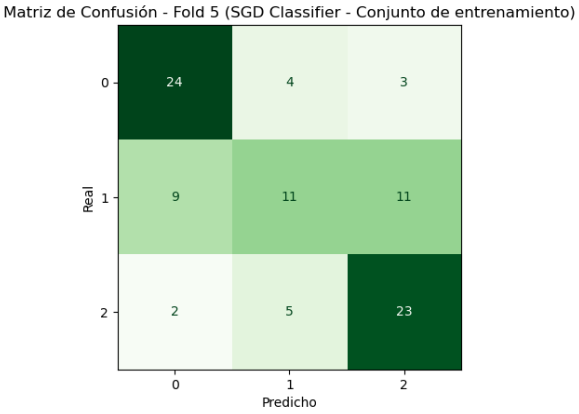

Figura D.159: SDG, Fold 5, C2, SMOTE

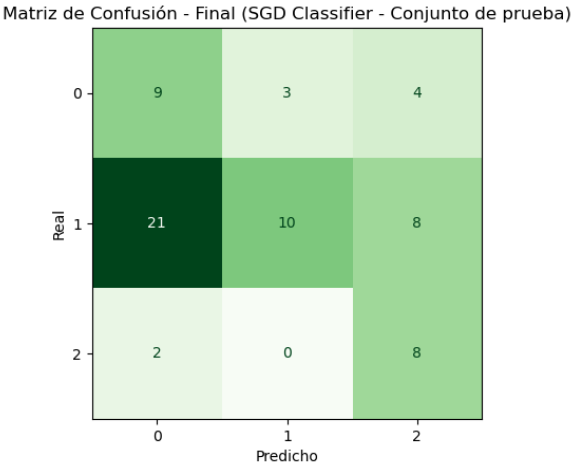

Figura D.160: SDG, Final, C2, SMOTE

Decision Tree

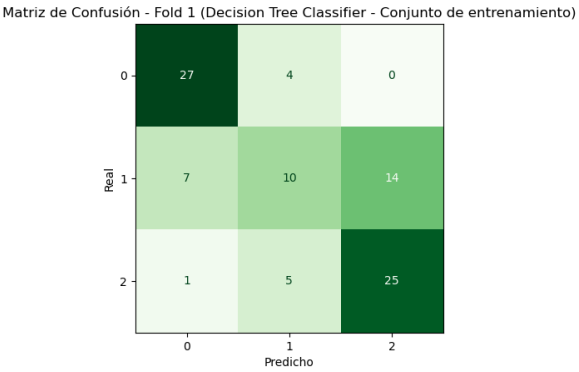

Figura D.161: Decision Tree, Fold 1, C2, SMOTE

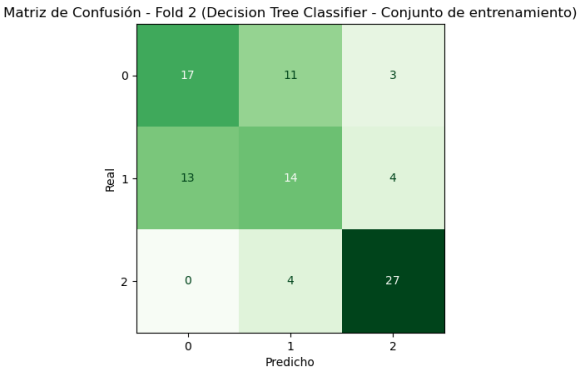

Figura D.162: Decision Tree, Fold 2, C2, SMOTE

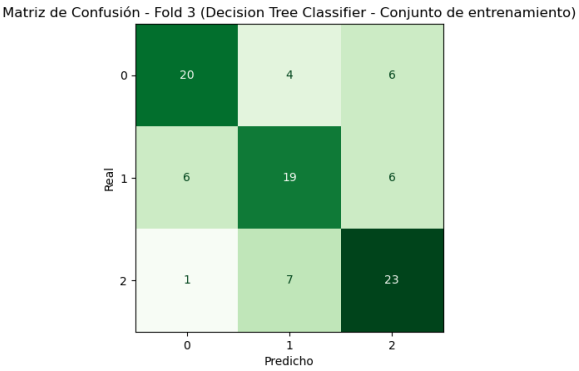

Figura D.162: Decision Tree, Fold 3, C2, SMOTE

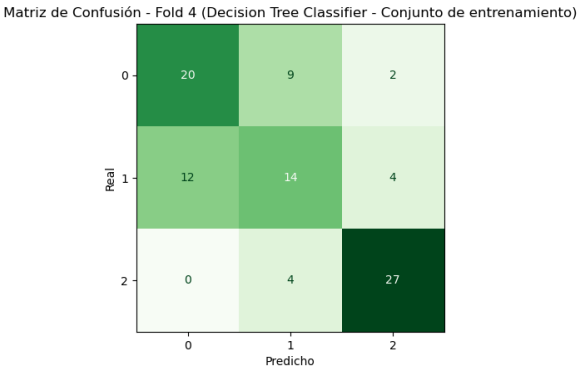

Figura D.163: Decision Tree, Fold 4, C2, SMOTE

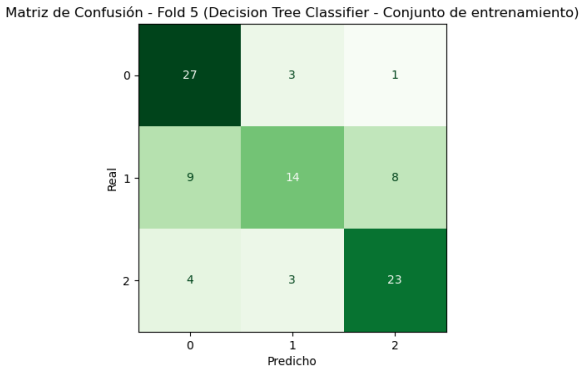

Figura D.163: Decision Tree, Fold 5, C2, SMOTE

TFG/imgs/anexo 4/dt_final_c2_s.png

Figura D.164: Decision Tree, Final, C2, SMOTE

Random Forest

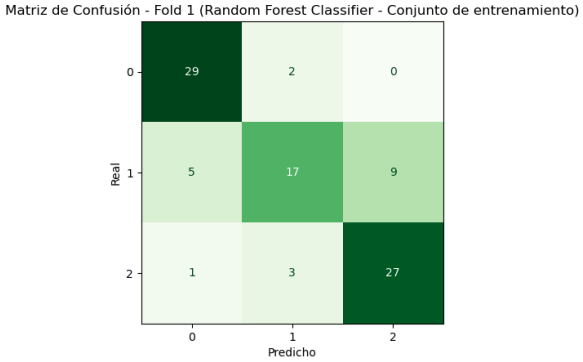

Figura D.165: Random Forest, Fold 1, C2, SMOTE

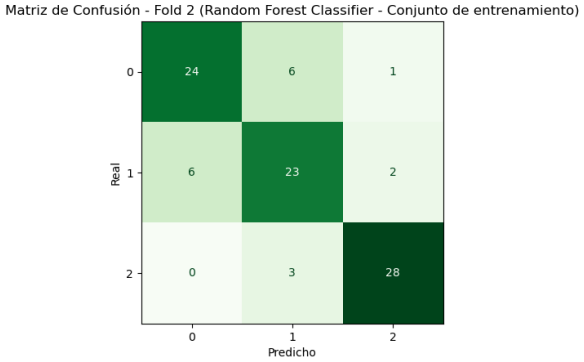

Figura D.166: Random Forest, Fold 2, C2, SMOTE

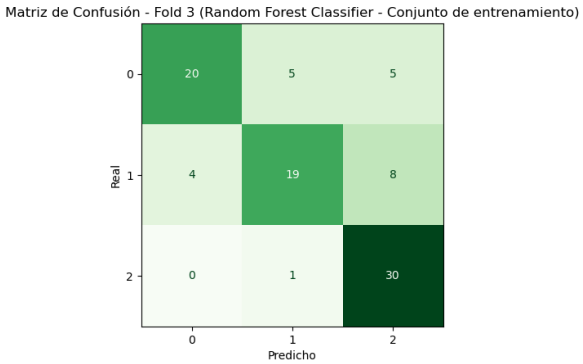

Figura D.166: Random Forest, Fold 3, C2, SMOTE

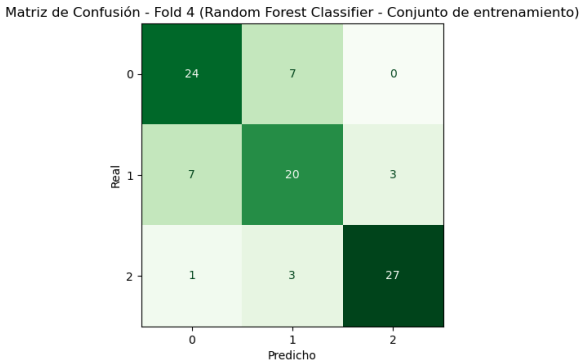

Figura D.167: Random Forest, Fold 4, C2, SMOTE

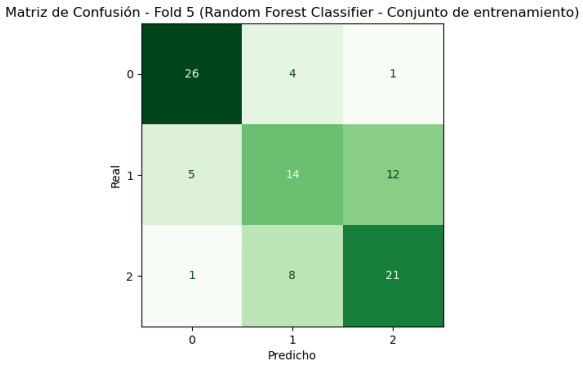

Figura D.167: Random Forest, Fold 5, C2, SMOTE

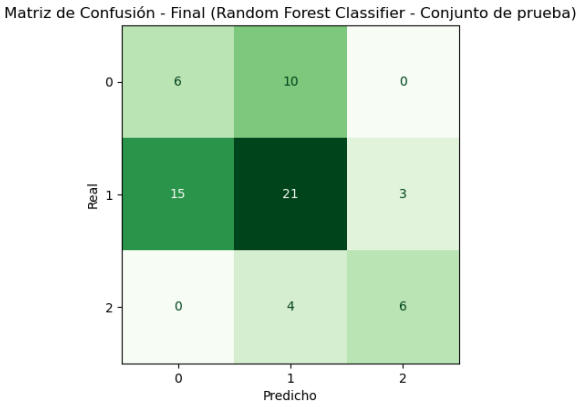

Figura D.168: Random Forest, Final, C2, SMOTE

AdaBoost

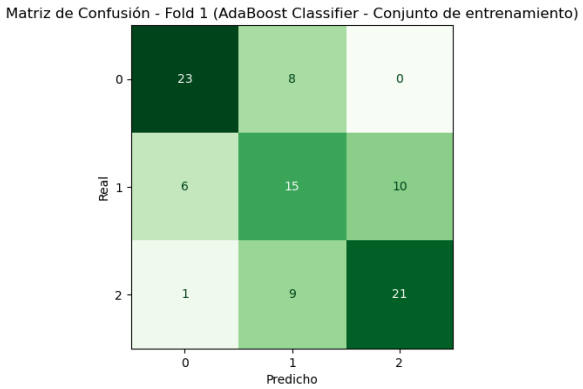

Figura D.169: AdaBoost, Fold 1, C2, SMOTE

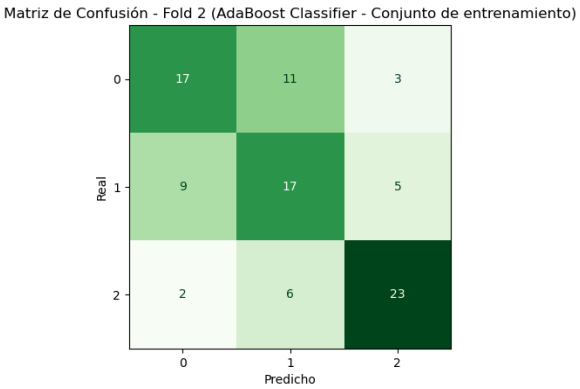

Figura D.170: AdaBoost, Fold 2, C2, SMOTE

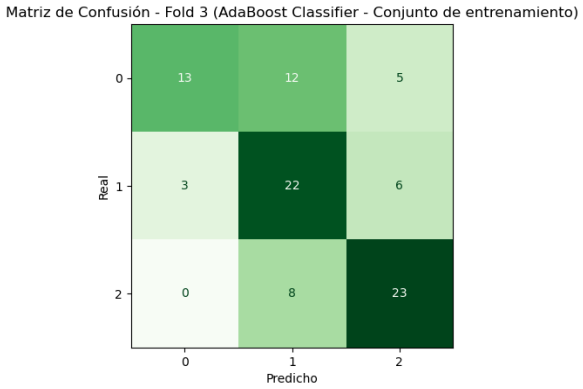

Figura D.170: AdaBoost, Fold 3, C2, SMOTE

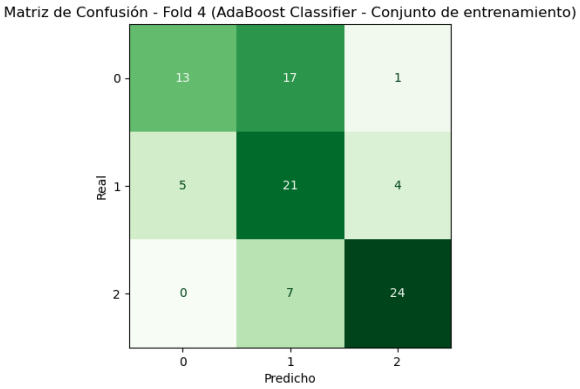

Figura D.171: AdaBoost, Fold 4, C2, SMOTE

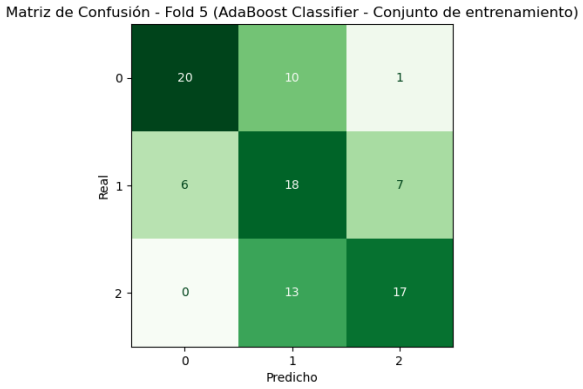

Figura D.171: AdaBoost, Fold 5, C2, SMOTE

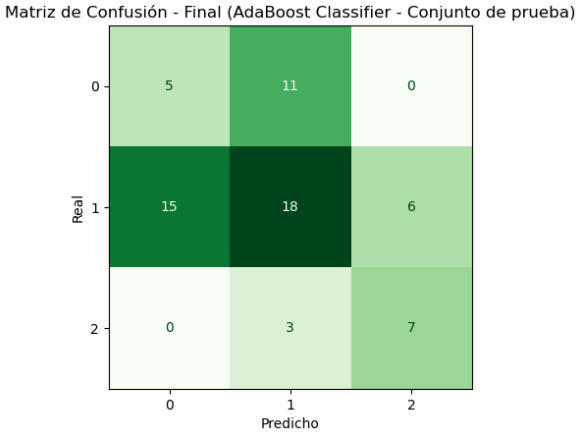

Figura D.172: AdaBoost, Final, C2, SMOTE

Gradient Boosting

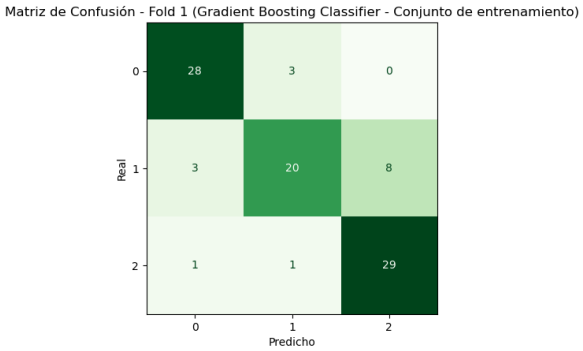

Figura D.173: Gradient Boosting, Fold 1, C2, SMOTE

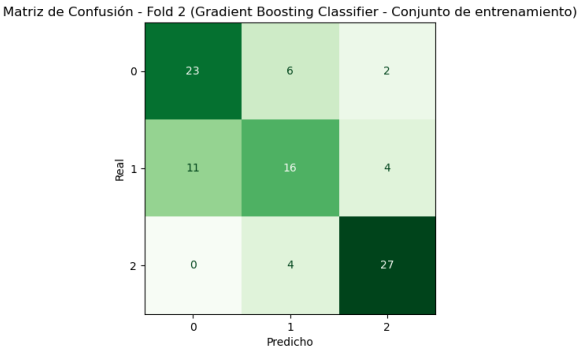

Figura D.174: Gradient Boosting, Fold 2, C2, SMOTE

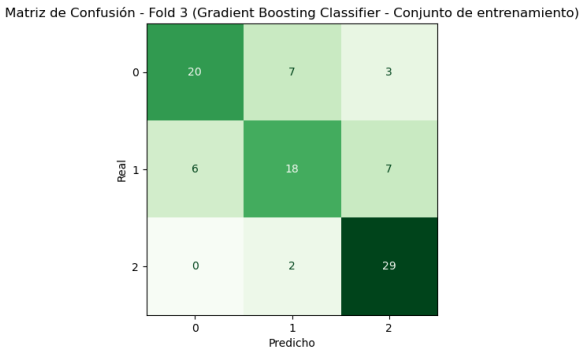

Figura D.174: Gradient Boosting, Fold 3, C2, SMOTE

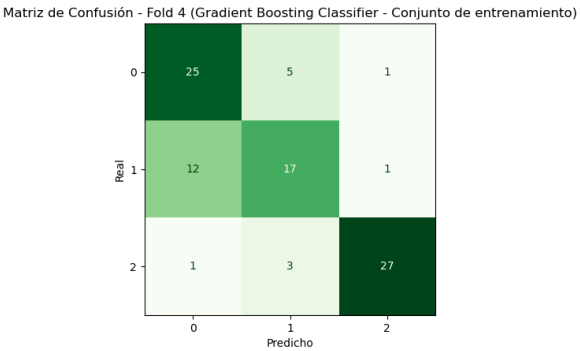

Figura D.175: Gradient Boosting, Fold 4, C2, SMOTE

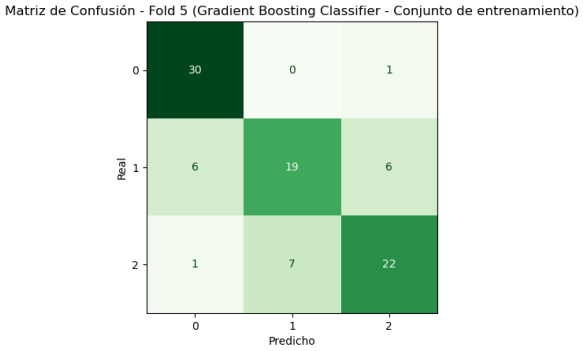

Figura D.175: Gradient Boosting, Fold 5, C2, SMOTE

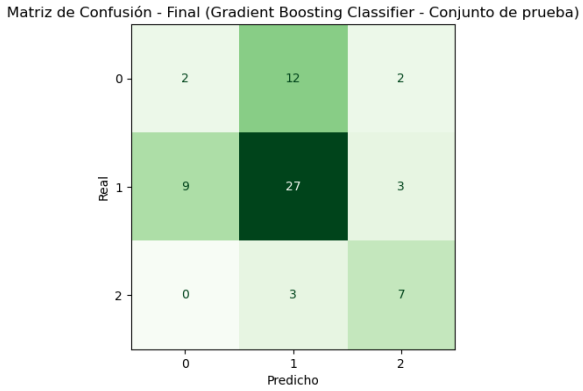

Figura D.176: Gradient Boosting, Final, C2, SMOTE

SVC

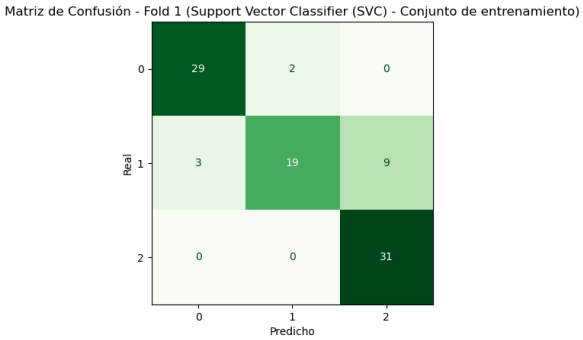

Figura D.177: SVC, Fold 1, C2, SMOTE

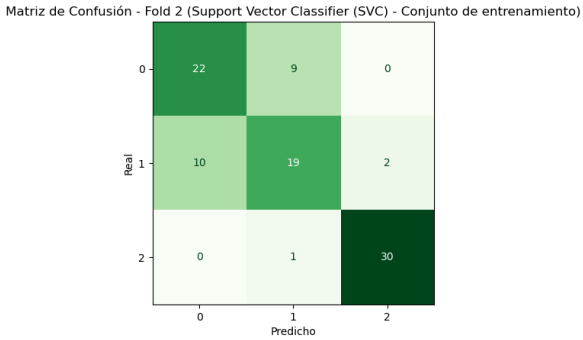

Figura D.178: SVC, Fold 2, C2, SMOTE

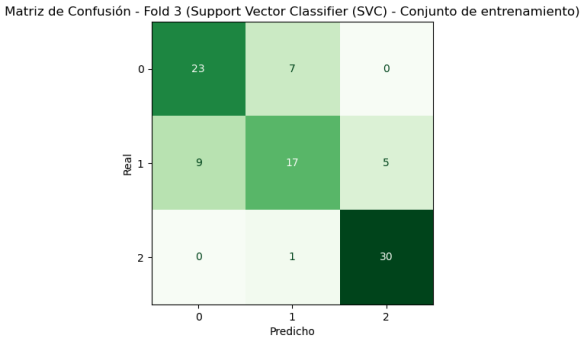

Figura D.178: SVC, Fold 3, C2, SMOTE

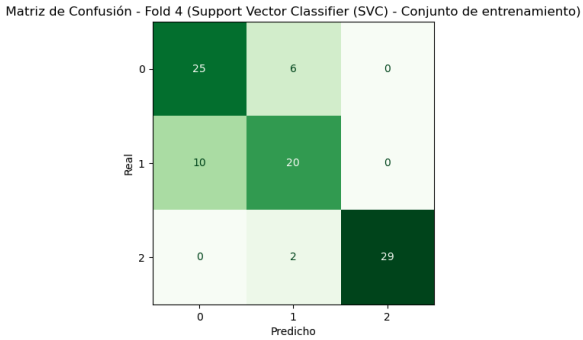

Figura D.179: SVC, Fold 4, C2, SMOTE

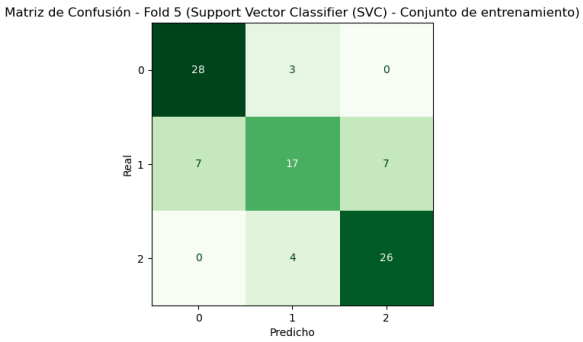

Figura D.179: SVC, Fold 5, C2, SMOTE

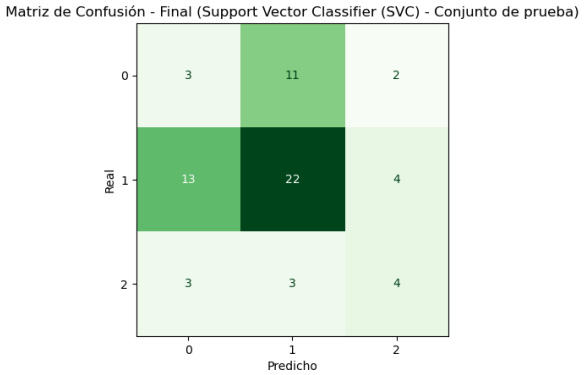

Figura D.180: SVC, Final, C2, SMOTE

k-Nearest Neighbors

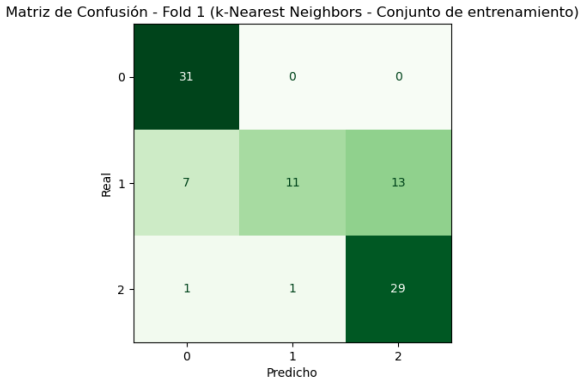

Figura D.181: k-Nearest Neighbors, Fold 1, C2, SMOTE

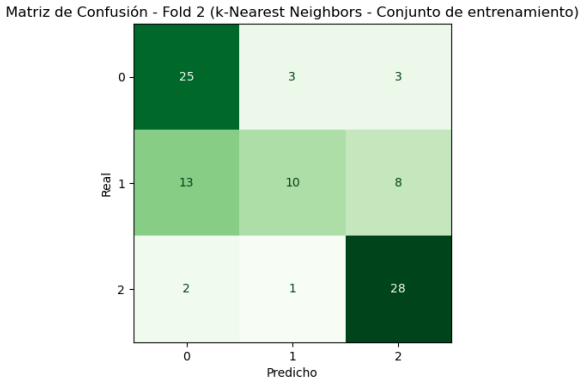

Figura D.182: k-Nearest Neighbors, Fold 2, C2, SMOTE

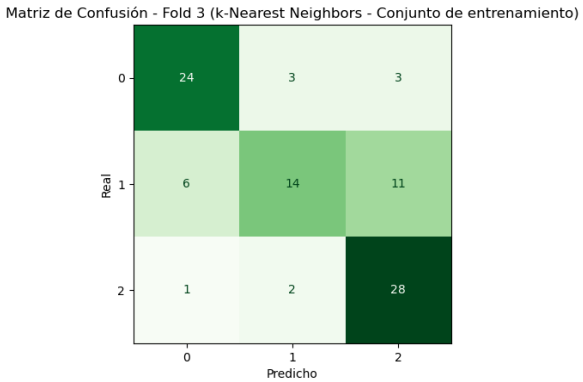

Figura D.182: k-Nearest Neighbors, Fold 3, C2, SMOTE

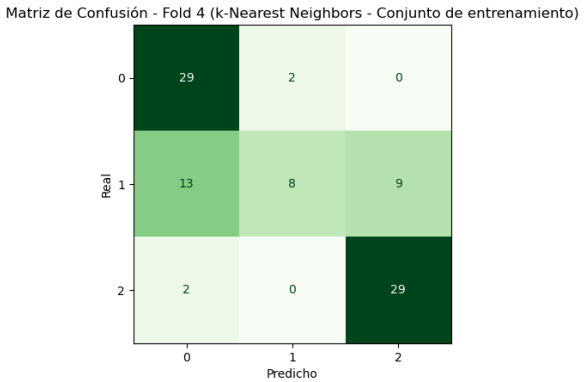

Figura D.183: k-Nearest Neighbors, Fold 4, C2, SMOTE

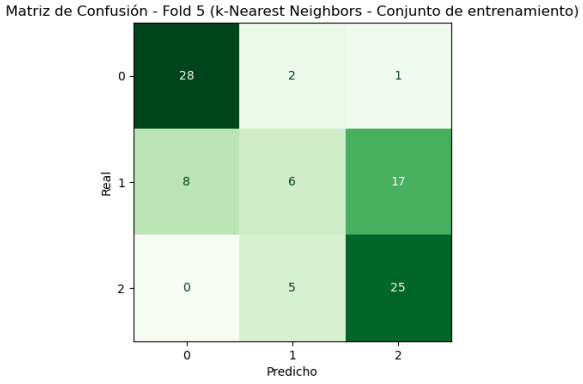

Figura D.183: k-Nearest Neighbors, Fold 5, C2, SMOTE

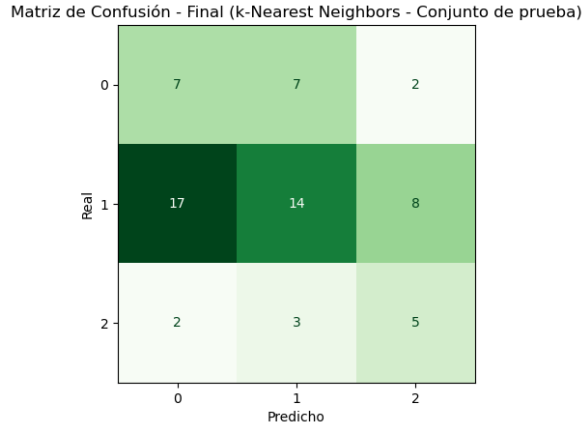

Figura D.184: k-Nearest Neighbors, Final, C2, SMOTE

Gaussian Naive Bayes

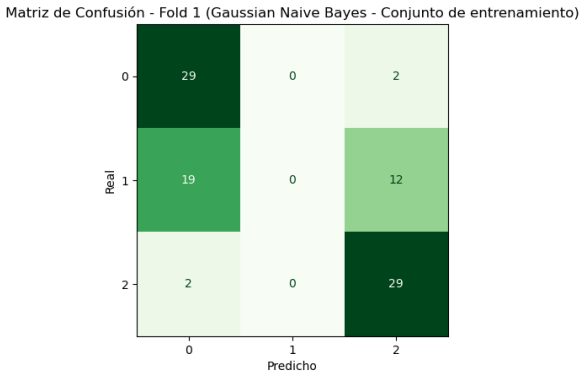

Figura D.185: Gaussian Naive Bayes, Fold 1, C2, SMOTE

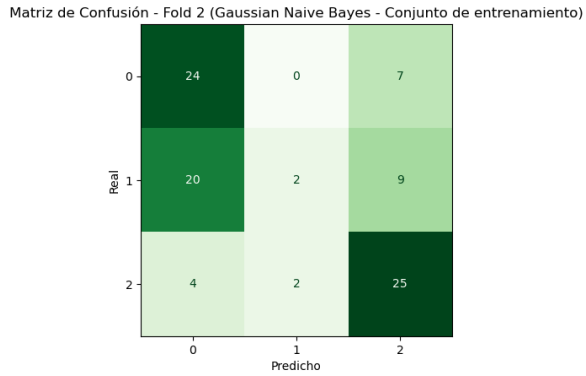

Figura D.186: Gaussian Naive Bayes, Fold 2, C2, SMOTE

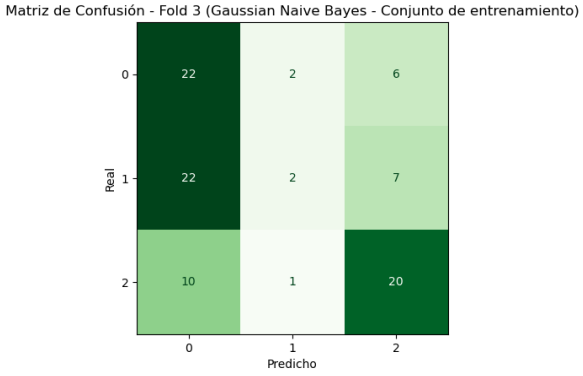

Figura D.186: Gaussian Naive Bayes, Fold 3, C2, SMOTE

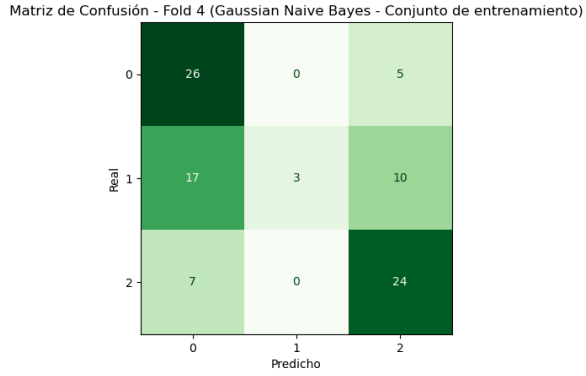

Figura D.187: Gaussian Naive Bayes, Fold 4, C2, SMOTE

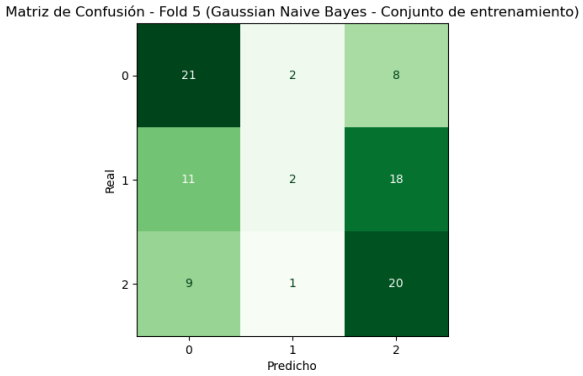

Figura D.187: Gaussian Naive Bayes, Fold 5, C2, SMOTE

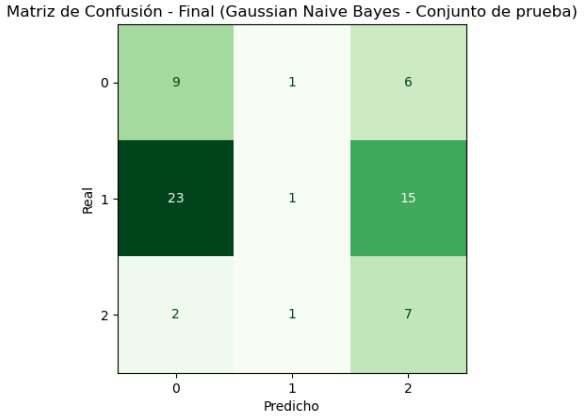

Figura D.188: Gaussian Naive Bayes, Final, C2, SMOTE

MLP

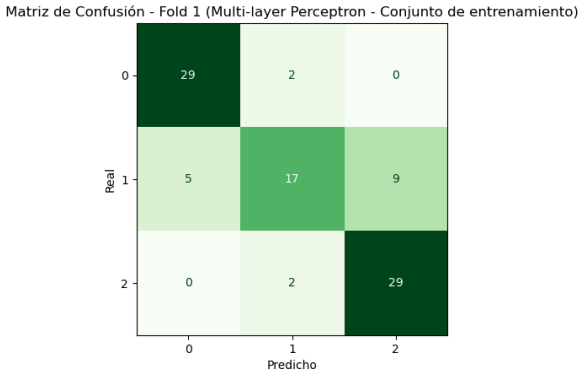

Figura D.189: MLP, Fold 1, C2, SMOTE

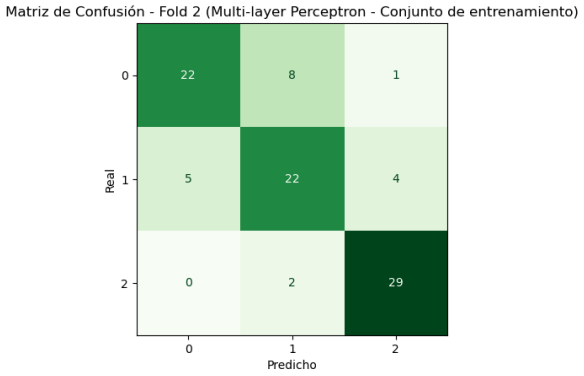

Figura D.190: MLP, Fold 2, C2, SMOTE

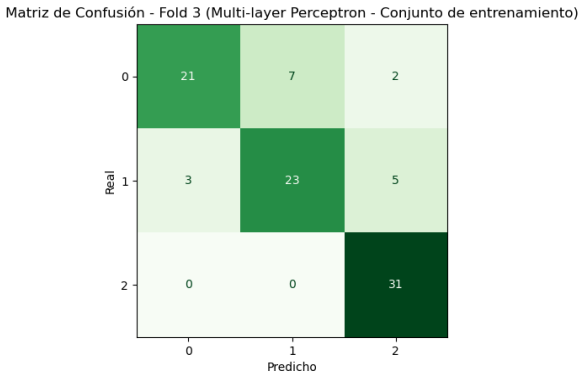

Figura D.190: MLP, Fold 3, C2, SMOTE

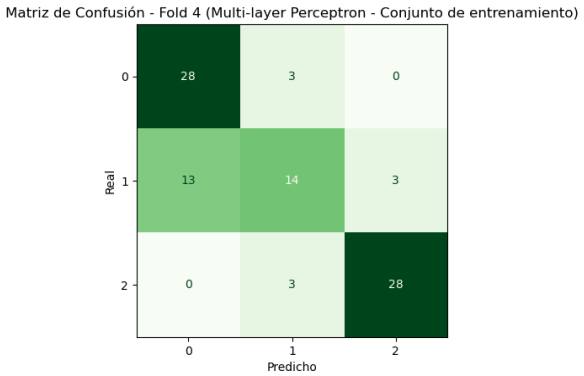

Figura D.191: MLP, Fold 4, C2, SMOTE

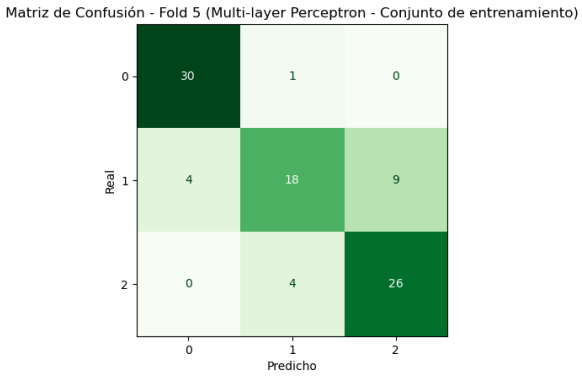

Figura D.191: MLP, Fold 5, C2, SMOTE

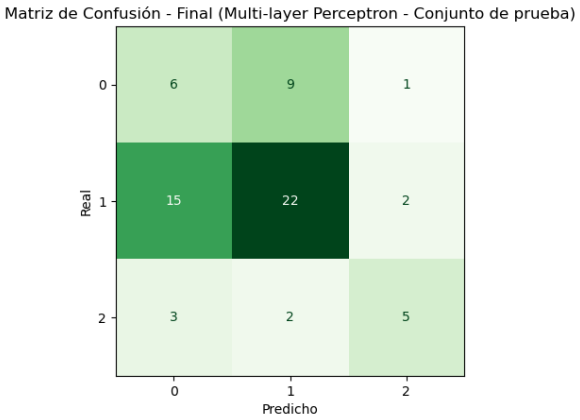

Figura D.192: MLP, Final, C2, SMOTE

D.3 Clasificación 3

D.3.1. ADASYN

Logistic Regression

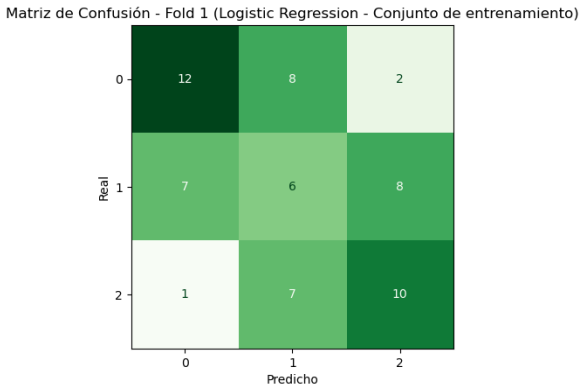

Figura D.193: Logistic Regression, Fold 1, C3, ADASYN

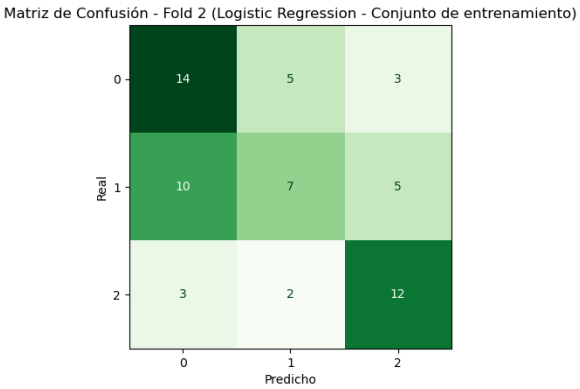

Figura D.194: Logistic Regression, Fold 2, C3, ADASYN

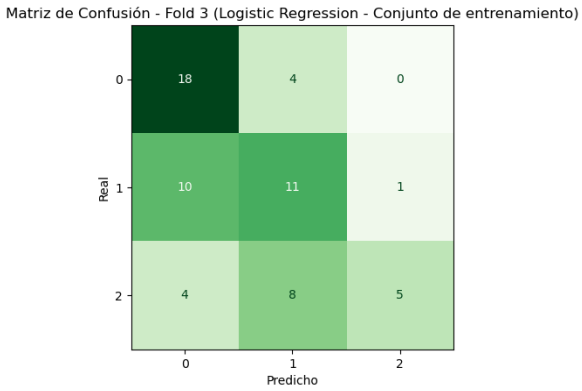

Figura D.194: Logistic Regression, Fold 3, C3, ADASYN

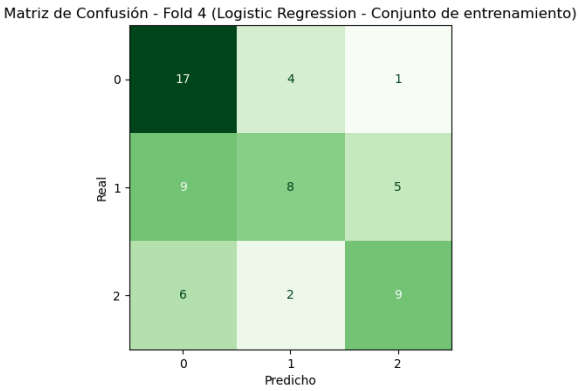

Figura D.195: Logistic Regression, Fold 4, C3, ADASYN

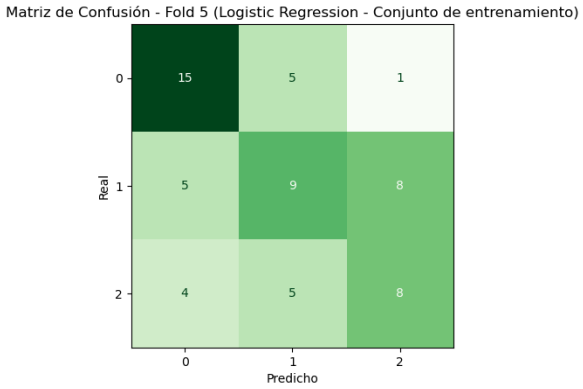

Figura D.195: Logistic Regression, Fold 5, C3, ADASYN

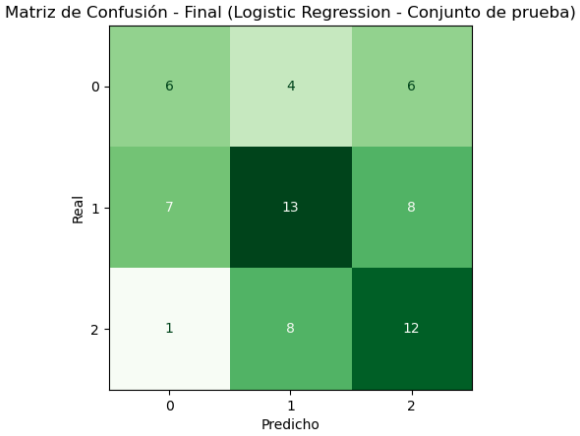

Figura D.196: Logistic Regression, Final, C3, ADASYN

Logistic Regression Lasso

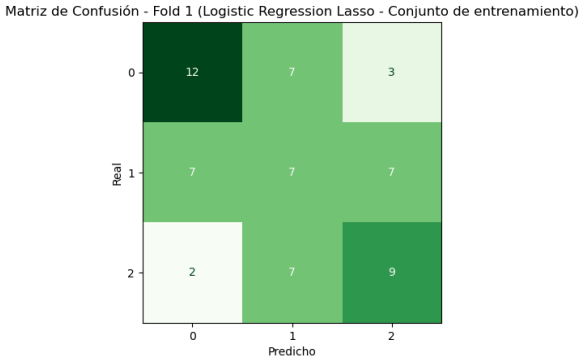

Figura D.197: Logistic Regression Lasso, Fold 1, C3, ADASYN

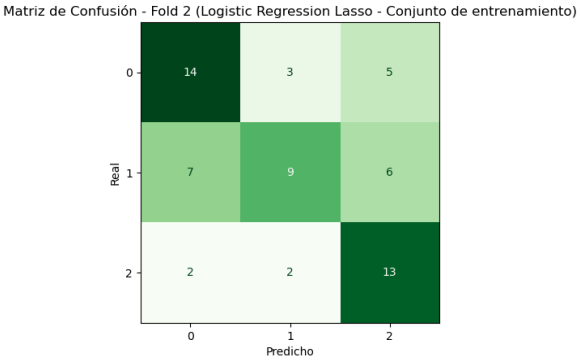

Figura D.198: Logistic Regression Lasso, Fold 2, C3, ADASYN

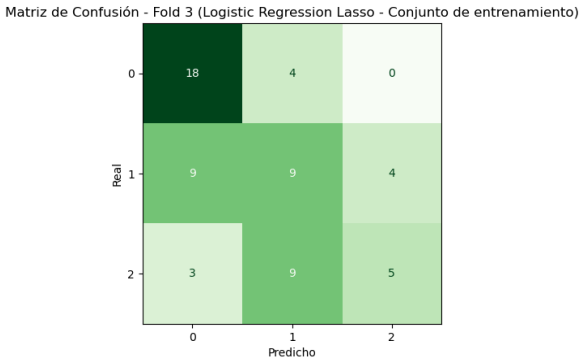

Figura D.198: Logistic Regression Lasso, Fold 3, C3, ADASYN

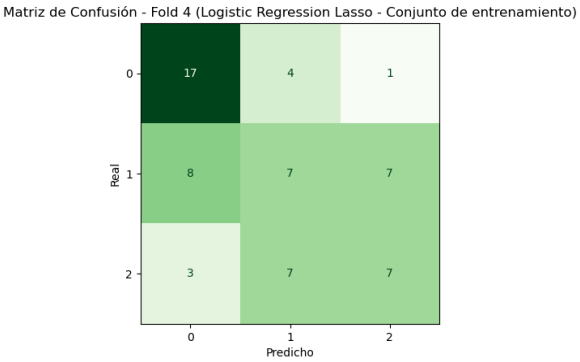

Figura D.199: Logistic Regression Lasso, Fold 4, C3, ADASYN

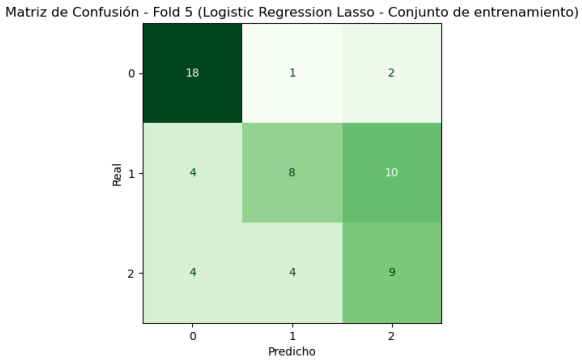

Figura D.199: Logistic Regression Lasso, Fold 5, C3, ADASYN

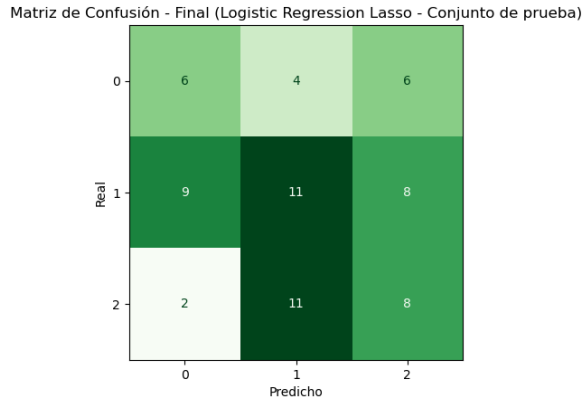

Figura D.200: Logistic Regression Lasso, Final, C3, ADASYN

Ridge

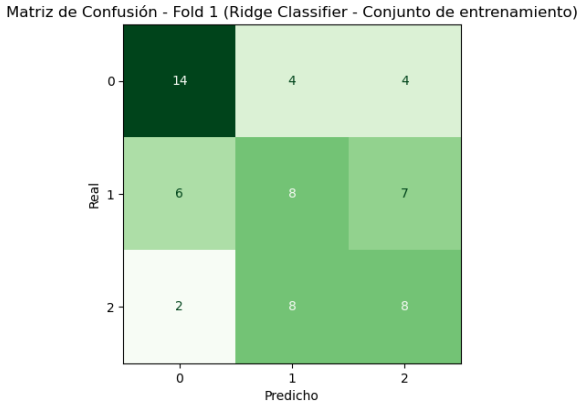

Figura D.201: Ridge, Fold 1, C3, ADASYN

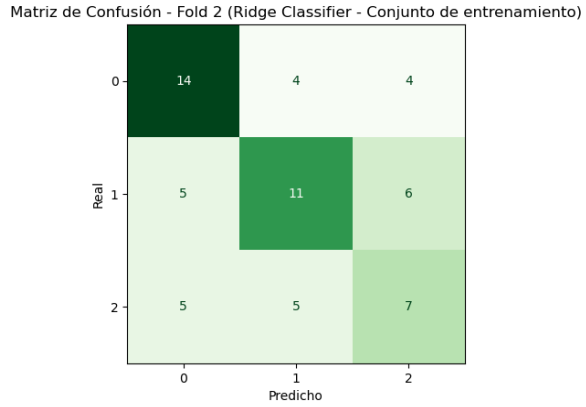

Figura D.202: Ridge, Fold 2, C3, ADASYN

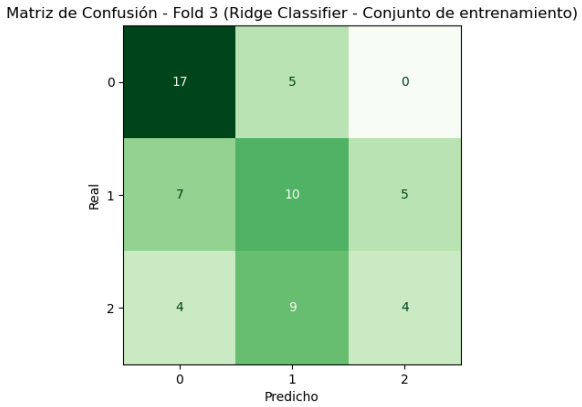

Figura D.202: Ridge, Fold 3, C3, ADASYN

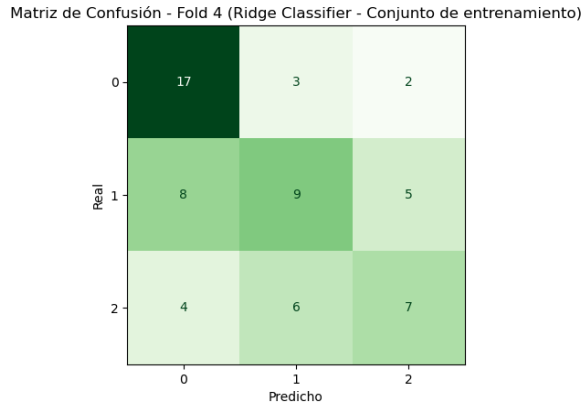

Figura D.203: Ridge, Fold 4, C3, ADASYN

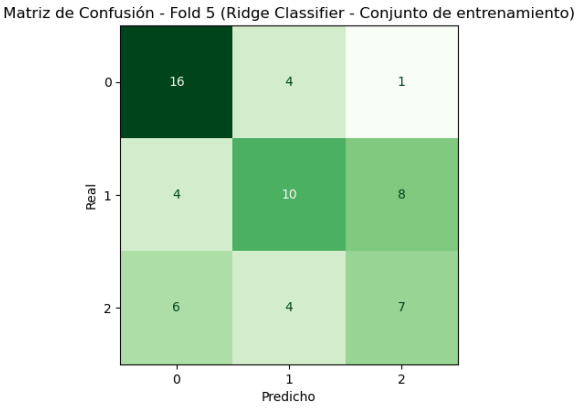

Figura D.203: Ridge, Fold 5, C3, ADASYN

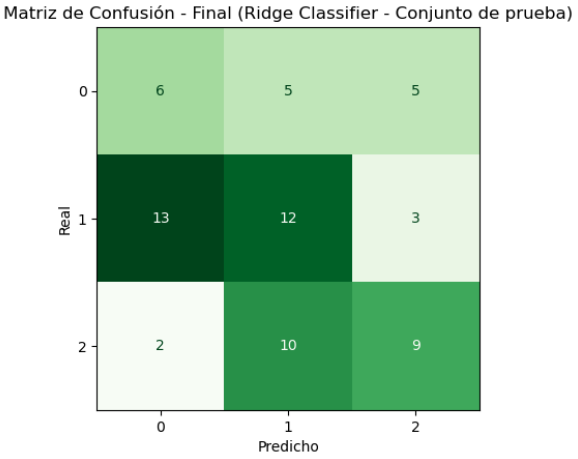

Figura D.204: Ridge, Final, C3, ADASYN

SDG

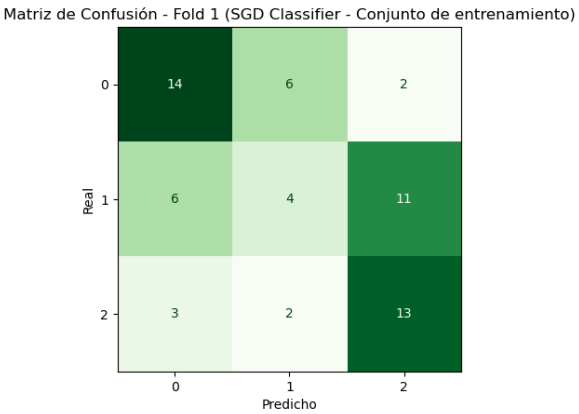

Figura D.205: SDG, Fold 1, C3, ADASYN

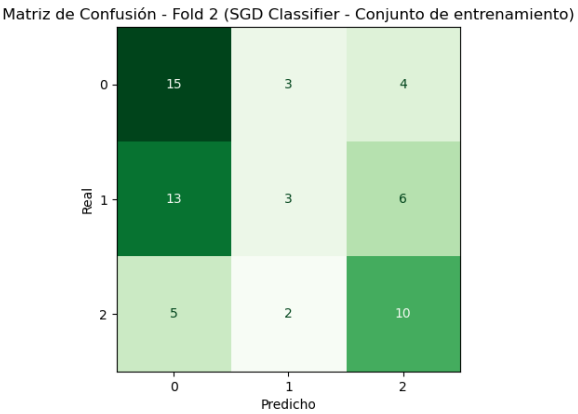

Figura D.206: SDG, Fold 2, C3, ADASYN

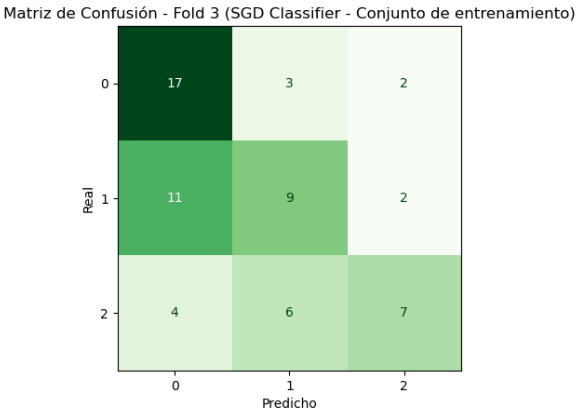

Figura D.206: SDG, Fold 3, C3, ADASYN

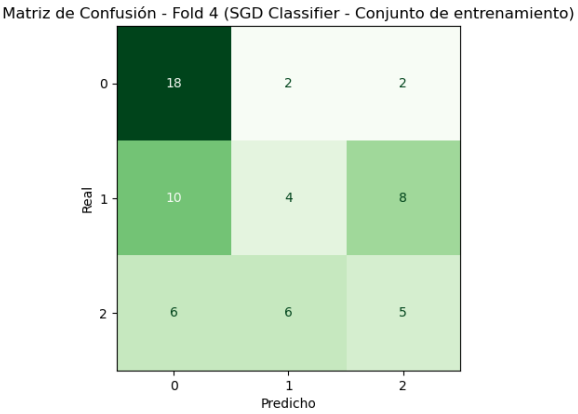

Figura D.207: SDG, Fold 4, C3, ADASYN

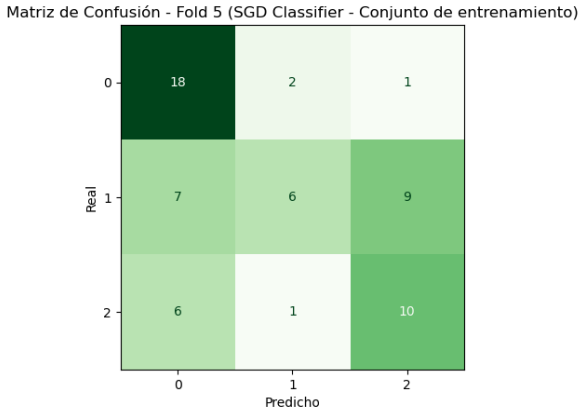

Figura D.207: SDG, Fold 5, C3, ADASYN

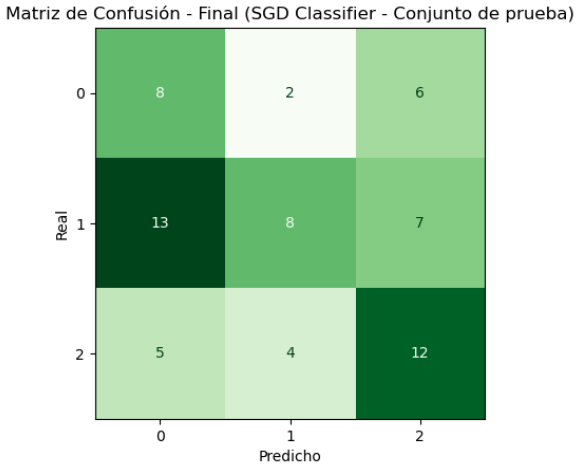

Figura D.208: SDG, Final, C3, ADASYN

Decision Tree

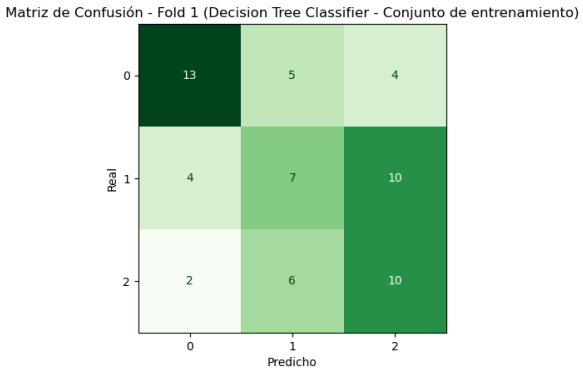

Figura D.209: Decision Tree, Fold 1, C3, ADASYN

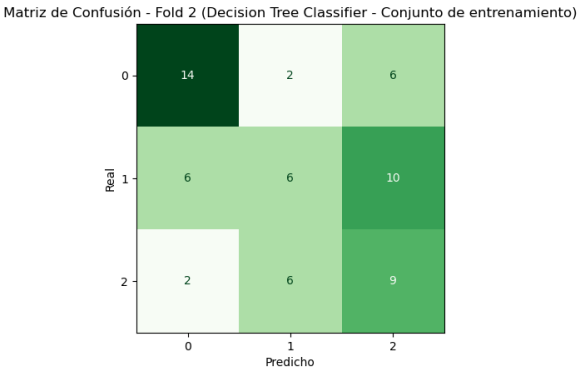

Figura D.210: Decision Tree, Fold 2, C3, ADASYN

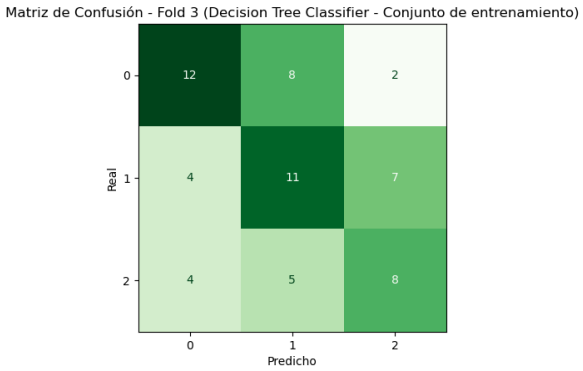

Figura D.210: Decision Tree, Fold 3, C3, ADASYN

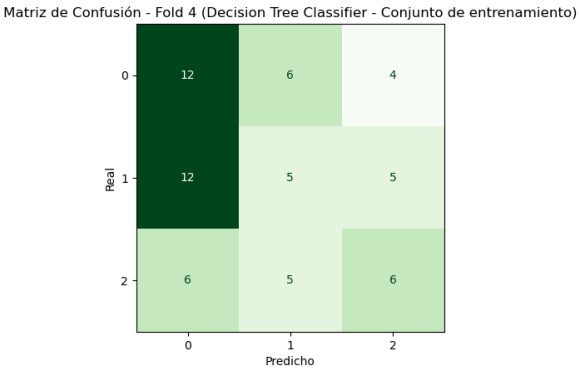

Figura D.211: Decision Tree, Fold 4, C3, ADASYN

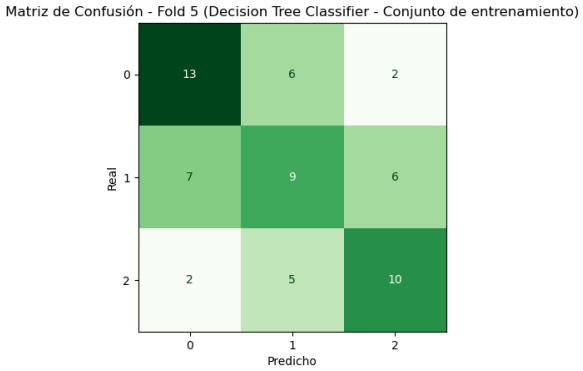

Figura D.211: Decision Tree, Fold 5, C3, ADASYN

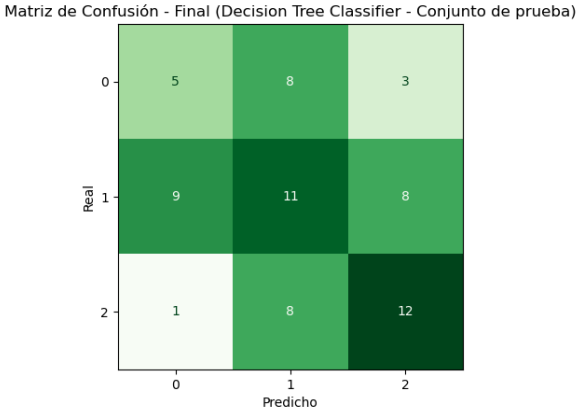

Figura D.212: Decision Tree, Final, C3, ADASYN

Random Forest

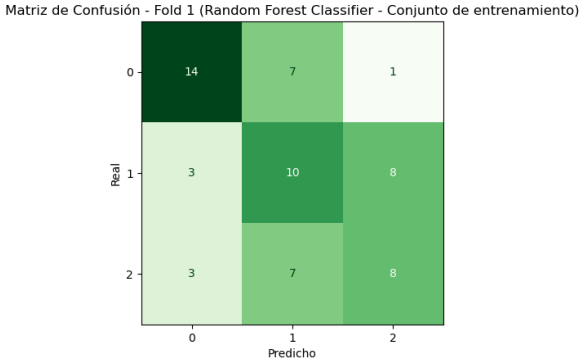

Figura D.213: Random Forest, Fold 1, C3, ADASYN

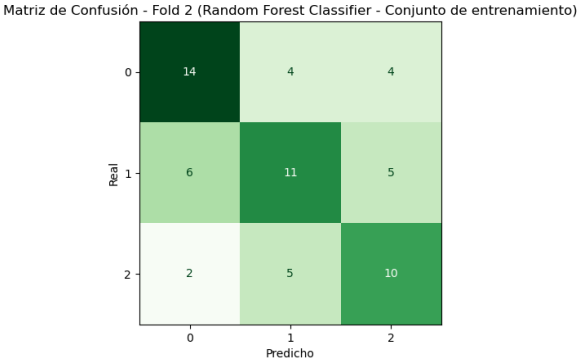

Figura D.214: Random Forest, Fold 2, C3, ADASYN

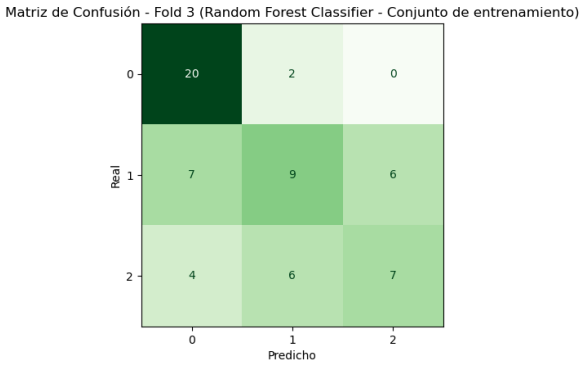

Figura D.214: Random Forest, Fold 3, C3, ADASYN

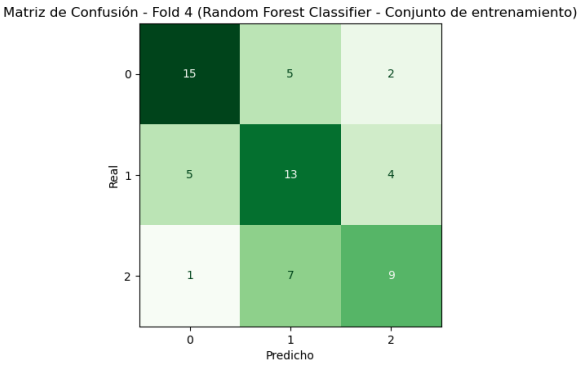

Figura D.215: Random Forest, Fold 4, C3, ADASYN

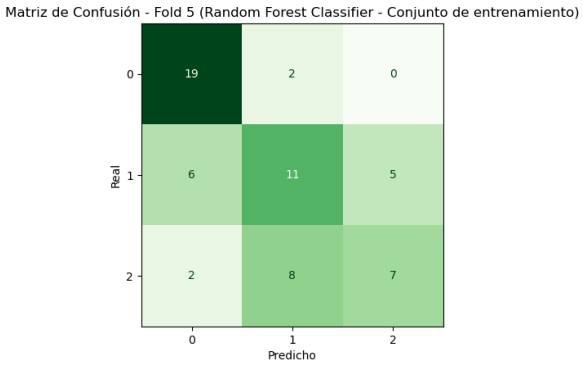

Figura D.215: Random Forest, Fold 5, C3, ADASYN

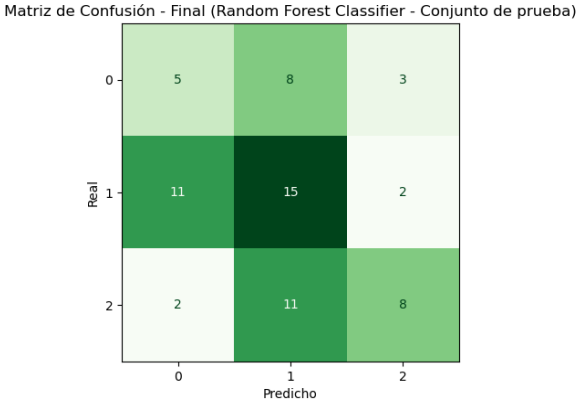

Figura D.216: Random Forest, Final, C3, ADASYN

AdaBoost

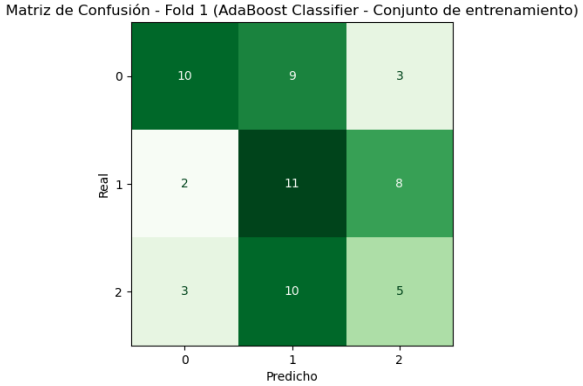

Figura D.217: AdaBoost, Fold 1, C3, ADASYN

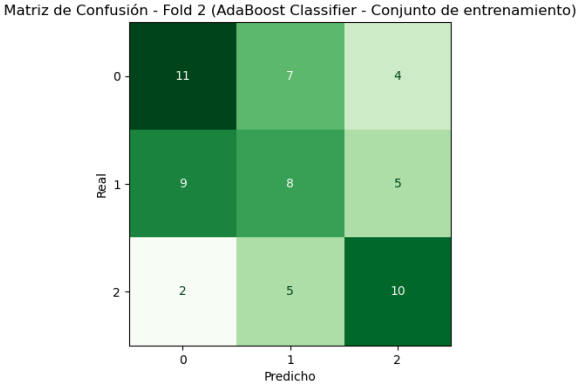

Figura D.218: AdaBoost, Fold 2, C3, ADASYN

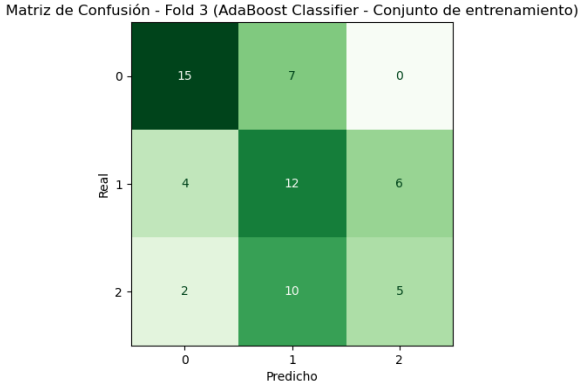

Figura D.218: AdaBoost, Fold 3, C3, ADASYN

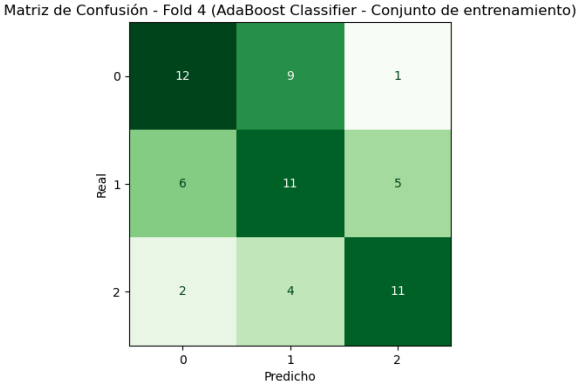

Figura D.219: AdaBoost, Fold 4, C3, ADASYN

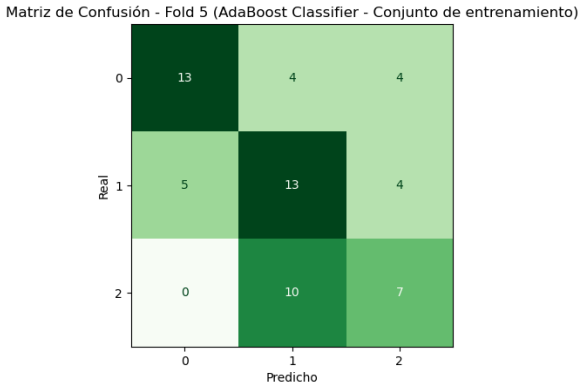

Figura D.219: AdaBoost, Fold 5, C3, ADASYN

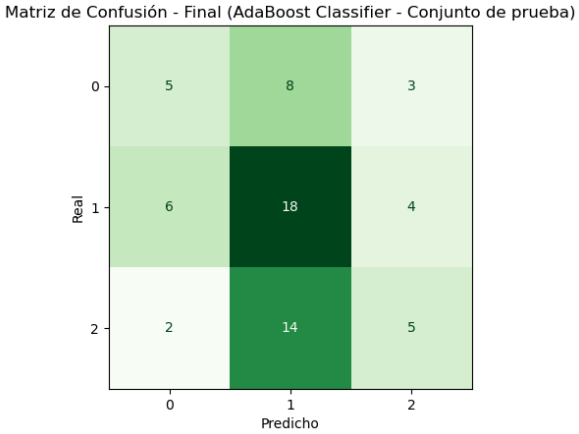

Figura D.220: AdaBoost, Final, C3, ADASYN

Gradient Boosting

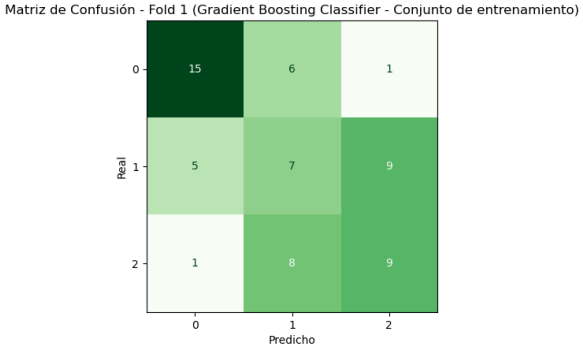

Figura D.221: Gradient Boosting, Fold 1, C3, ADASYN

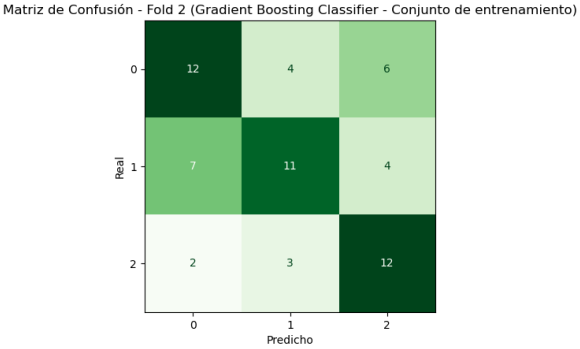

Figura D.222: Gradient Boosting, Fold 2, C3, ADASYN

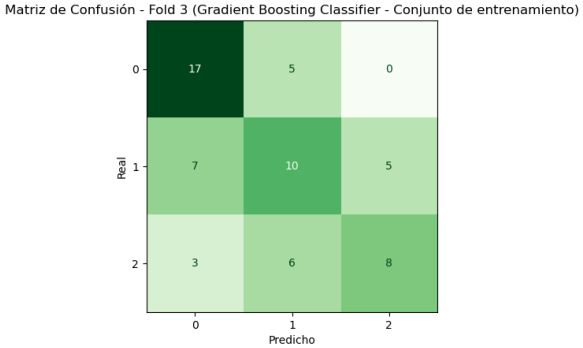

Figura D.222: Gradient Boosting, Fold 3, C3, ADASYN

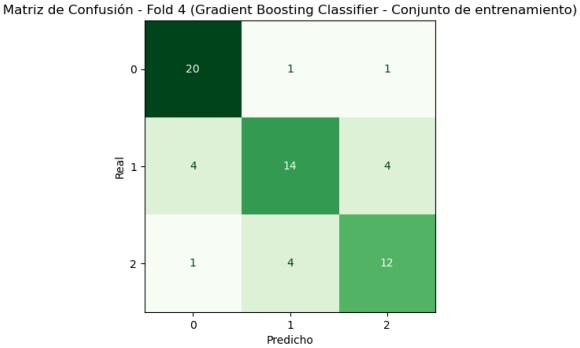

Figura D.223: Gradient Boosting, Fold 4, C3, ADASYN

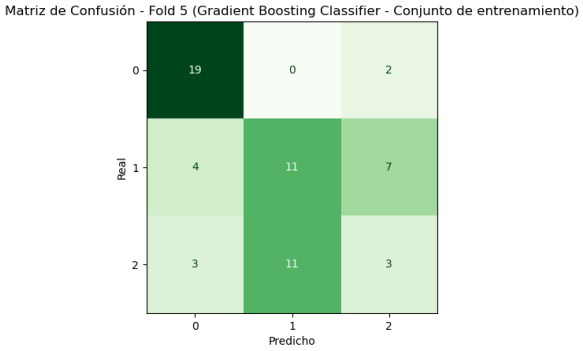

Figura D.223: Gradient Boosting, Fold 5, C3, ADASYN

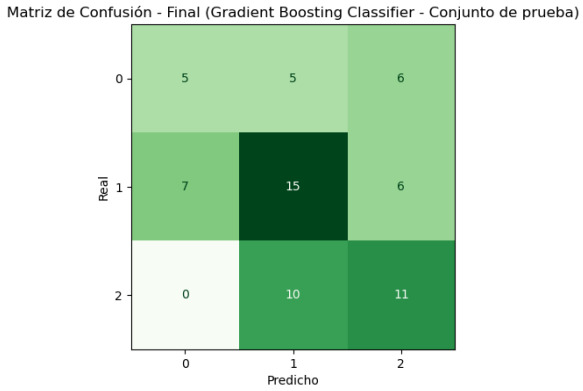

Figura D.224: Gradient Boosting, Final, C3, ADASYN

SVC

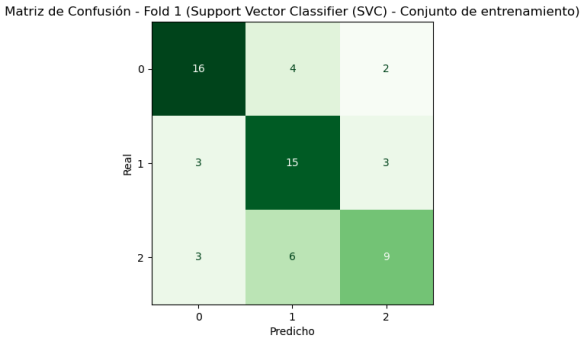

Figura D.225: SVC, Fold 1, C3, ADASYN

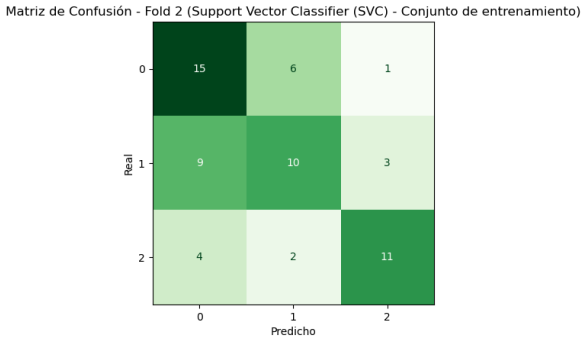

Figura D.226: SVC, Fold 2, C3, ADASYN

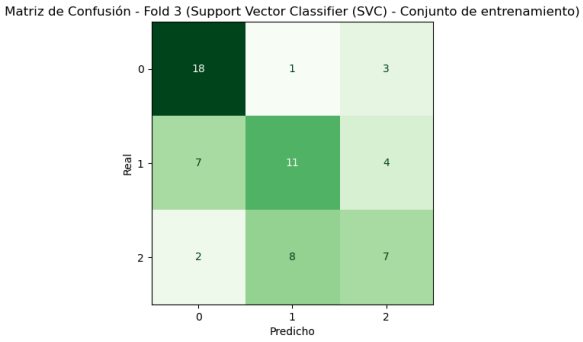

Figura D.226: SVC, Fold 3, C3, ADASYN

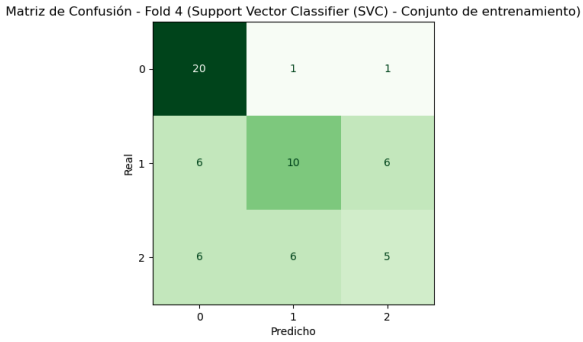

Figura D.227: SVC, Fold 4, C3, ADASYN

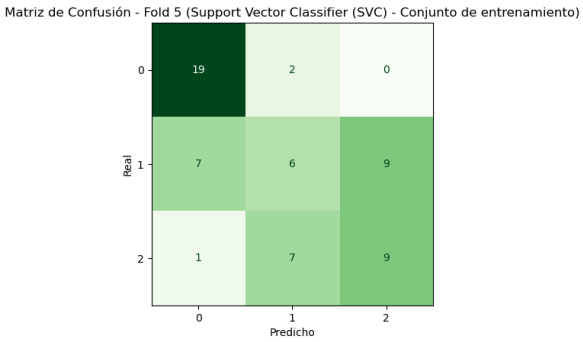

Figura D.227: SVC, Fold 5, C3, ADASYN

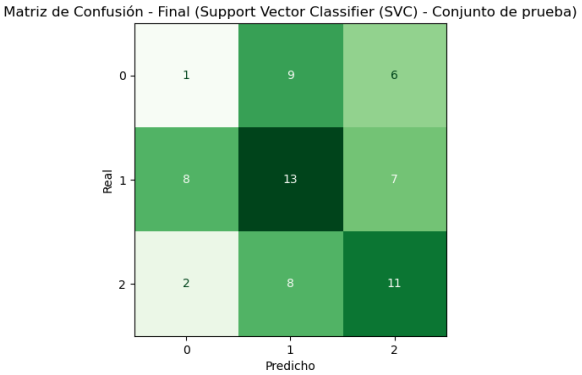

Figura D.228: SVC, Final, C3, ADASYN

k-Nearest Neighbors

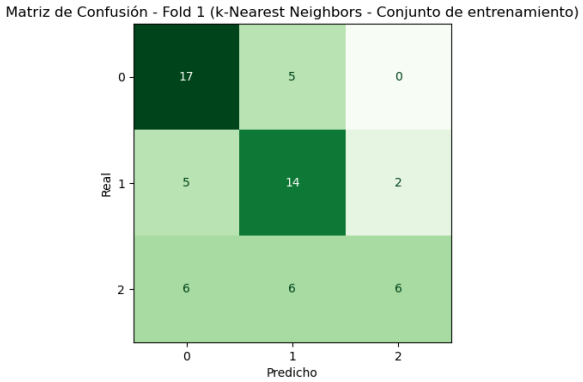

Figura D.229: k-Nearest Neighbors, Fold 1, C3, ADASYN

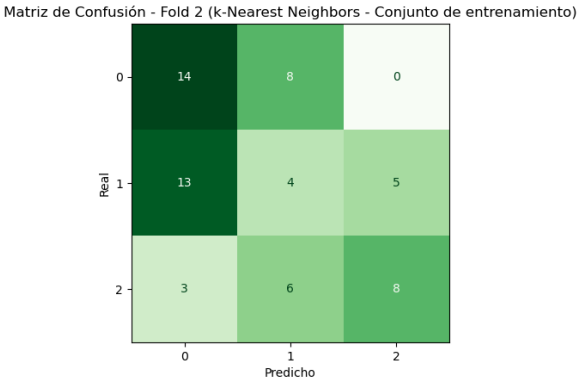

Figura D.230: k-Nearest Neighbors, Fold 2, C3, ADASYN

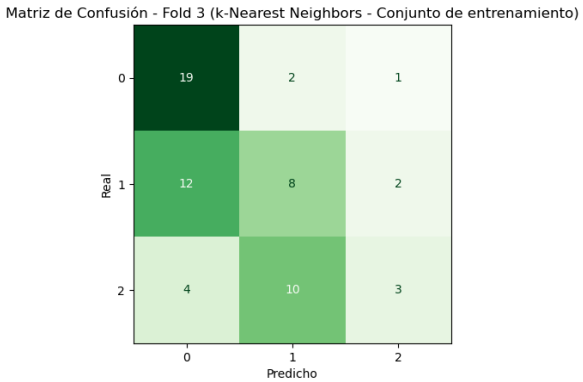

Figura D.230: k-Nearest Neighbors, Fold 3, C3, ADASYN

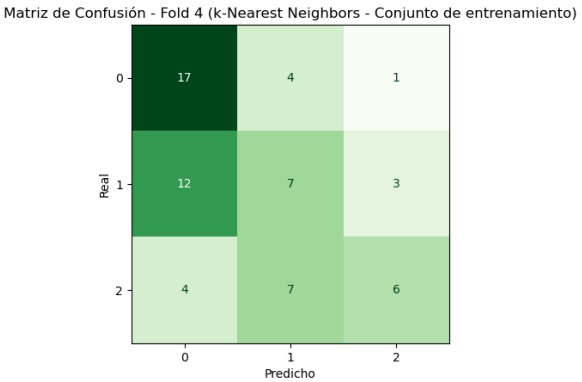

Figura D.231: k-Nearest Neighbors, Fold 4, C3, ADASYN

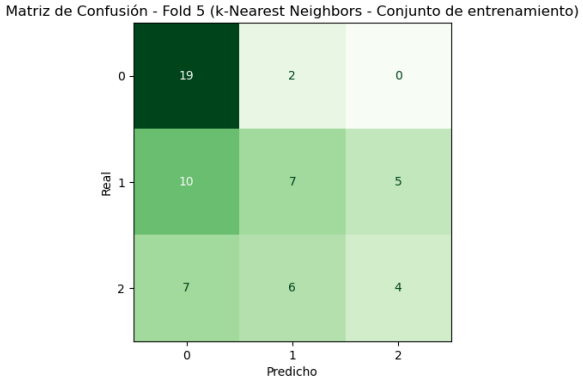

Figura D.231: k-Nearest Neighbors, Fold 5, C3, ADASYN

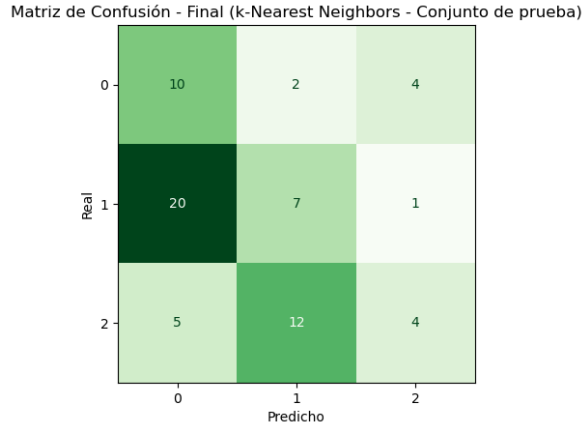

Figura D.232: k-Nearest Neighbors, Final, C3, ADASYN

Gaussian Naive Bayes

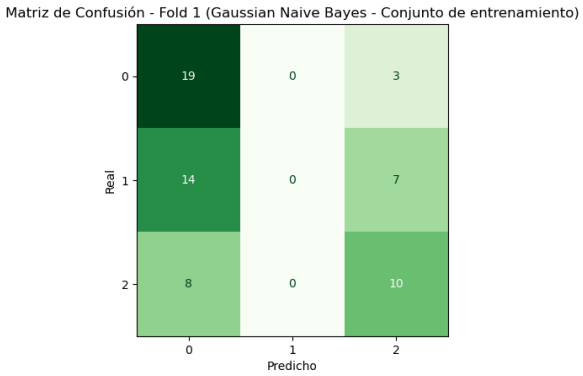

Figura D.233: Gaussian Naive Bayes, Fold 1, C3, ADASYN

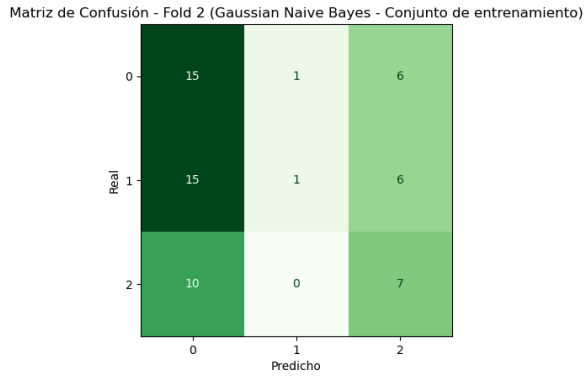

Figura D.234: Gaussian Naive Bayes, Fold 2, C3, ADASYN

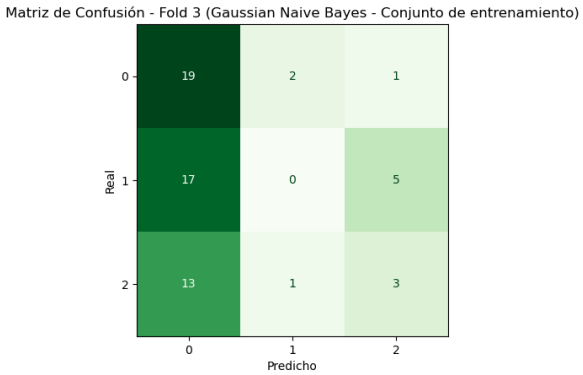

Figura D.234: Gaussian Naive Bayes, Fold 3, C3, ADASYN

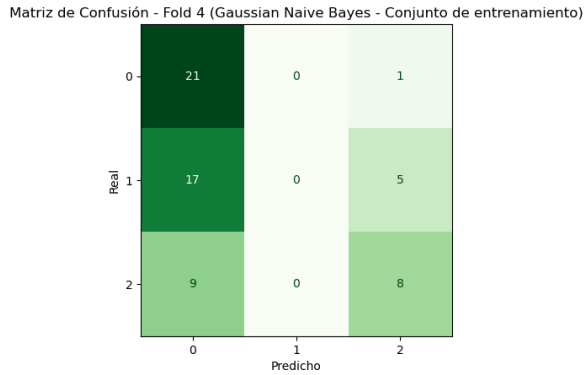

Figura D.235: Gaussian Naive Bayes, Fold 4, C3, ADASYN

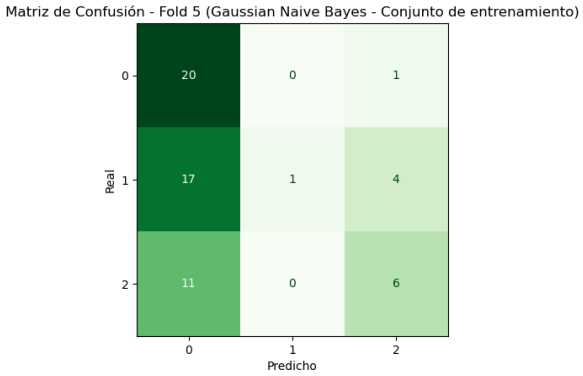

Figura D.235: Gaussian Naive Bayes, Fold 5, C3, ADASYN

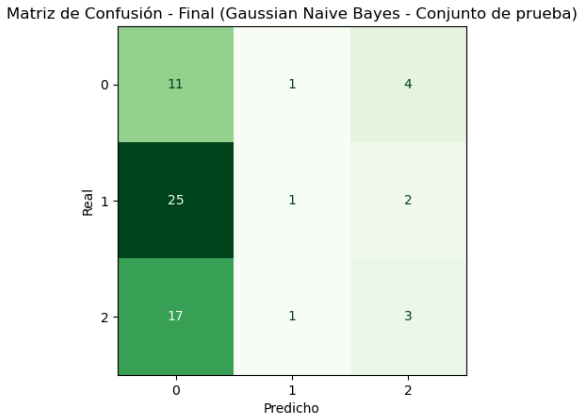

Figura D.236: Gaussian Naive Bayes, Final, C3, ADASYN

MLP

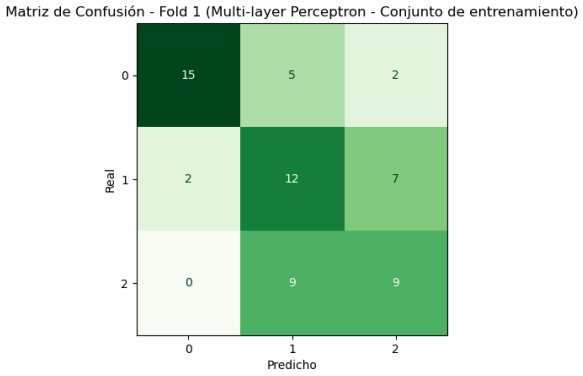

Figura D.237: MLP, Fold 1, C3, ADASYN

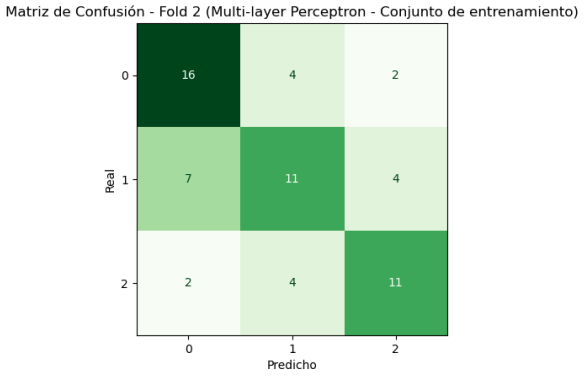

Figura D.238: MLP, Fold 2, C3, ADASYN

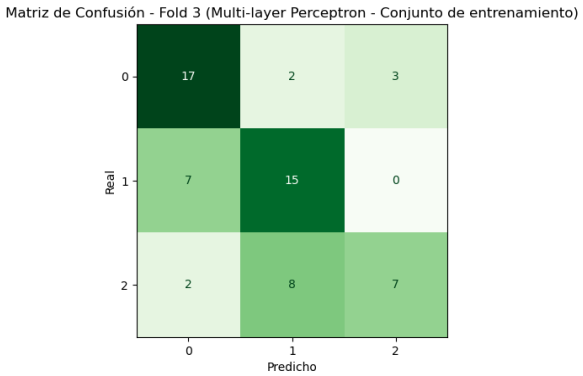

Figura D.238: MLP, Fold 3, C3, ADASYN

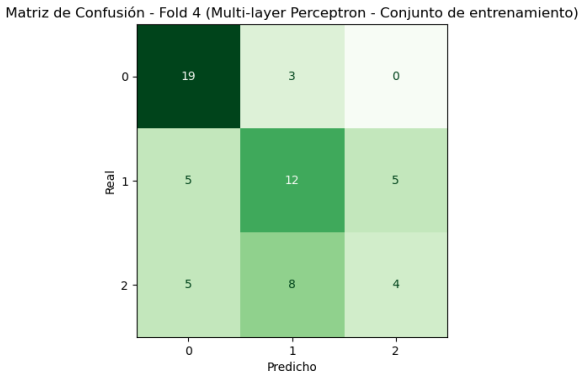

Figura D.239: MLP, Fold 4, C3, ADASYN

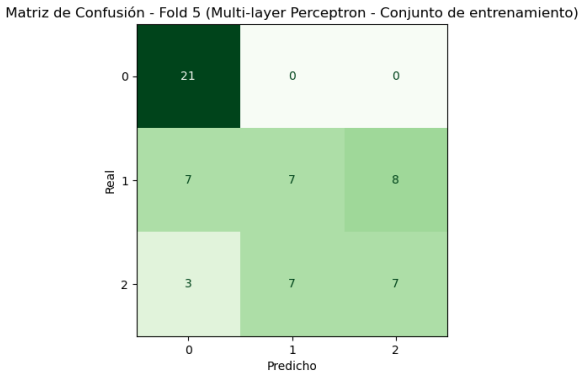

Figura D.239: MLP, Fold 5, C3, ADASYN

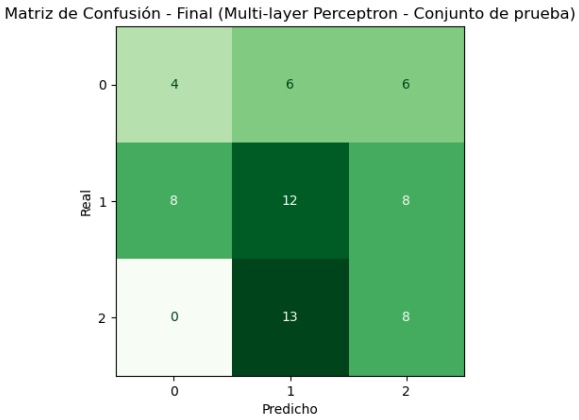

Figura D.240: MLP, Final, C3, ADASYN

D.3.2. SMOTE

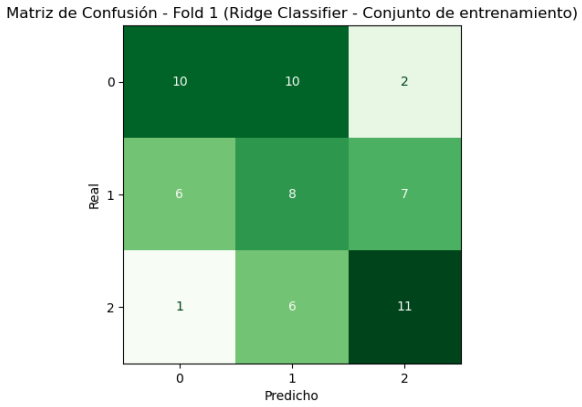

Figura D.241: Ridge, Fold 1, C3, SMOTE

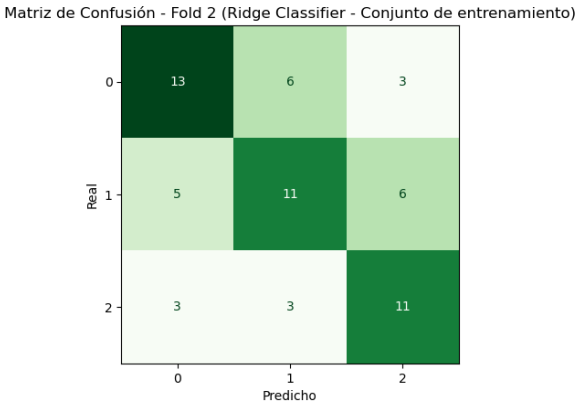

Figura D.242: Ridge, Fold 2, C3, SMOTE

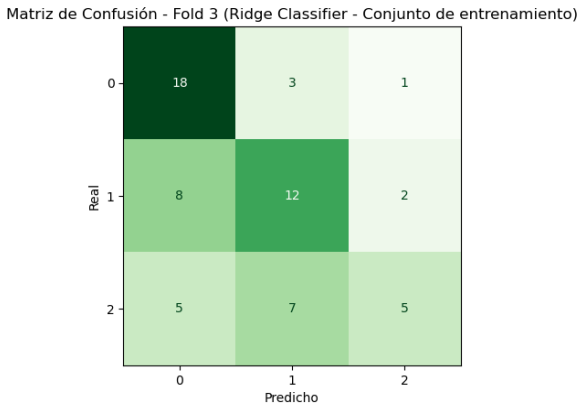

Figura D.242: Ridge, Fold 3, C3, SMOTE

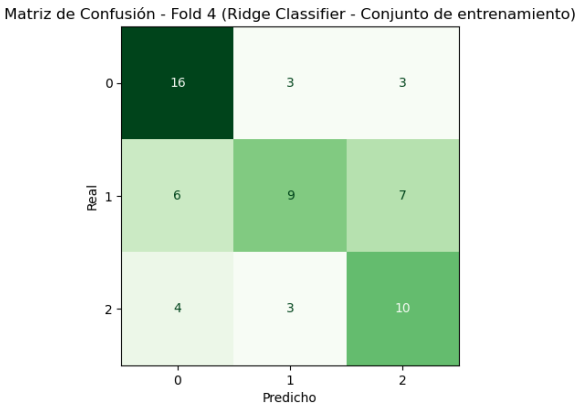

Figura D.243: Ridge, Fold 4, C3, SMOTE

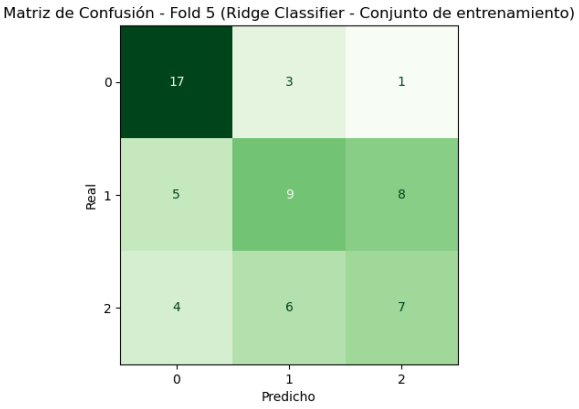

Figura D.243: Ridge, Fold 5, C3, SMOTE

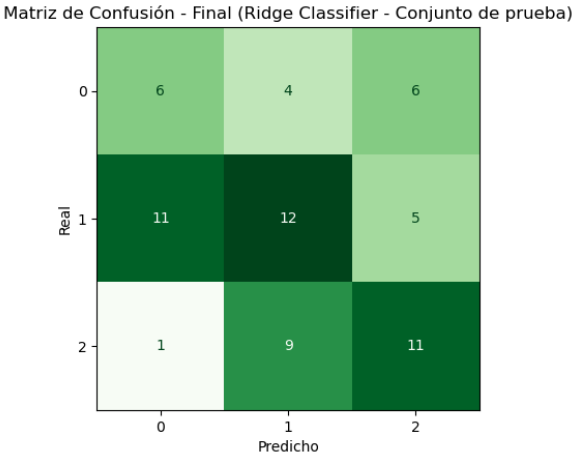

Figura D.244: Ridge, Final, C3, SMOTE

SDG

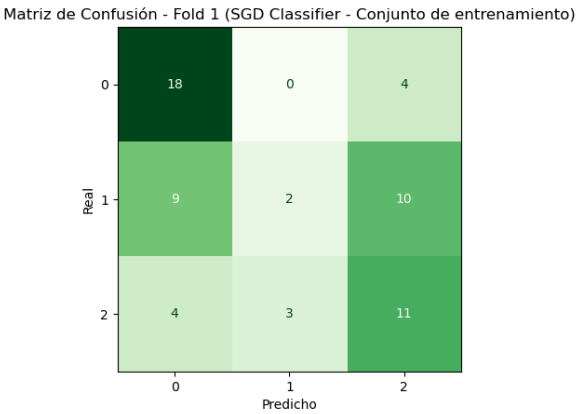

Figura D.245: SDG, Fold 1, C3, SMOTE

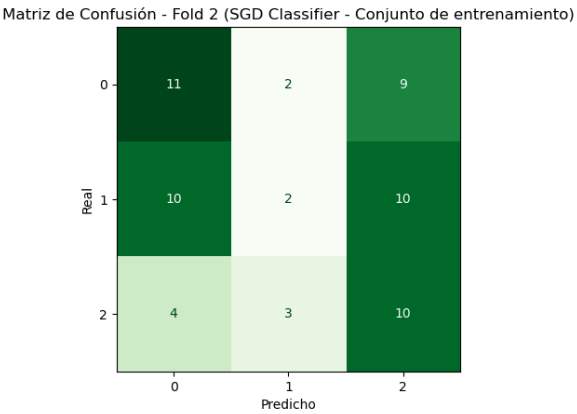

Figura D.246: SDG, Fold 2, C3, SMOTE

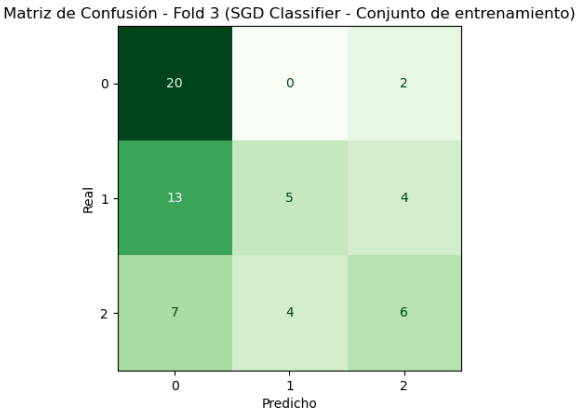

Figura D.246: SDG, Fold 3, C3, SMOTE

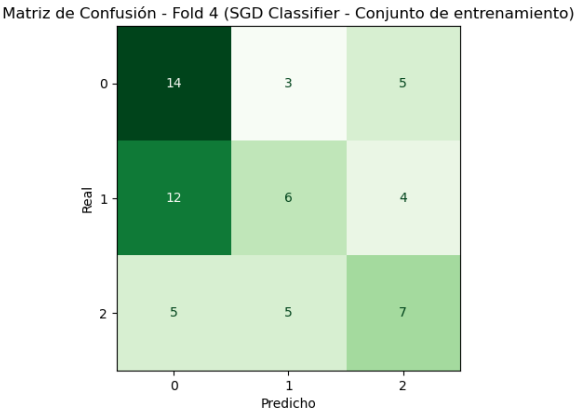

Figura D.247: SDG, Fold 4, C3, SMOTE

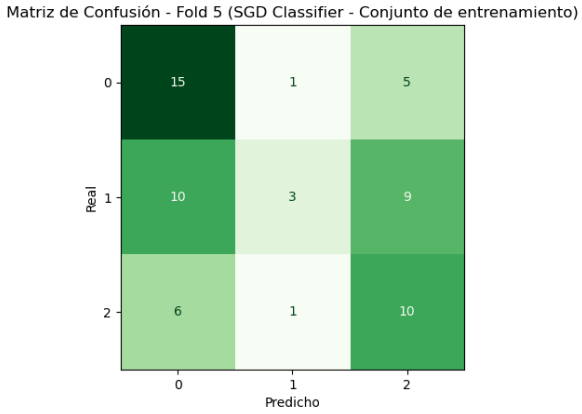

Figura D.247: SDG, Fold 5, C3, SMOTE

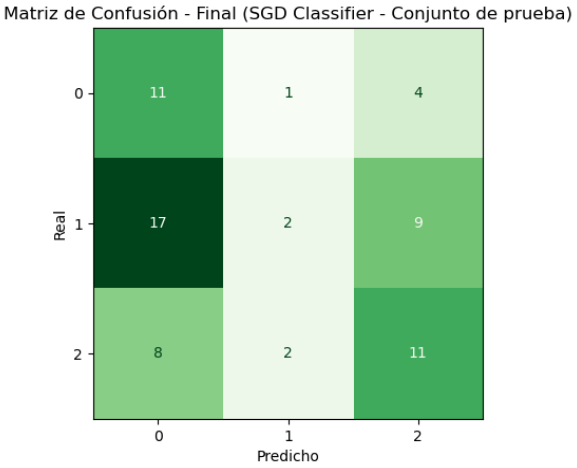

Figura D.248: SDG, Final, C3, SMOTE

Decision Tree

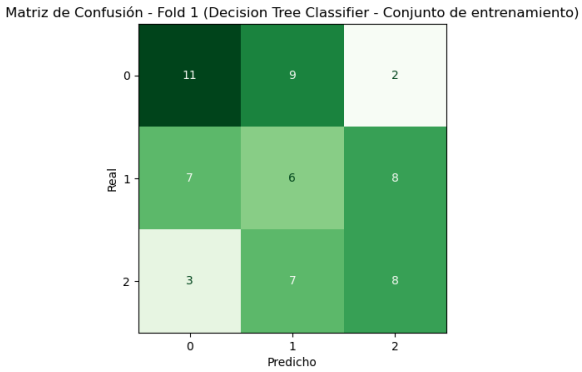

Figura D.249: Decision Tree, Fold 1, C3, SMOTE

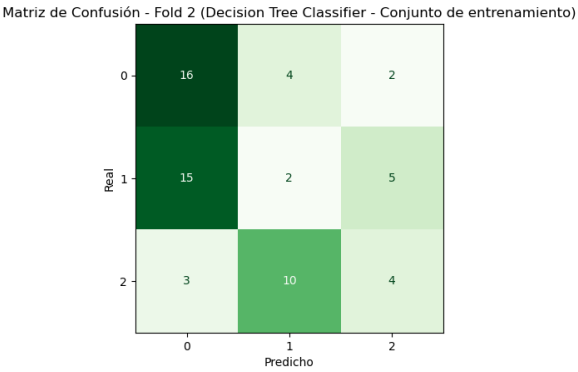

Figura D.250: Decision Tree, Fold 2, C3, SMOTE

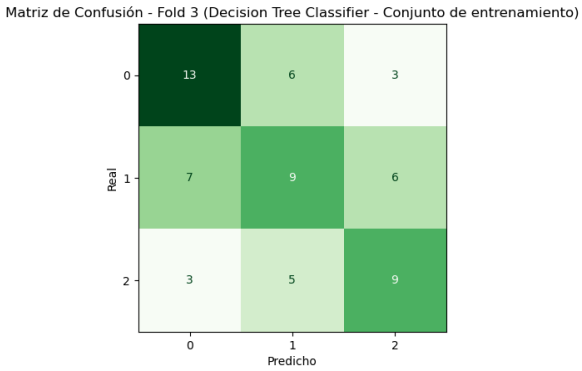

Figura D.250: Decision Tree, Fold 3, C3, SMOTE

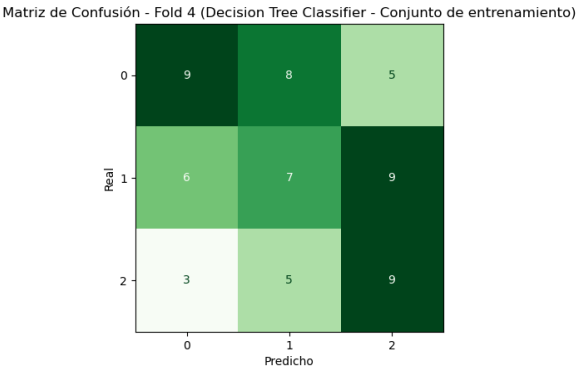

Figura D.251: Decision Tree, Fold 4, C3, SMOTE

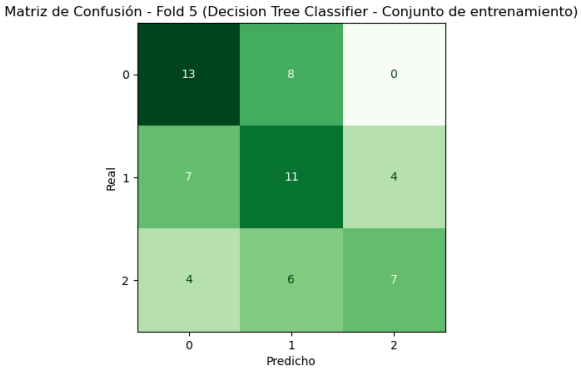

Figura D.251: Decision Tree, Fold 5, C3, SMOTE

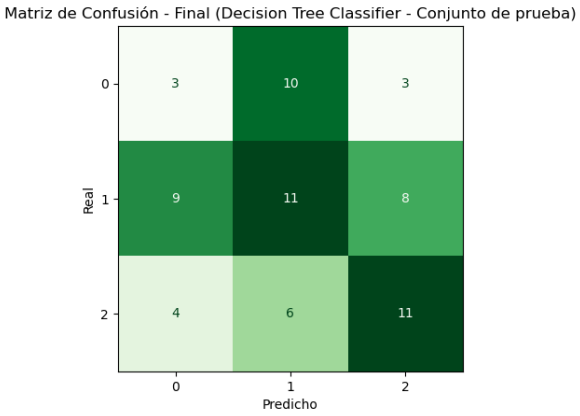

Figura D.252: Decision Tree, Final, C3, SMOTE

Random Forest

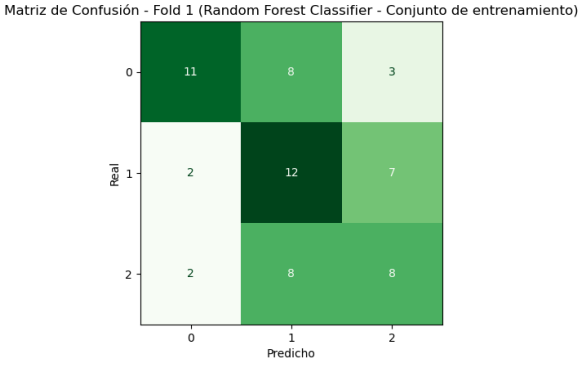

Figura D.253: Random Forest, Fold 1, C3, SMOTE

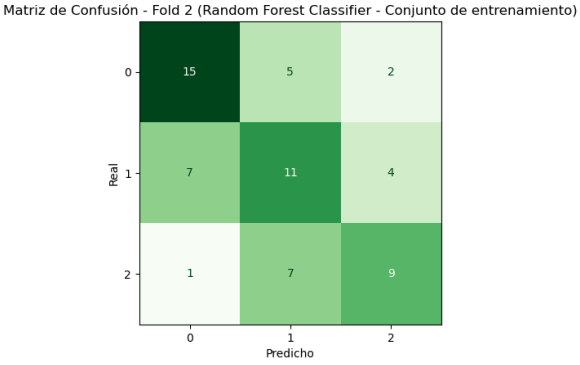

Figura D.254: Random Forest, Fold 2, C3, SMOTE

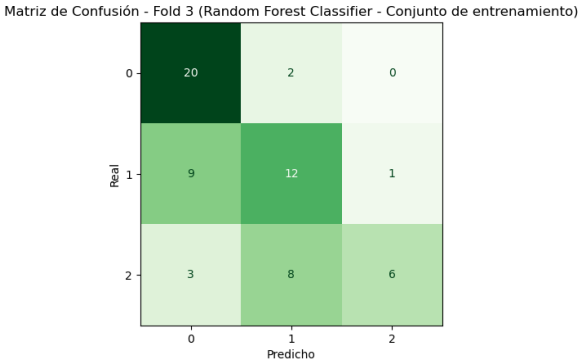

Figura D.254: Random Forest, Fold 3, C3, SMOTE

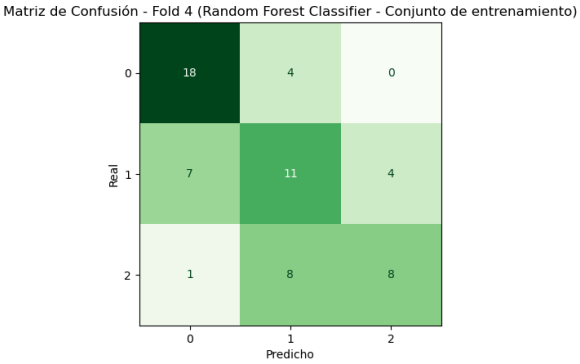

Figura D.255: Random Forest, Fold 4, C3, SMOTE

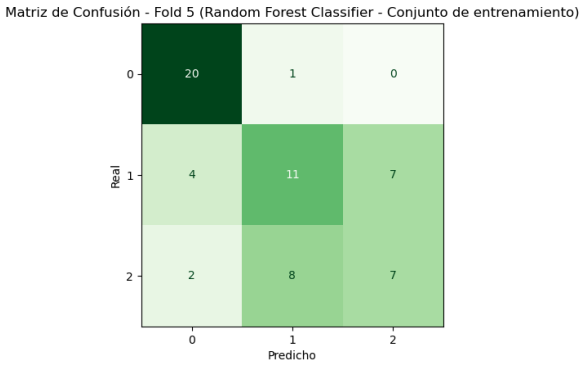

Figura D.255: Random Forest, Fold 5, C3, SMOTE

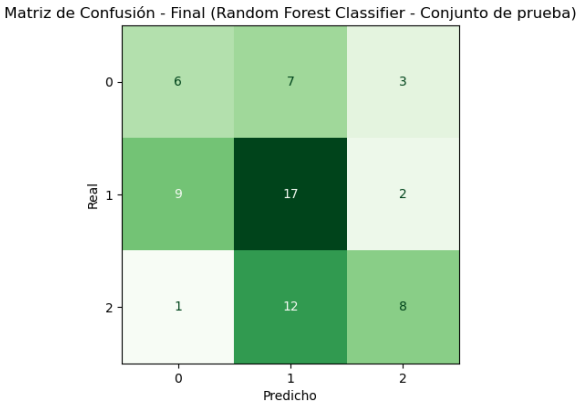

Figura D.256: Random Forest, Final, C3, SMOTE

AdaBoost

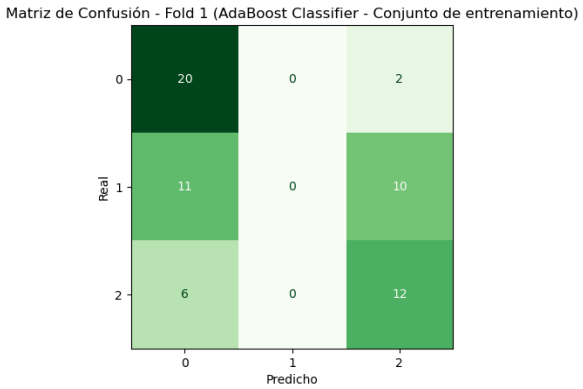

Figura D.257: AdaBoost, Fold 1, C3, SMOTE

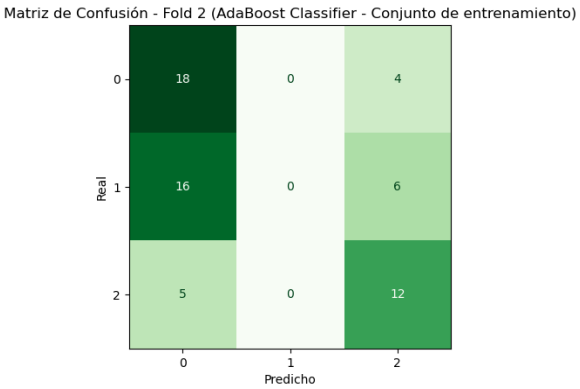

Figura D.258: AdaBoost, Fold 2, C3, SMOTE

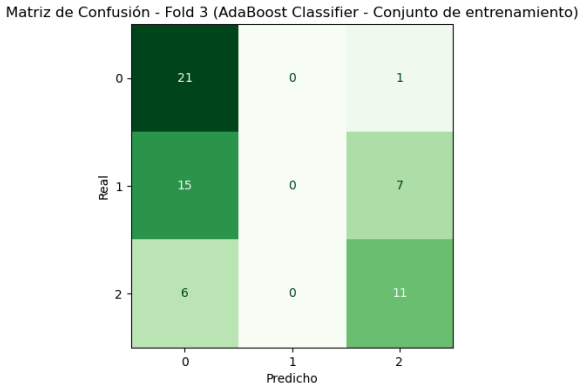

Figura D.258: AdaBoost, Fold 3, C3, SMOTE

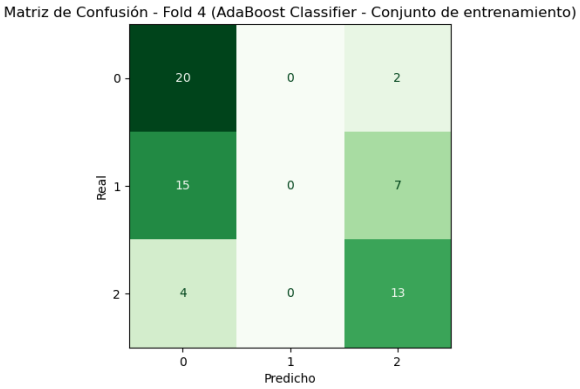

Figura D.259: AdaBoost, Fold 4, C3, SMOTE

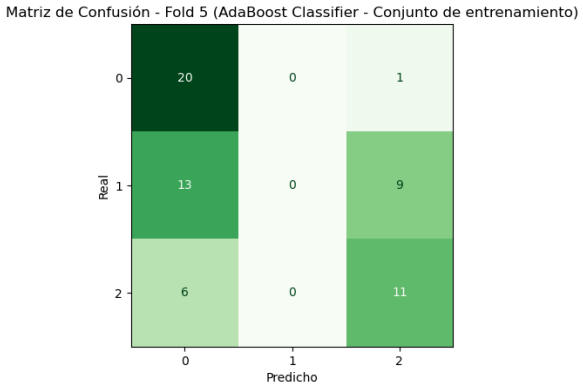

Figura D.259: AdaBoost, Fold 5, C3, SMOTE

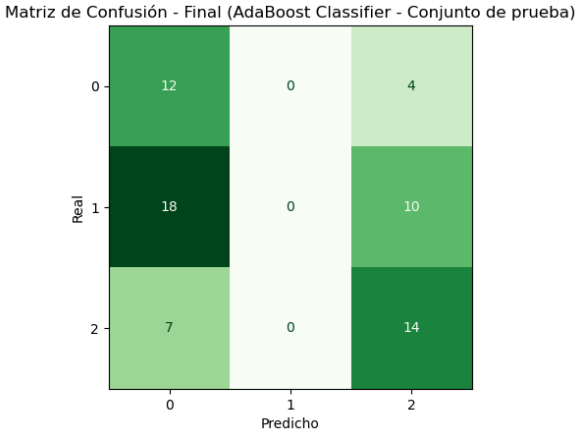

Figura D.260: AdaBoost, Final, C3, SMOTE

Gradient Boosting

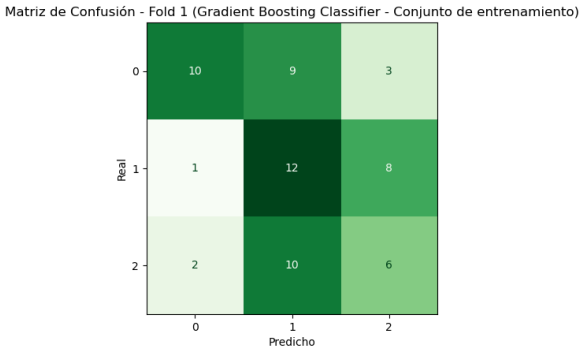

Figura D.261: Gradient Boosting, Fold 1, C3, SMOTE

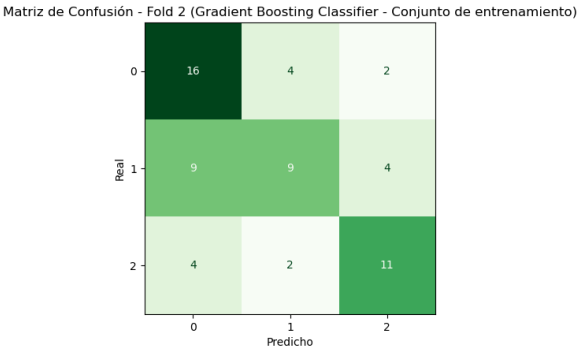

Figura D.262: Gradient Boosting, Fold 2, C3, SMOTE

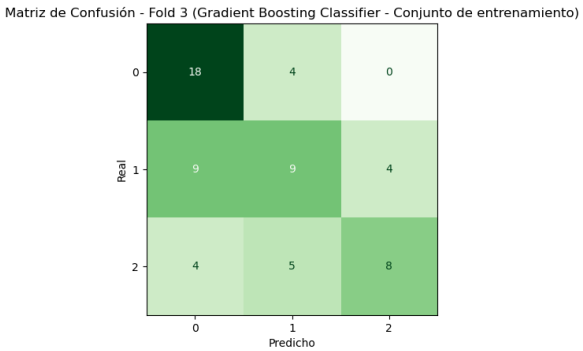

Figura D.262: Gradient Boosting, Fold 3, C3, SMOTE

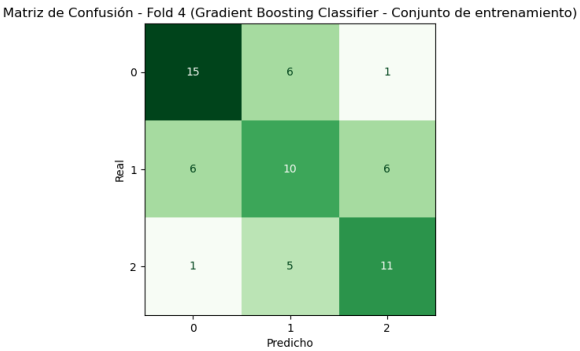

Figura D.263: Gradient Boosting, Fold 4, C3, SMOTE

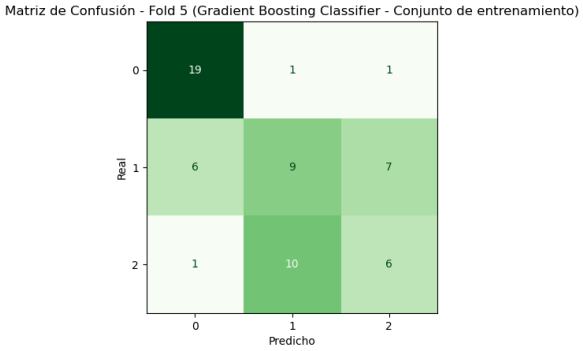

Figura D.263: Gradient Boosting, Fold 5, C3, SMOTE

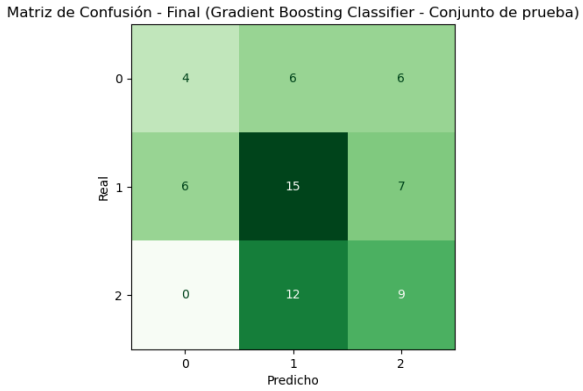

Figura D.264: Gradient Boosting, Final, C3, SMOTE

SVC

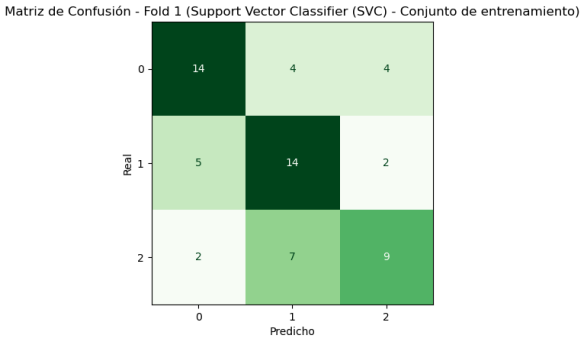

Figura D.265: SVC, Fold 1, C3, SMOTE

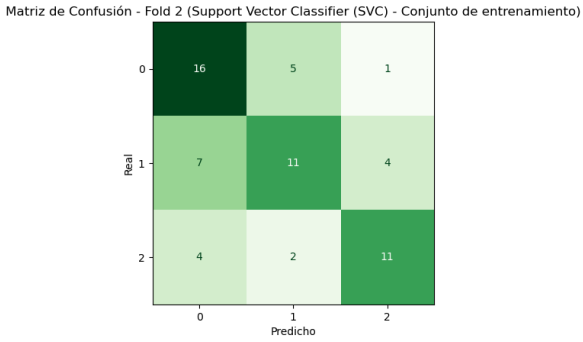

Figura D.266: SVC, Fold 2, C3, SMOTE

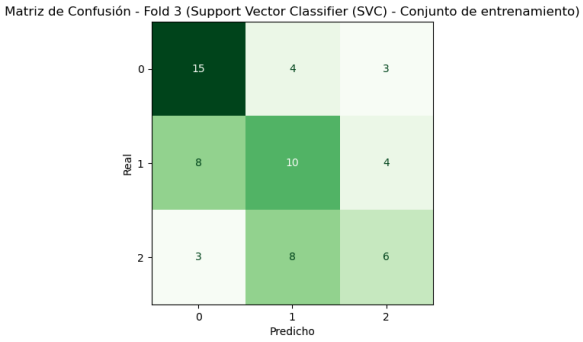

Figura D.266: SVC, Fold 3, C3, SMOTE

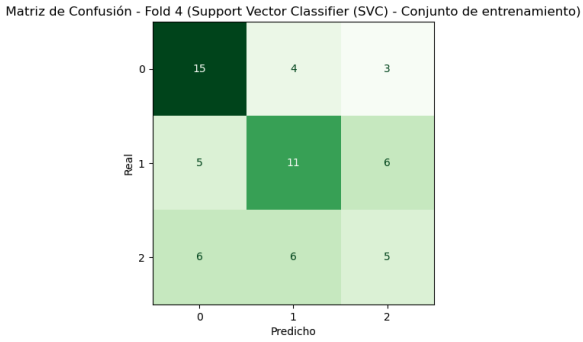

Figura D.267: SVC, Fold 4, C3, SMOTE

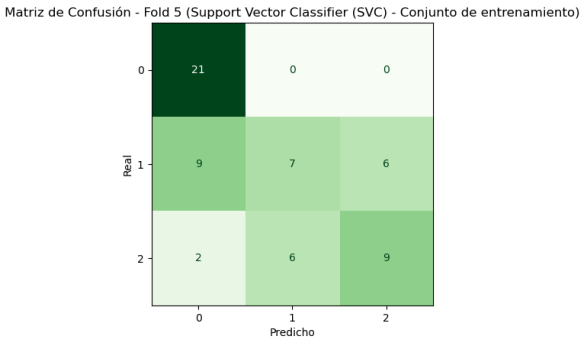

Figura D.267: SVC, Fold 5, C3, SMOTE

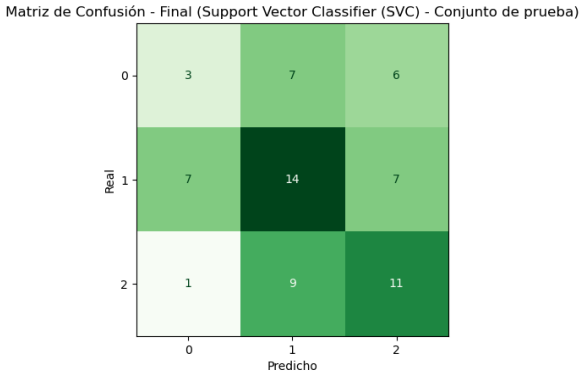

Figura D.268: SVC, Final, C3, SMOTE

k-Nearest Neighbors

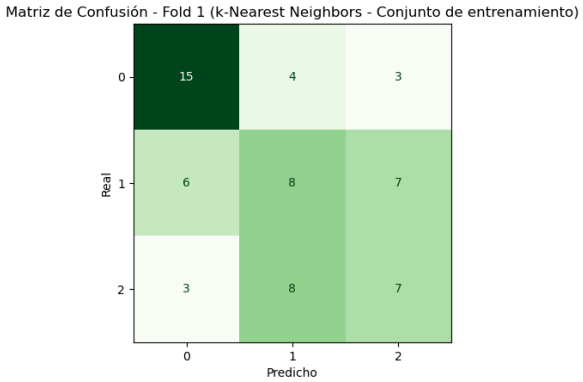

Figura D.269: k-Nearest Neighbors, Fold 1, C3, SMOTE

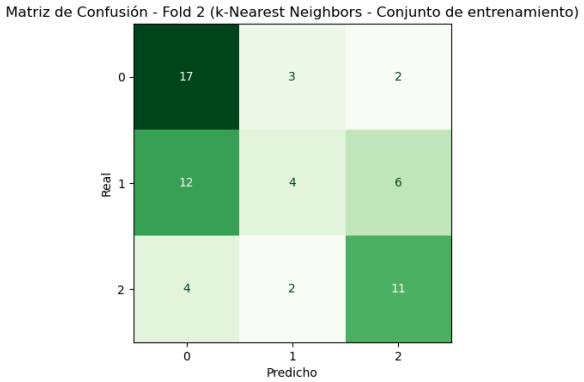

Figura D.270: k-Nearest Neighbors, Fold 2, C3, SMOTE

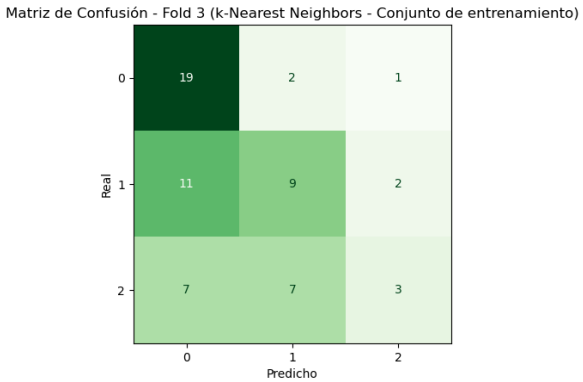

Figura D.270: k-Nearest Neighbors, Fold 3, C3, SMOTE

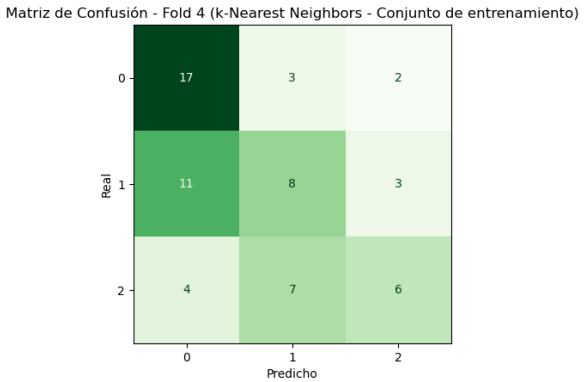

Figura D.271: k-Nearest Neighbors, Fold 4, C3, SMOTE

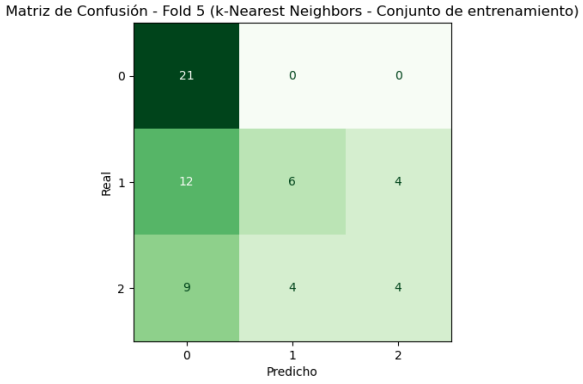

Figura D.271: k-Nearest Neighbors, Fold 5, C3, SMOTE

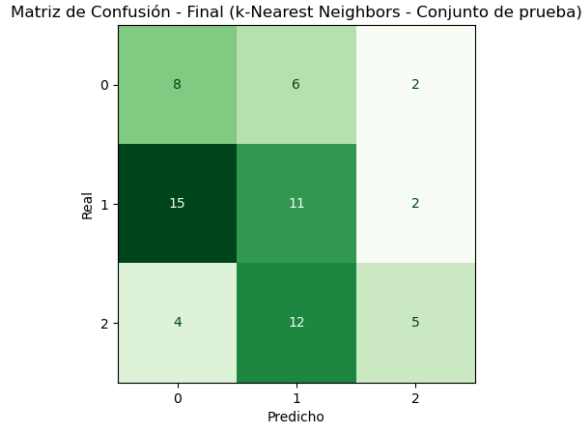

Figura D.272: k-Nearest Neighbors, Final, C3, SMOTE

Gaussian Naive Bayes

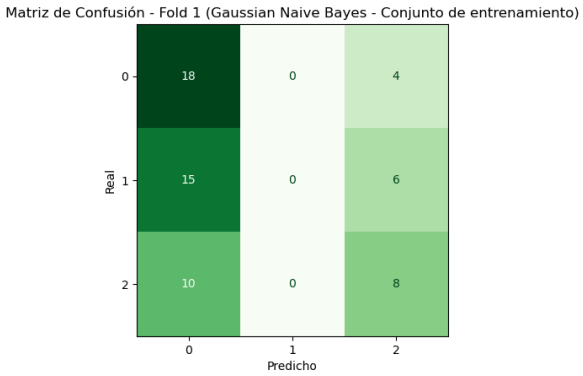

Figura D.273: Gaussian Naive Bayes, Fold 1, C3, SMOTE

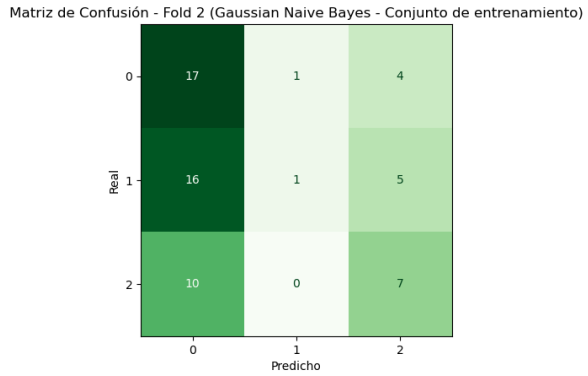

Figura D.274: Gaussian Naive Bayes, Fold 2, C3, SMOTE

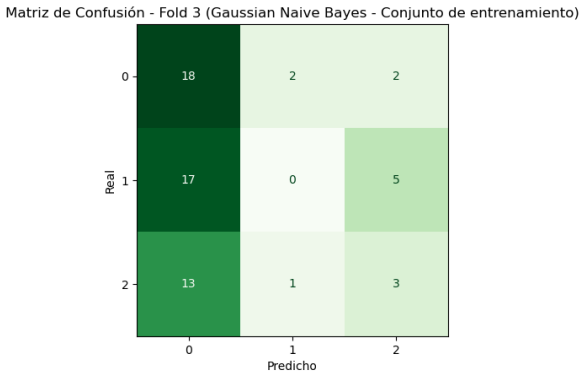

Figura D.274: Gaussian Naive Bayes, Fold 3, C3, SMOTE

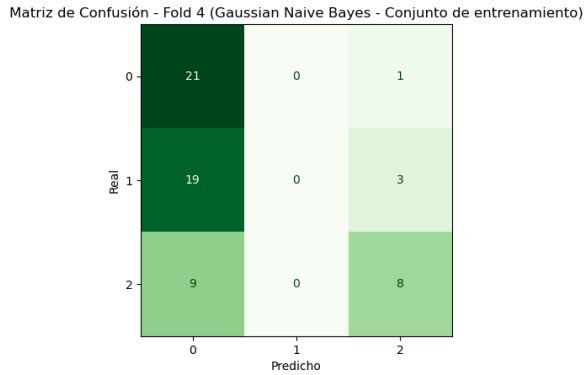

Figura D.275: Gaussian Naive Bayes, Fold 4, C3, SMOTE

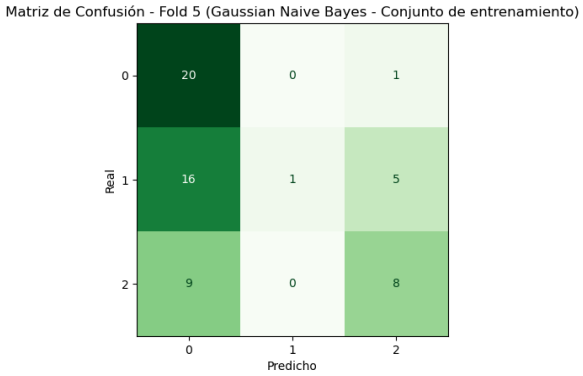

Figura D.275: Gaussian Naive Bayes, Fold 5, C3, SMOTE

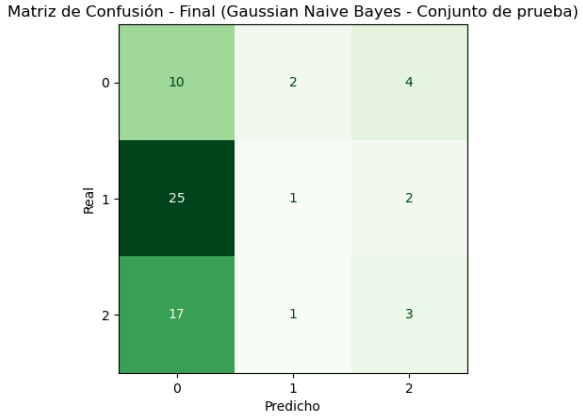

Figura D.276: Gaussian Naive Bayes, Final, C3, SMOTE

MLP

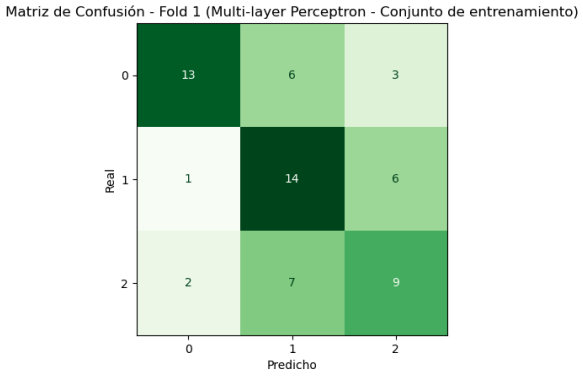

Figura D.277: MLP, Fold 1, C3, SMOTE

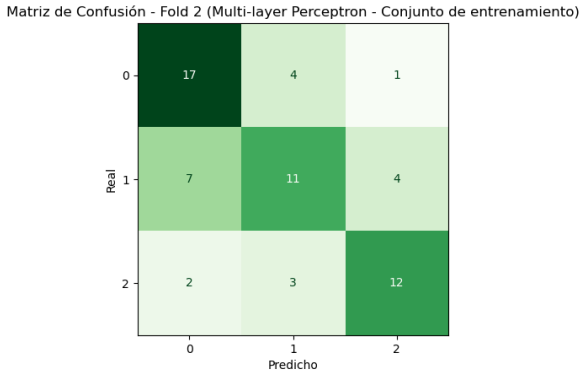

Figura D.278: MLP, Fold 2, C3, SMOTE

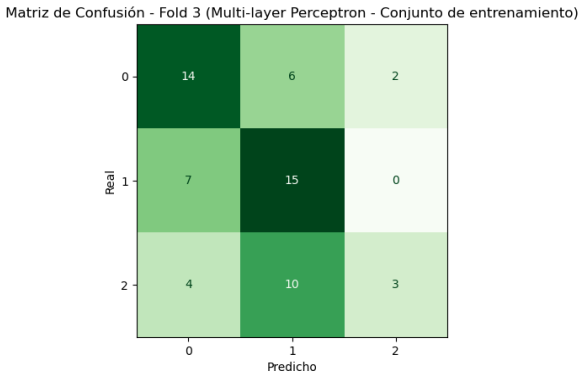

Figura D.278: MLP, Fold 3, C3, SMOTE

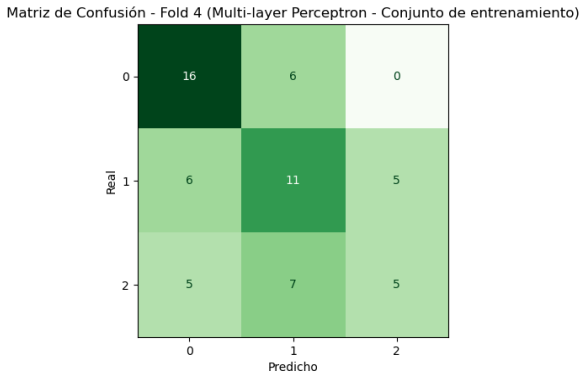

Figura D.279: MLP, Fold 4, C3, SMOTE

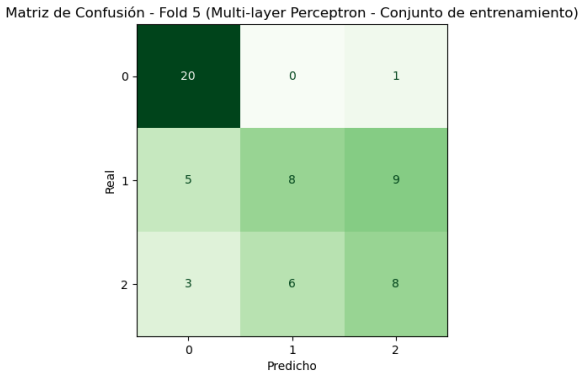

Figura D.279: MLP, Fold 5, C3, SMOTE

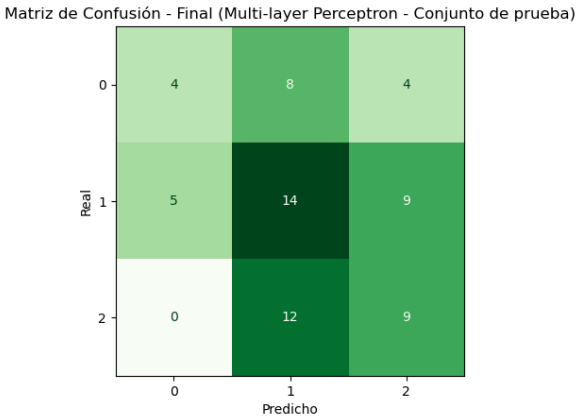

Figura D.280: MLP, Final, C3, SMOTE

D.4 Clasificación 4

D.4.1. ADASYN

Logistic Regression

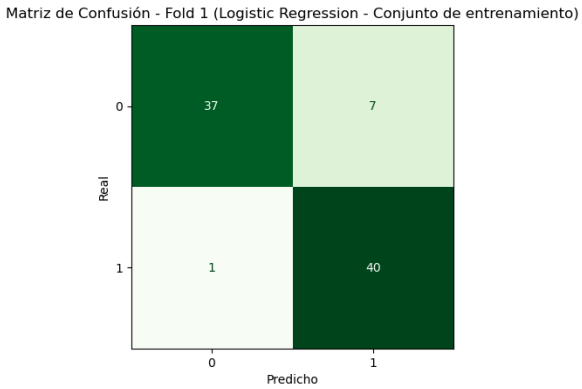

Figura D.281: Logistic Regression, Fold 1, C4, ADASYN

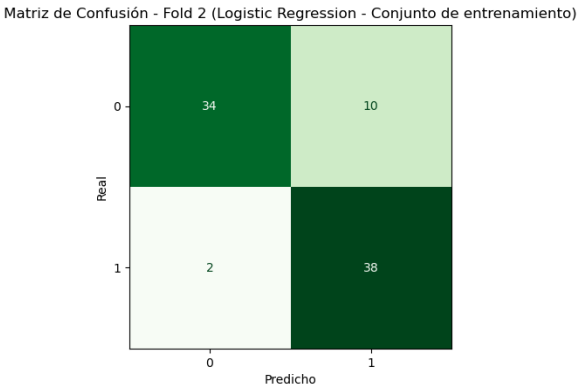

Figura D.282: Logistic Regression, Fold 2, C4, ADASYN

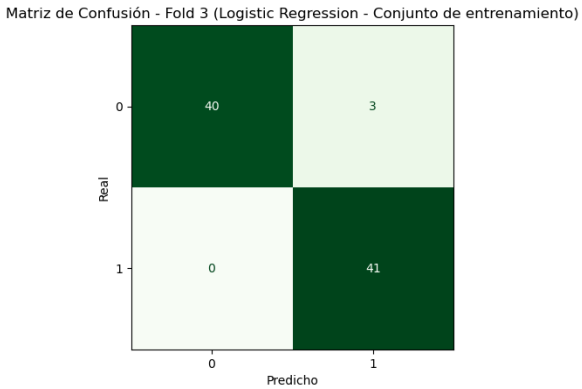

Figura D.282: Logistic Regression, Fold 3, C4, ADASYN

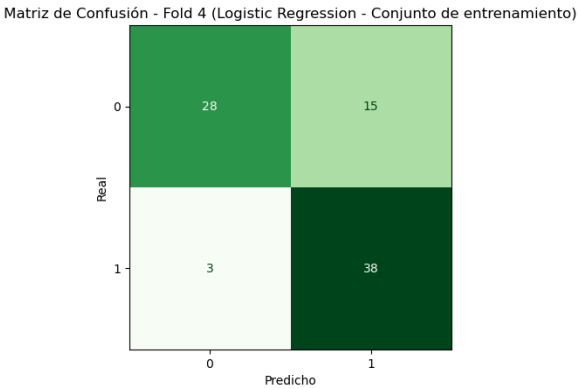

Figura D.283: Logistic Regression, Fold 4, C4, ADASYN

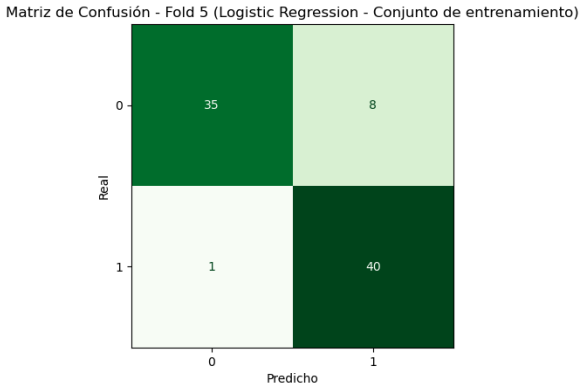

Figura D.283: Logistic Regression, Fold 5, C4, ADASYN

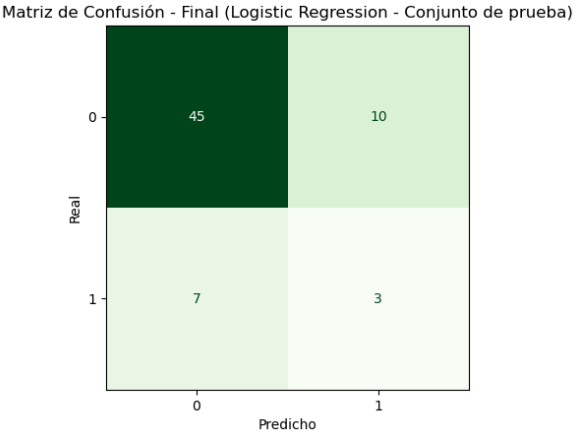

Figura D.284: Logistic Regression, Final, C4, ADASYN

Logistic Regression Lasso

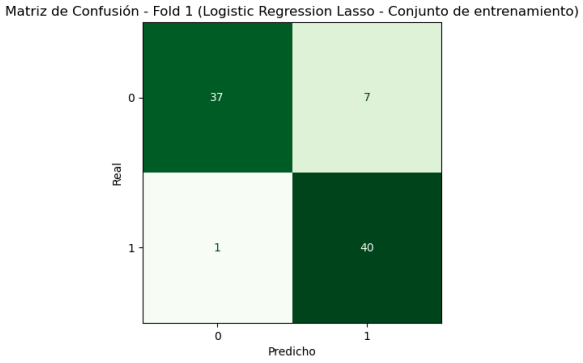

Figura D.285: Logistic Regression Lasso, Fold 1, C4, ADASYN

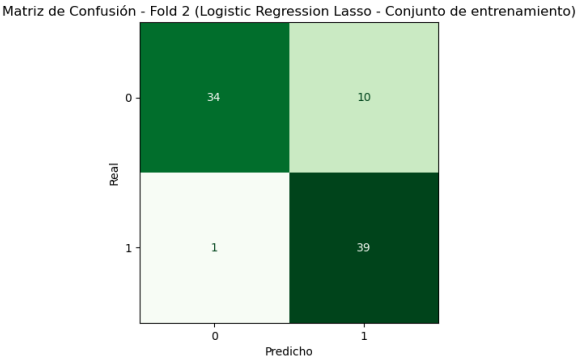

Figura D.286: Logistic Regression Lasso, Fold 2, C4, ADASYN

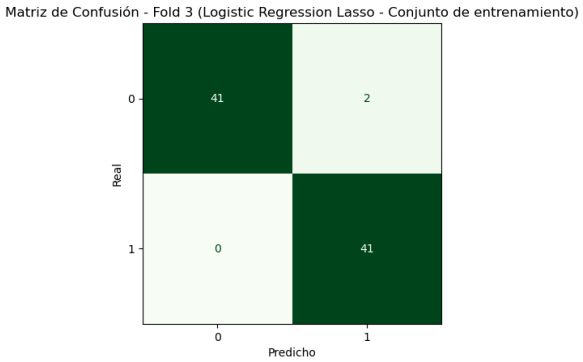

Figura D.286: Logistic Regression Lasso, Fold 3, C4, ADASYN

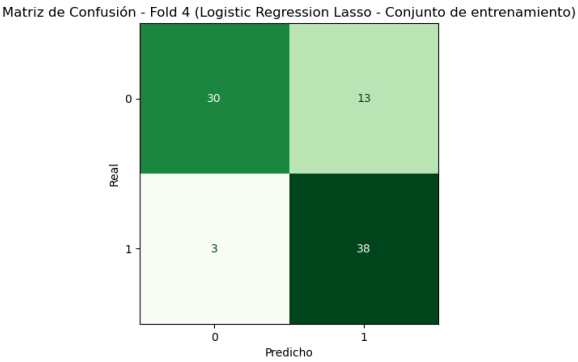

Figura D.287: Logistic Regression Lasso, Fold 4, C4, ADASYN

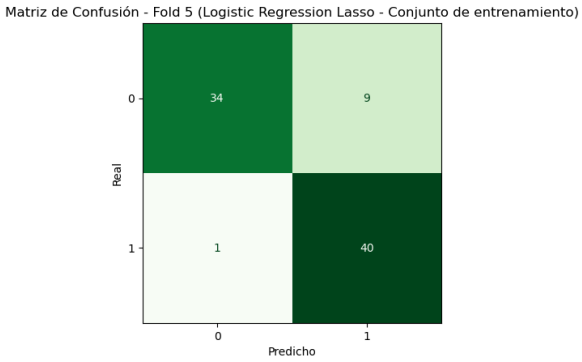

Figura D.287: Logistic Regression Lasso, Fold 5, C4, ADASYN

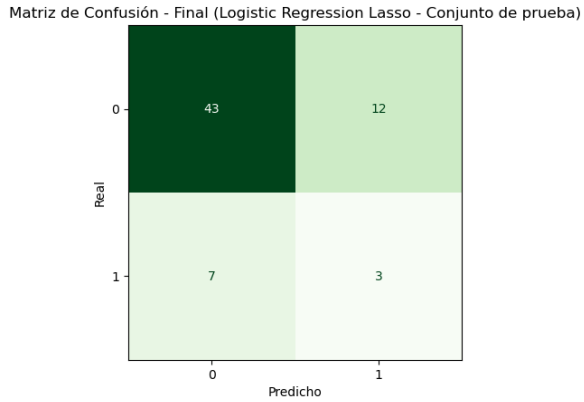

Figura D.288: Logistic Regression Lasso, Final, C4, ADASYN

Ridge

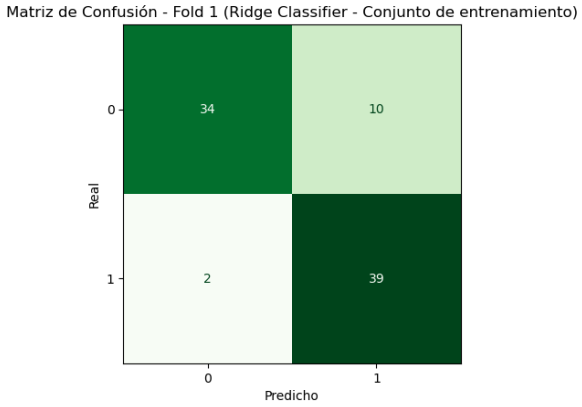

Figura D.289: Ridge, Fold 1, C4, ADASYN

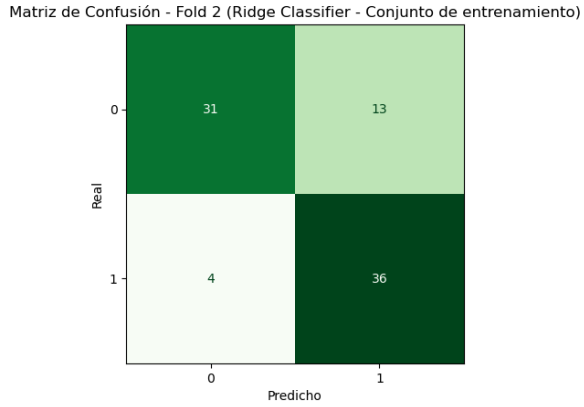

Figura D.290: Ridge, Fold 2, C4, ADASYN

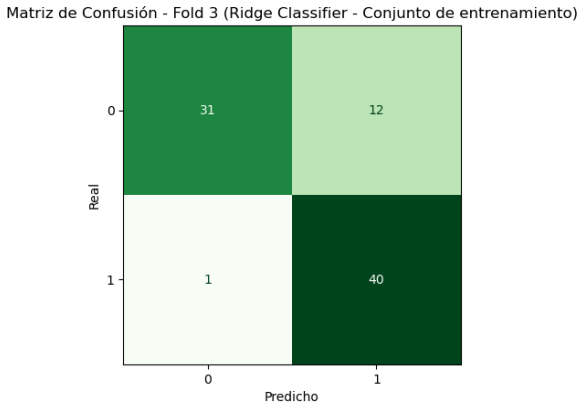

Figura D.290: Ridge, Fold 3, C4, ADASYN

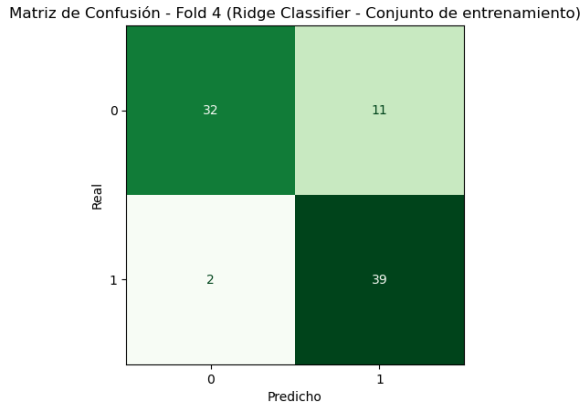

Figura D.291: Ridge, Fold 4, C4, ADASYN

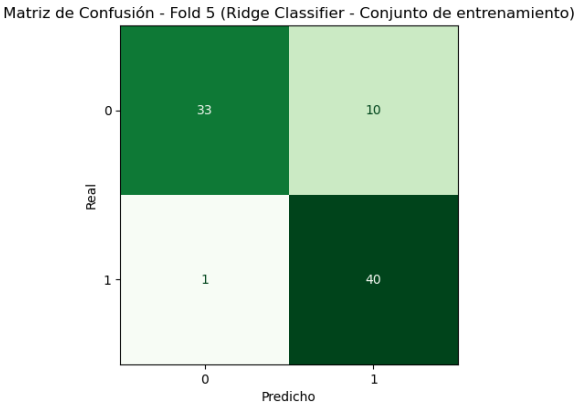

Figura D.291: Ridge, Fold 5, C4, ADASYN

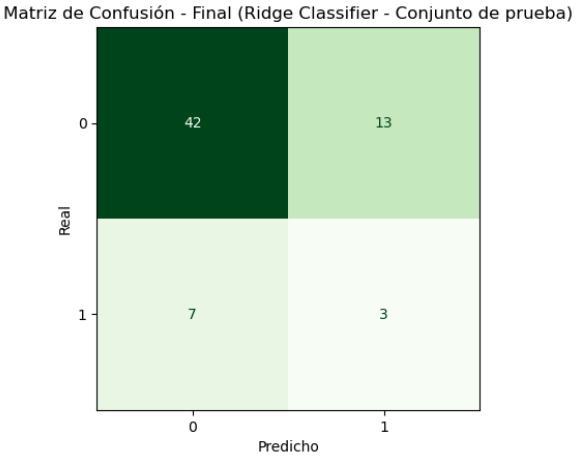

Figura D.292: Ridge, Final, C4, ADASYN

SDG

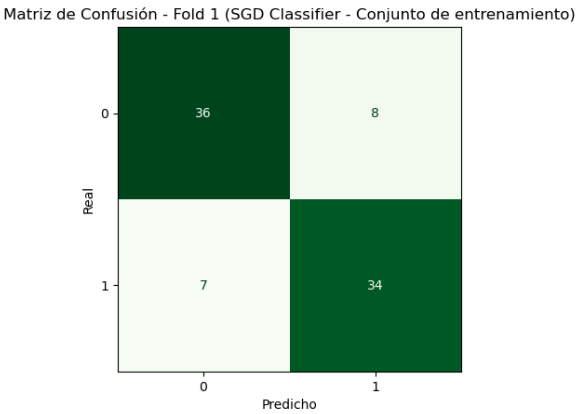

Figura D.293: SDG, Fold 1, C4, ADASYN

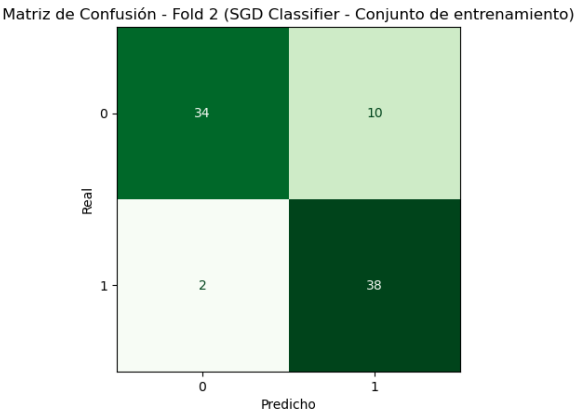

Figura D.294: SDG, Fold 2, C4, ADASYN

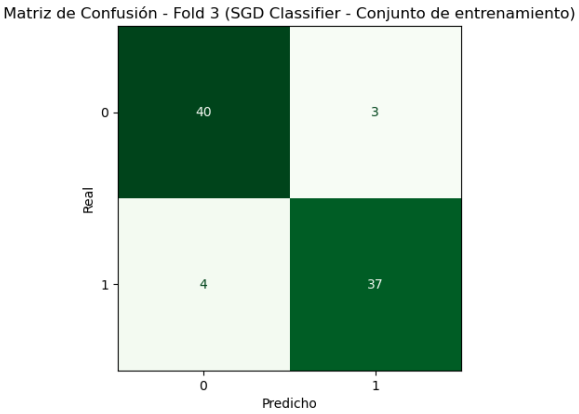

Figura D.294: SDG, Fold 3, C4, ADASYN

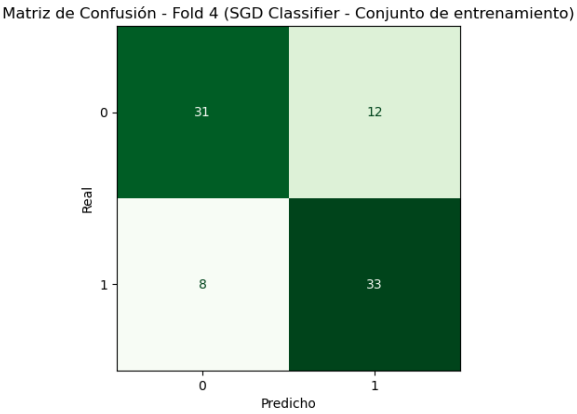

Figura D.295: SDG, Fold 4, C4, ADASYN

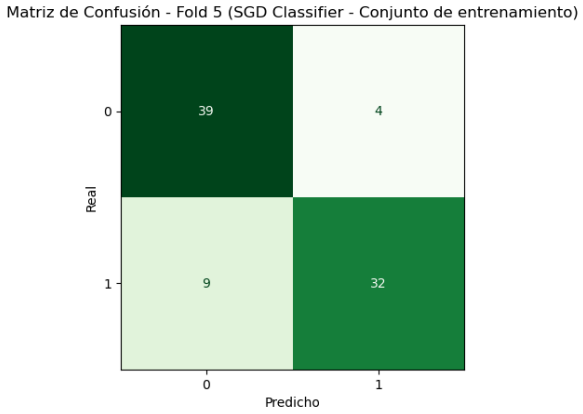

Figura D.295: SDG, Fold 5, C4, ADASYN

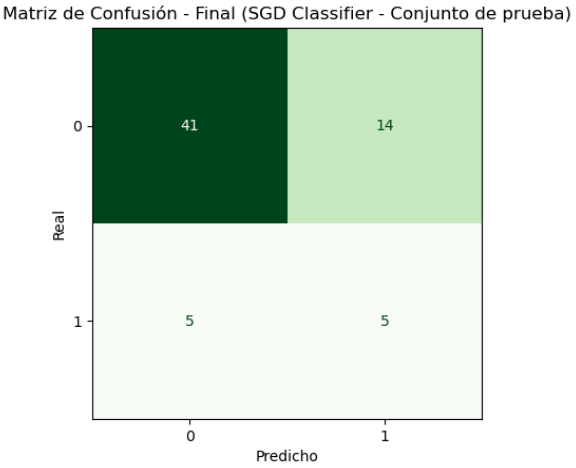

Figura D.296: SDG, Final, C4, ADASYN

Decision Tree

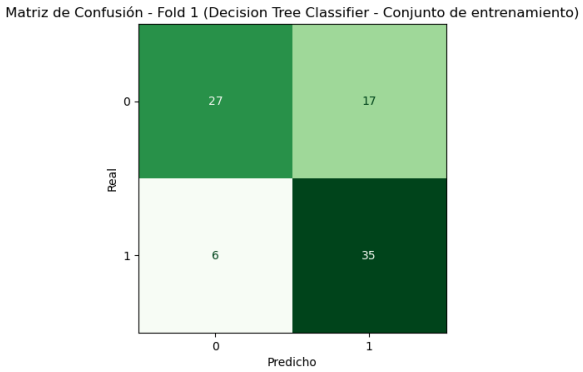

Figura D.297: Decision Tree, Fold 1, C4, ADASYN

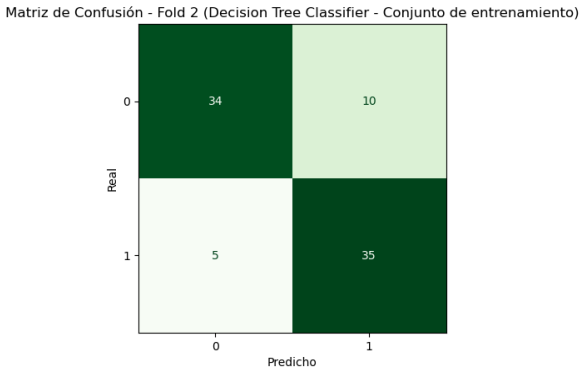

Figura D.298: Decision Tree, Fold 2, C4, ADASYN

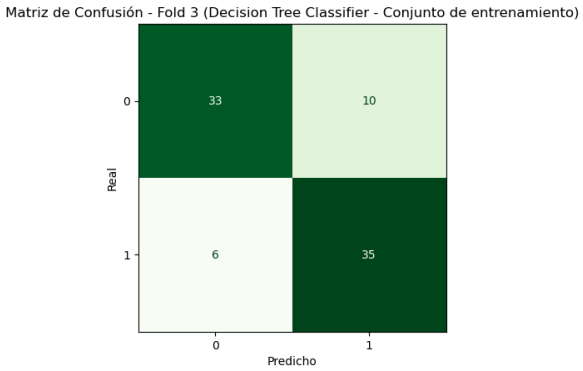

Figura D.298: Decision Tree, Fold 3, C4, ADASYN

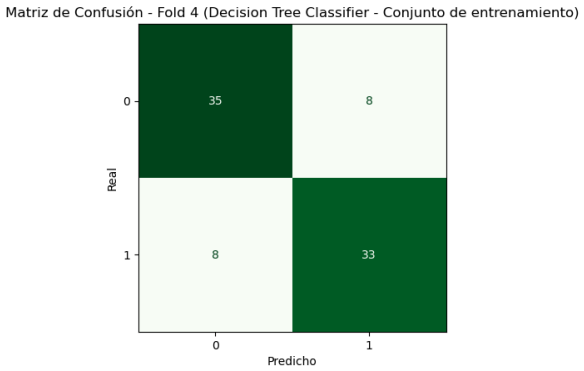

Figura D.299: Decision Tree, Fold 4, C4, ADASYN

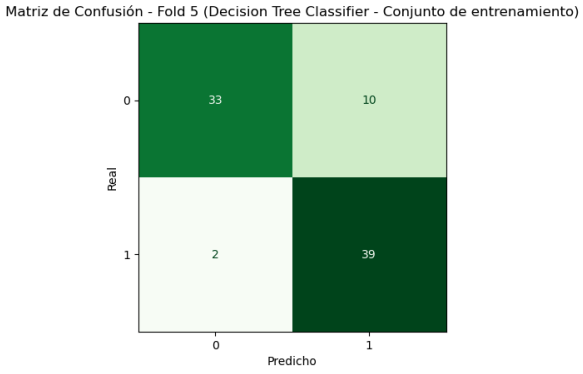

Figura D.299: Decision Tree, Fold 5, C4, ADASYN

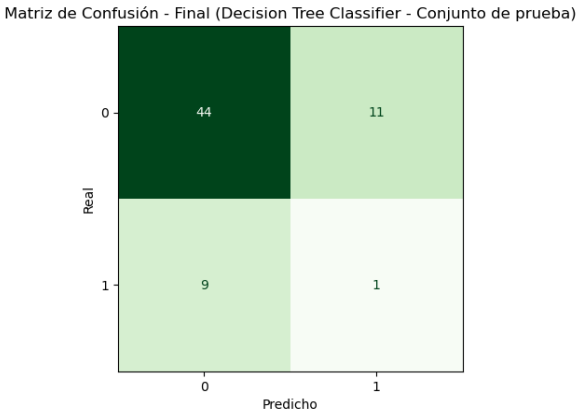

Figura D.300: Decision Tree, Final, C4, ADASYN

Random Forest

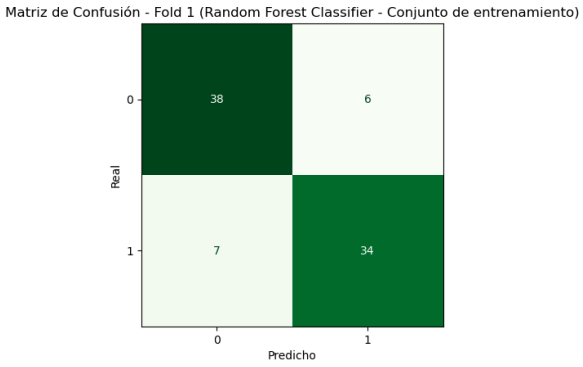

Figura D.301: Random Forest, Fold 1, C4, ADASYN

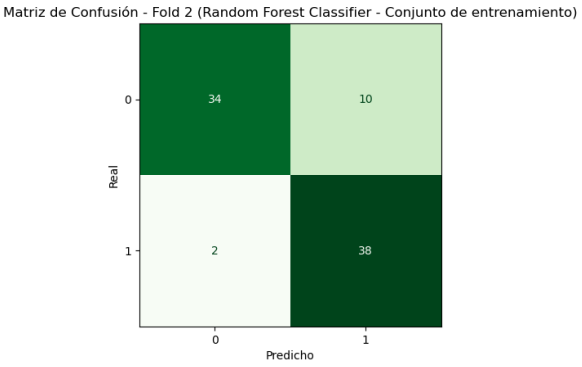

Figura D.302: Random Forest, Fold 2, C4, ADASYN

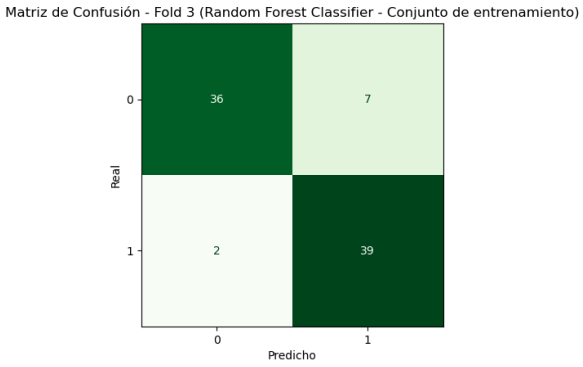

Figura D.302: Random Forest, Fold 3, C4, ADASYN

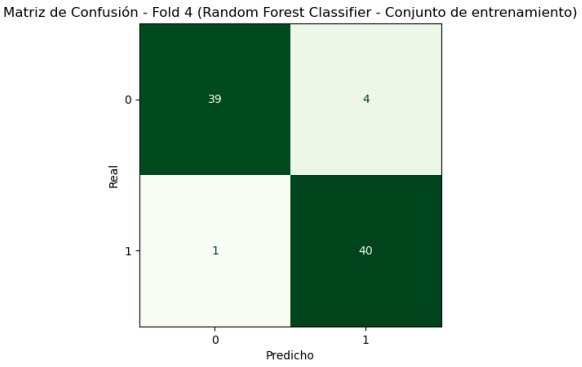

Figura D.303: Random Forest, Fold 4, C4, ADASYN

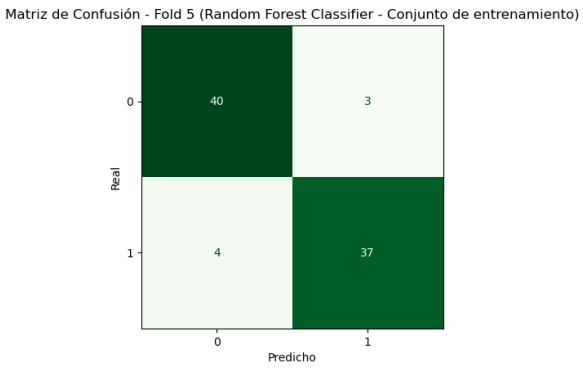

Figura D.303: Random Forest, Fold 5, C4, ADASYN

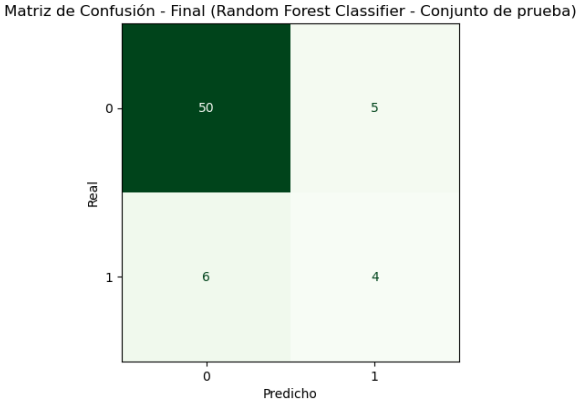

Figura D.304: Random Forest, Final, C4, ADASYN

AdaBoost

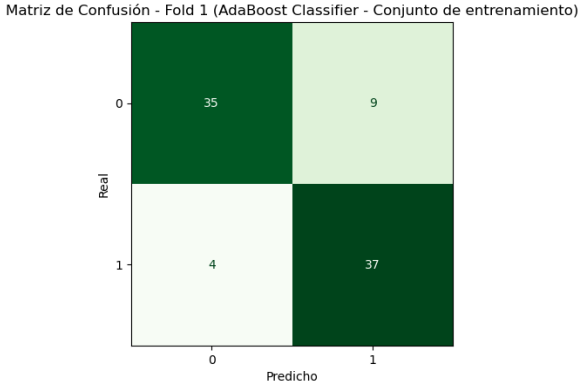

Figura D.305: AdaBoost, Fold 1, C4, ADASYN

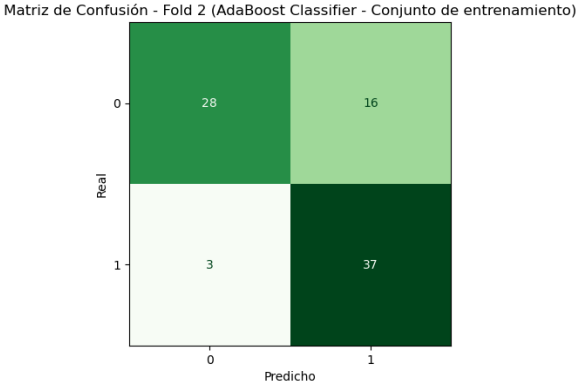

Figura D.306: AdaBoost, Fold 2, C4, ADASYN

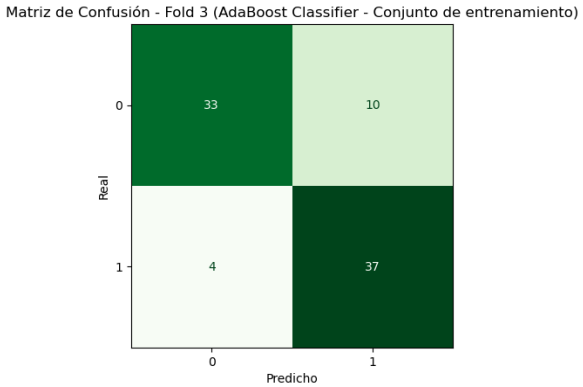

Figura D.306: AdaBoost, Fold 3, C4, ADASYN

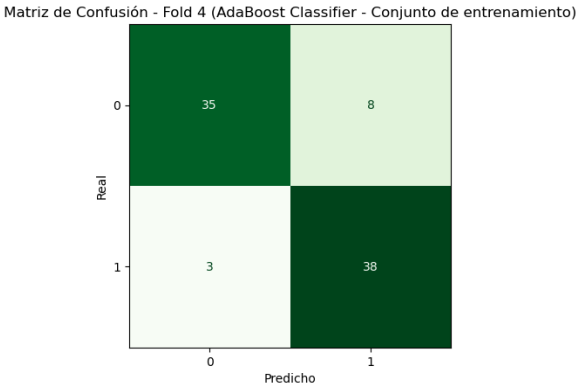

Figura D.307: AdaBoost, Fold 4, C4, ADASYN

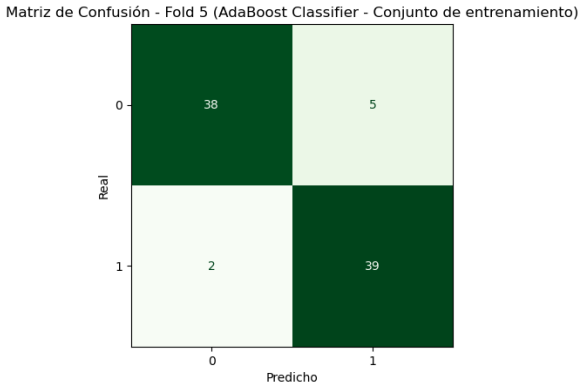

Figura D.307: AdaBoost, Fold 5, C4, ADASYN

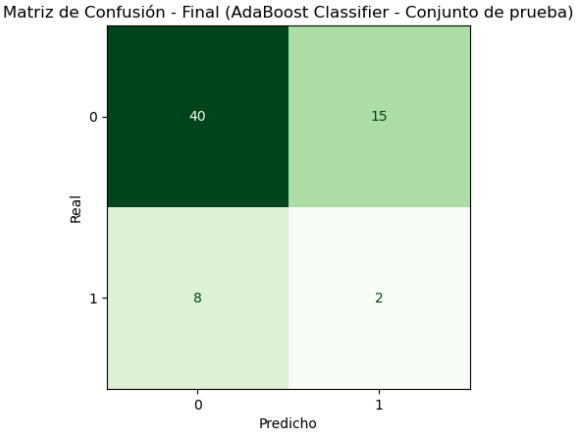

Figura D.308: AdaBoost, Final, C4, ADASYN

Gradient Boosting

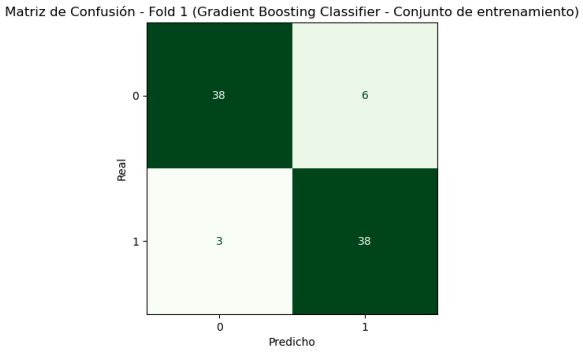

Figura D.309: Gradient Boosting, Fold 1, C4, ADASYN

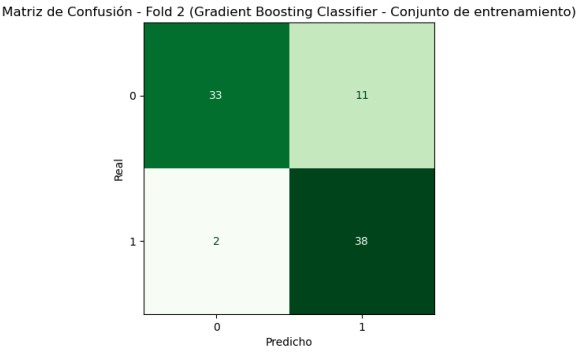

Figura D.310: Gradient Boosting, Fold 2, C4, ADASYN

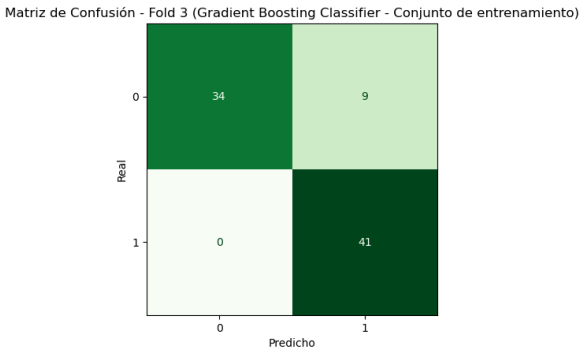

Figura D.310: Gradient Boosting, Fold 3, C4, ADASYN

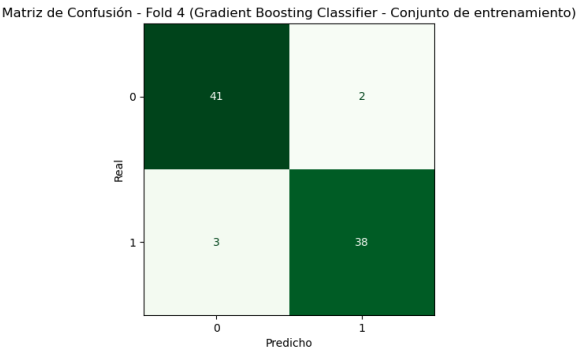

Figura D.311: Gradient Boosting, Fold 4, C4, ADASYN

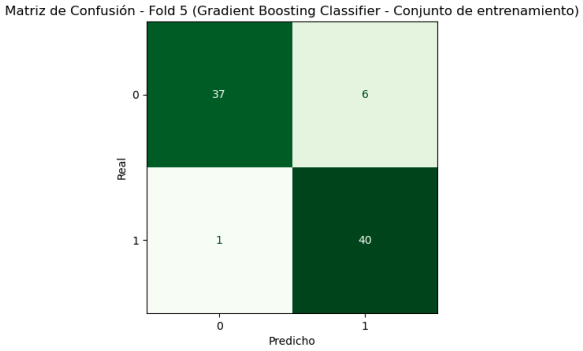

Figura D.311: Gradient Boosting, Fold 5, C4, ADASYN

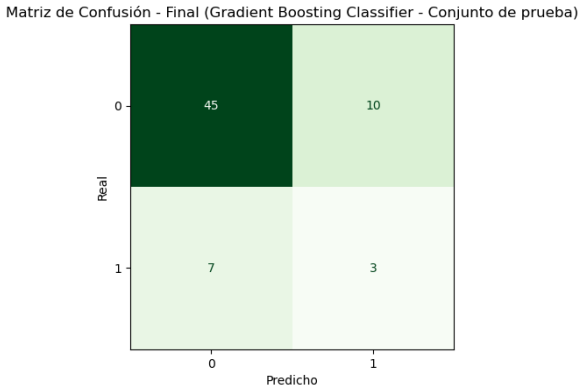

Figura D.312: Gradient Boosting, Final, C4, ADASYN

SVC

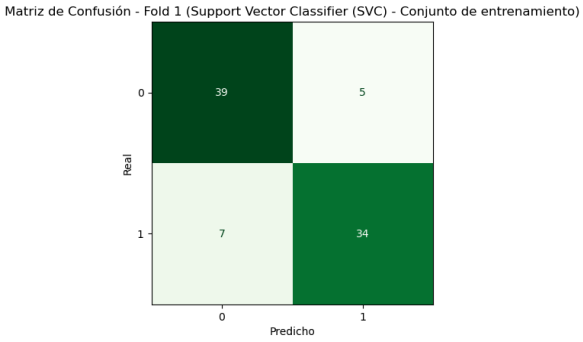

Figura D.313: SVC, Fold 1, C4, ADASYN

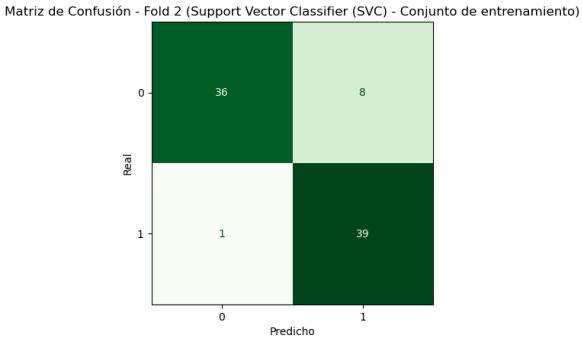

Figura D.314: SVC, Fold 2, C4, ADASYN

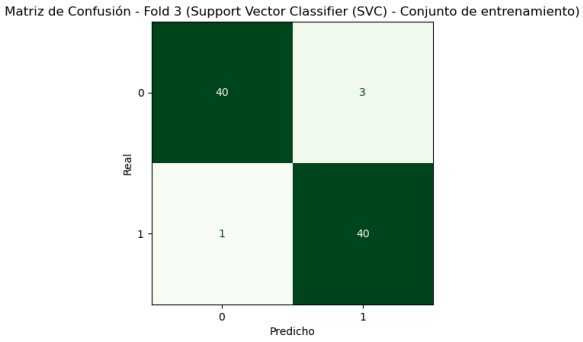

Figura D.314: SVC, Fold 3, C4, ADASYN

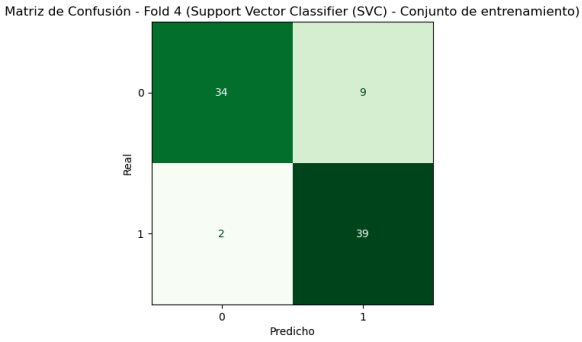

Figura D.315: SVC, Fold 4, C4, ADASYN

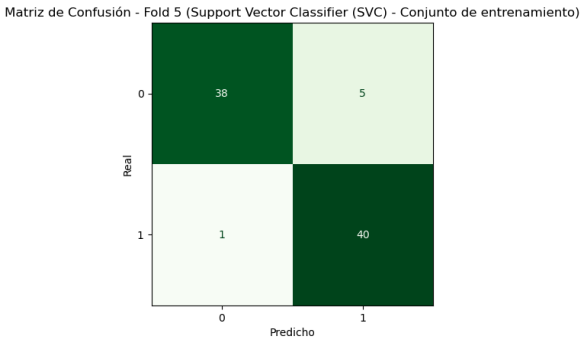

Figura D.315: SVC, Fold 5, C4, ADASYN

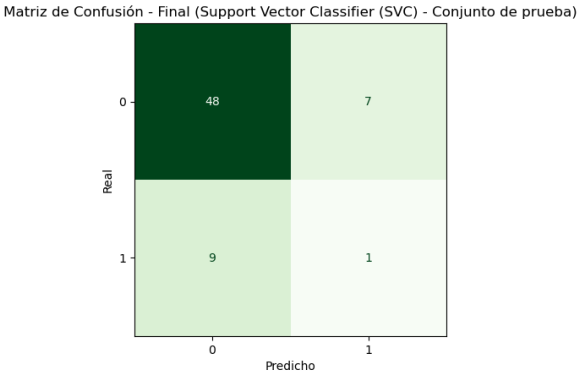

Figura D.316: SVC, Final, C4, ADASYN

k-Nearest Neighbors

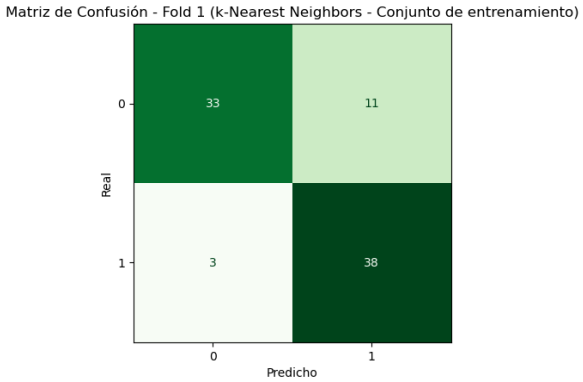

Figura D.317: k-Nearest Neighbors, Fold 1, C4, ADASYN

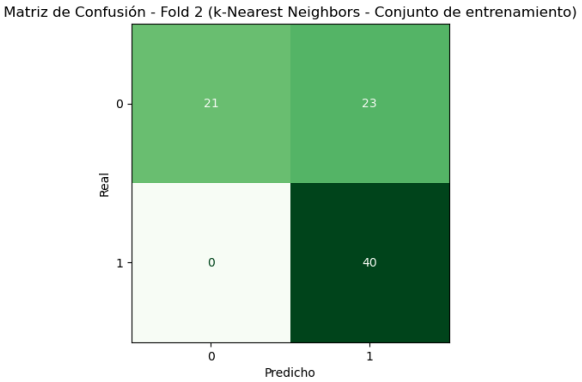

Figura D.318: k-Nearest Neighbors, Fold 2, C4, ADASYN

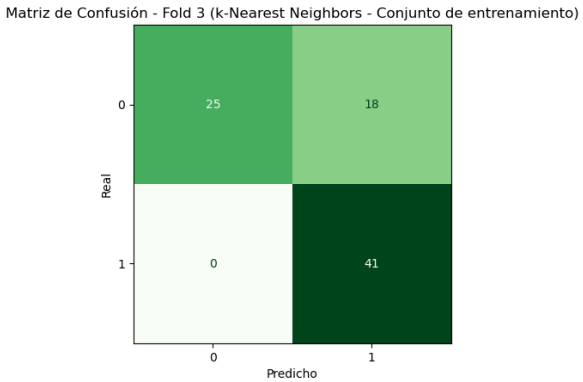

Figura D.318: k-Nearest Neighbors, Fold 3, C4, ADASYN

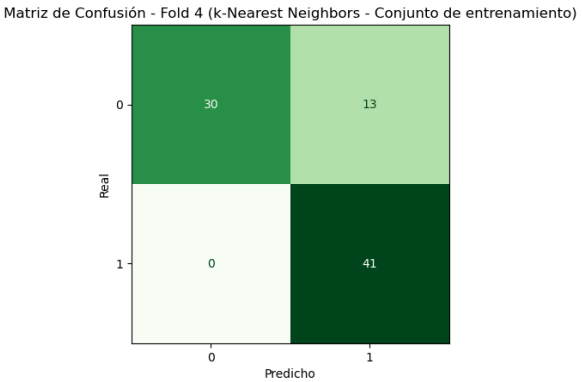

Figura D.319: k-Nearest Neighbors, Fold 4, C4, ADASYN

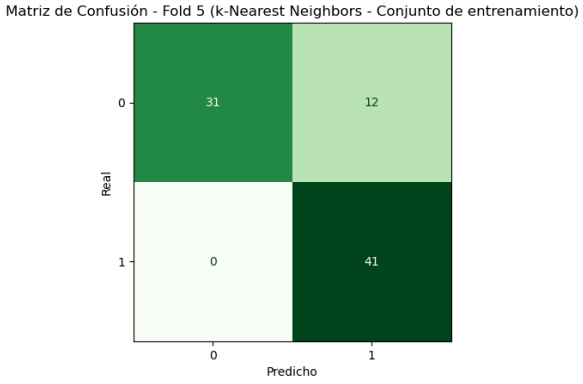

Figura D.319: k-Nearest Neighbors, Fold 5, C4, ADASYN

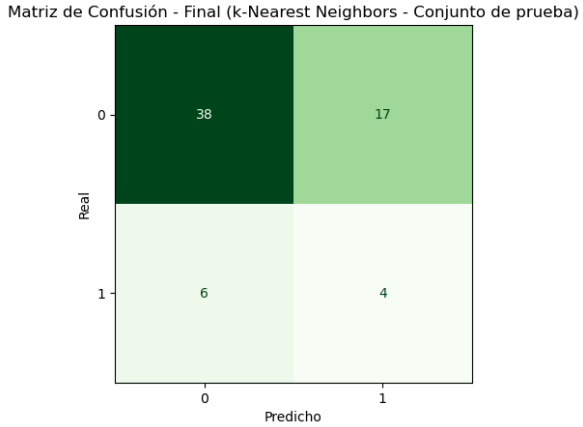

Figura D.320: k-Nearest Neighbors, Final, C4, ADASYN

Gaussian Naive Bayes

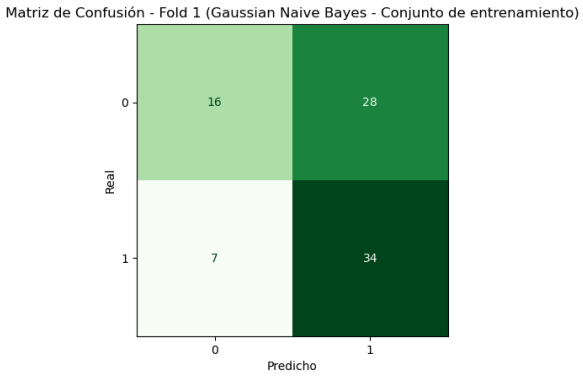

Figura D.321: Gaussian Naive Bayes, Fold 1, C4, ADASYN

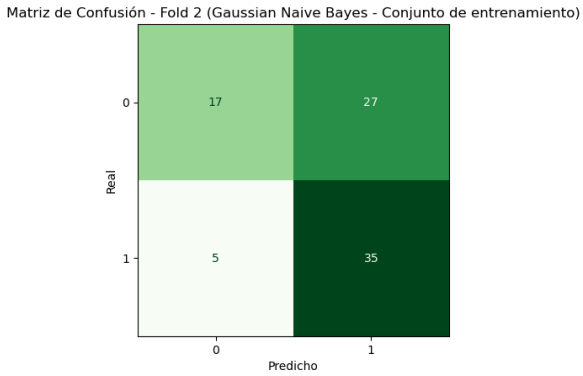

Figura D.322: Gaussian Naive Bayes, Fold 2, C4, ADASYN

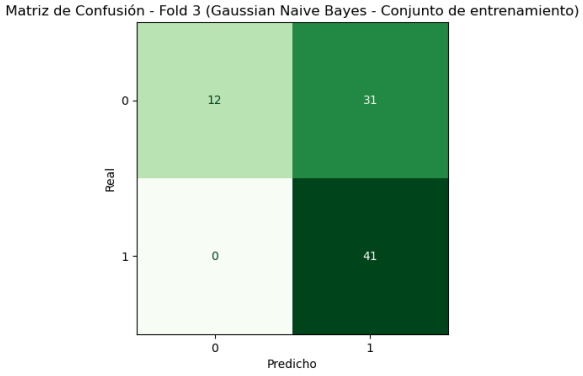

Figura D.322: Gaussian Naive Bayes, Fold 3, C4, ADASYN

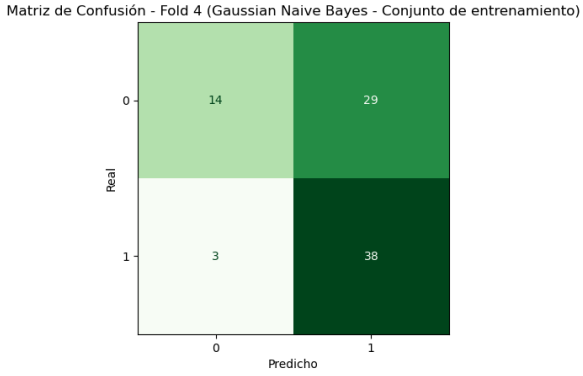

Figura D.323: Gaussian Naive Bayes, Fold 4, C4, ADASYN

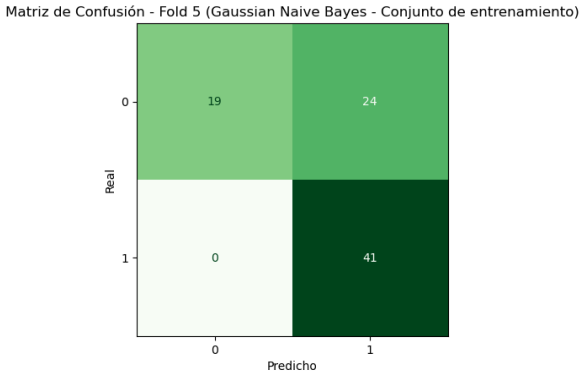

Figura D.323: Gaussian Naive Bayes, Fold 5, C4, ADASYN

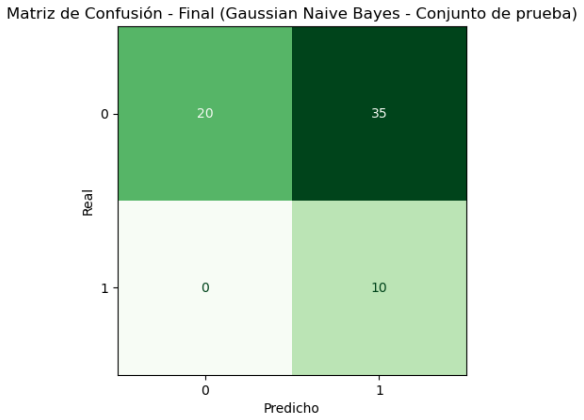

Figura D.324: Gaussian Naive Bayes, Final, C4, ADASYN

MLP

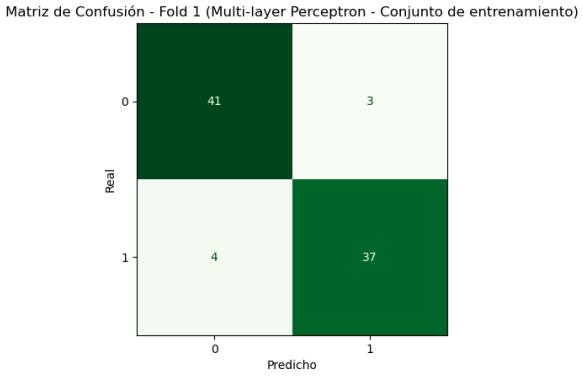

Figura D.325: MLP, Fold 1, C4, ADASYN

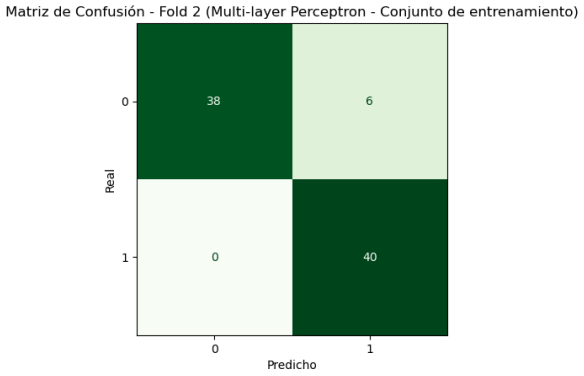

Figura D.326: MLP, Fold 2, C4, ADASYN

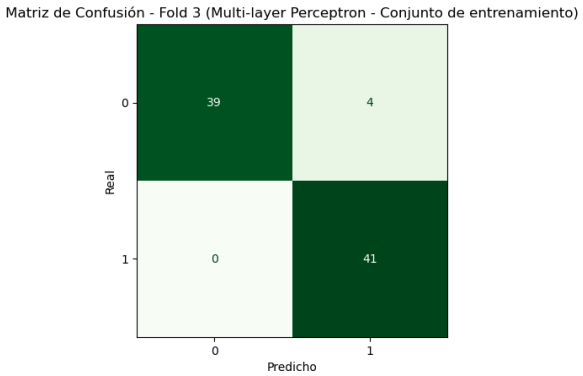

Figura D.326: MLP, Fold 3, C4, ADASYN

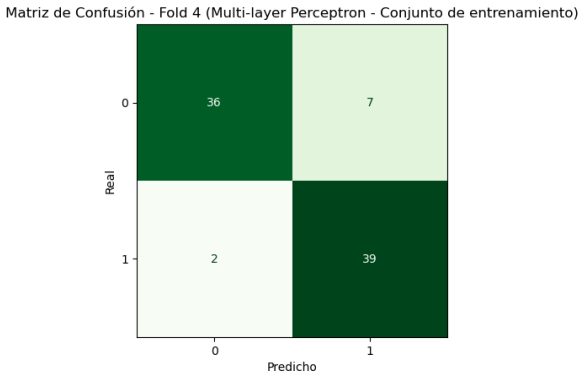

Figura D.327: MLP, Fold 4, C4, ADASYN

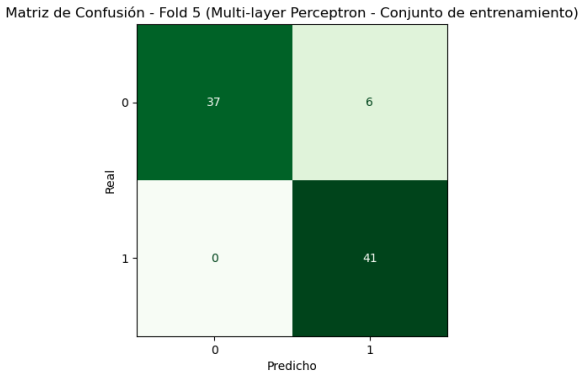

Figura D.327: MLP, Fold 5, C4, ADASYN

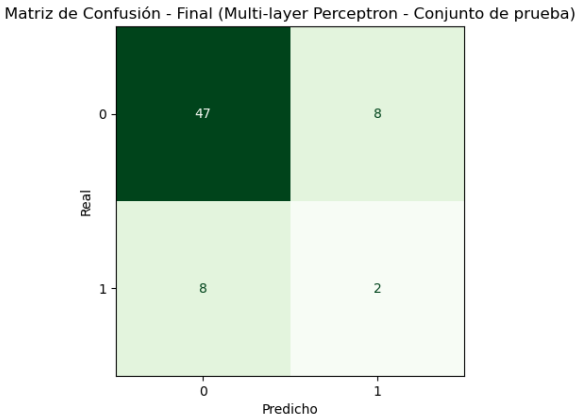

Figura D.328: MLP, Final, C4, ADASYN

D.4.2. SMOTE

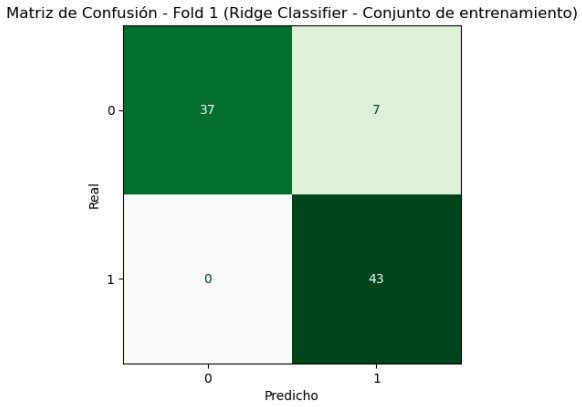

Figura D.329: Ridge, Fold 1, C4, SMOTE

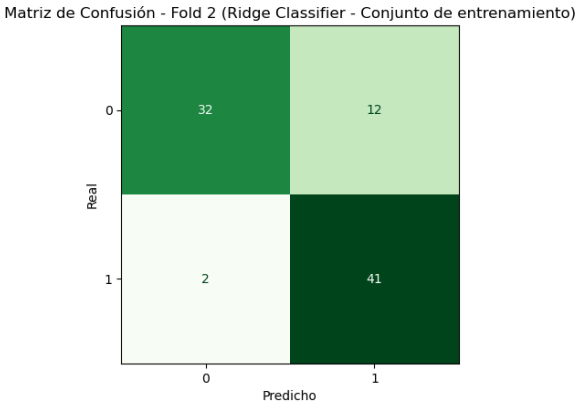

Figura D.330: Ridge, Fold 2, C4, SMOTE

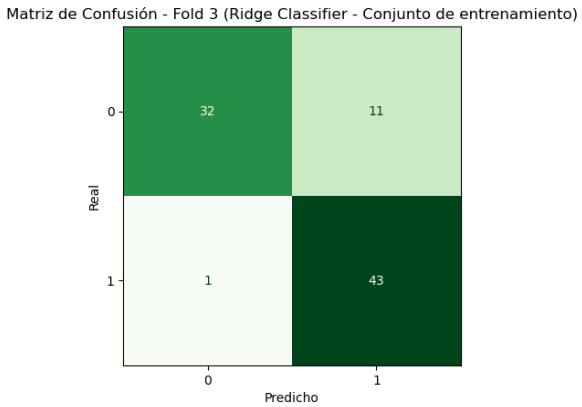

Figura D.330: Ridge, Fold 3, C4, SMOTE

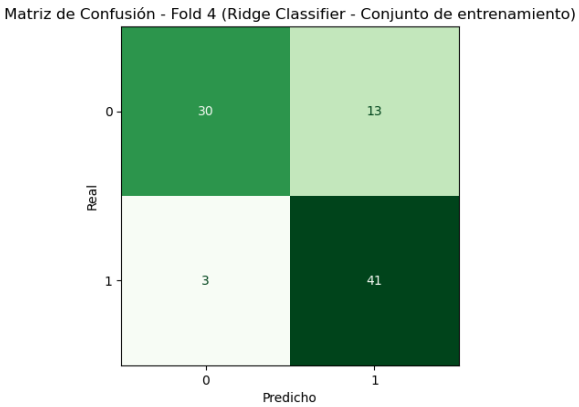

Figura D.331: Ridge, Fold 4, C4, SMOTE

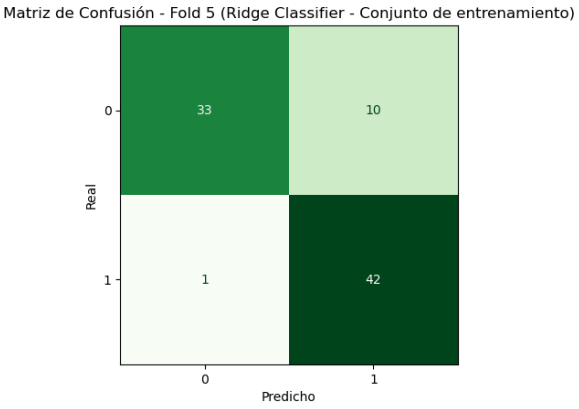

Figura D.331: Ridge, Fold 5, C4, SMOTE

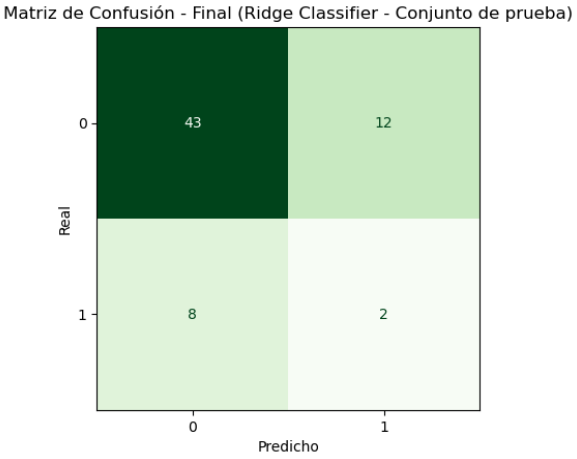

Figura D.332: Ridge, Final, C4, SMOTE

SDG

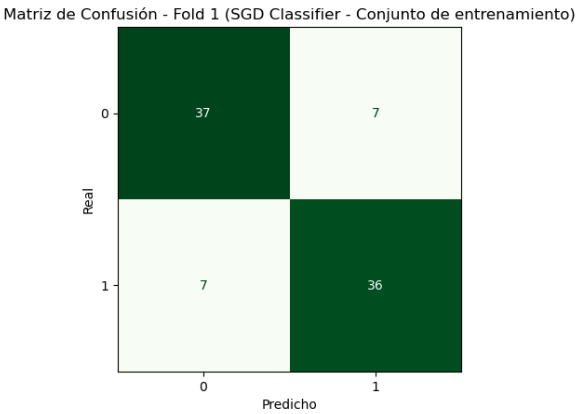

Figura D.333: SDG, Fold 1, C4, SMOTE

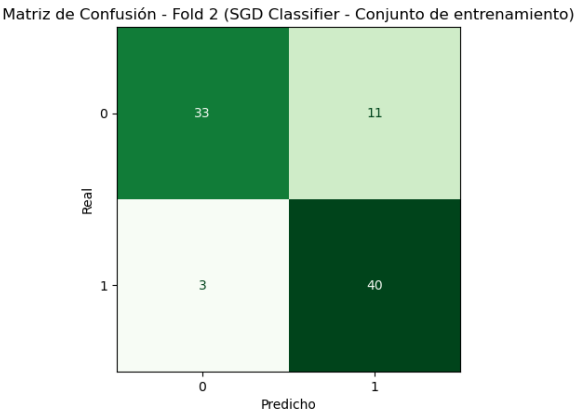

Figura D.334: SDG, Fold 2, C4, SMOTE

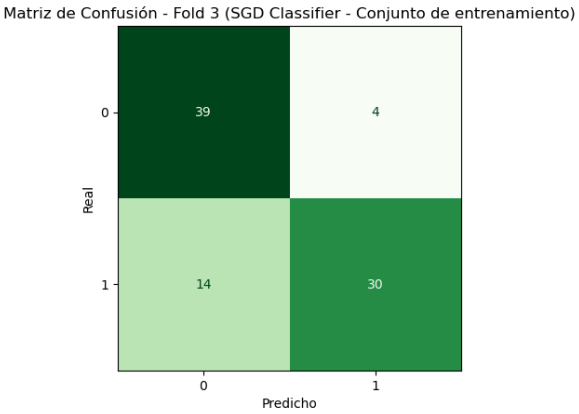

Figura D.334: SDG, Fold 3, C4, SMOTE

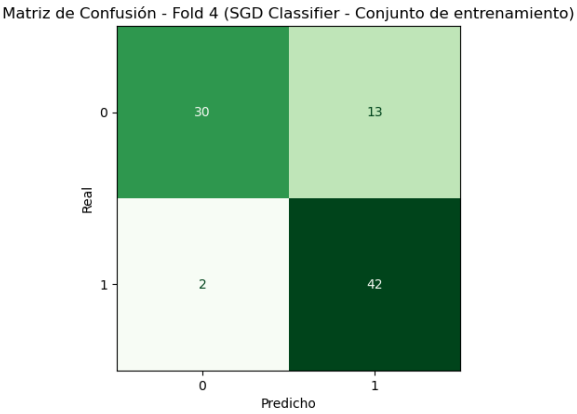

Figura D.335: SDG, Fold 4, C4, SMOTE

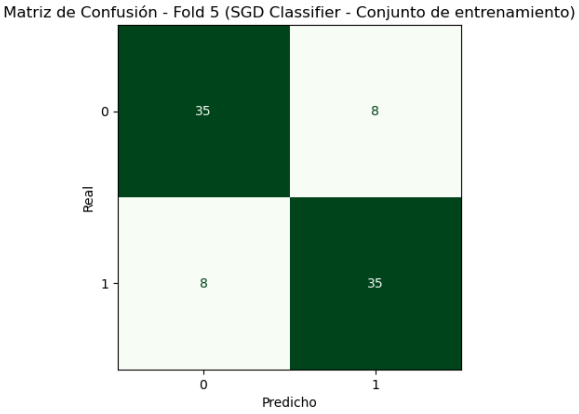

Figura D.335: SDG, Fold 5, C4, SMOTE

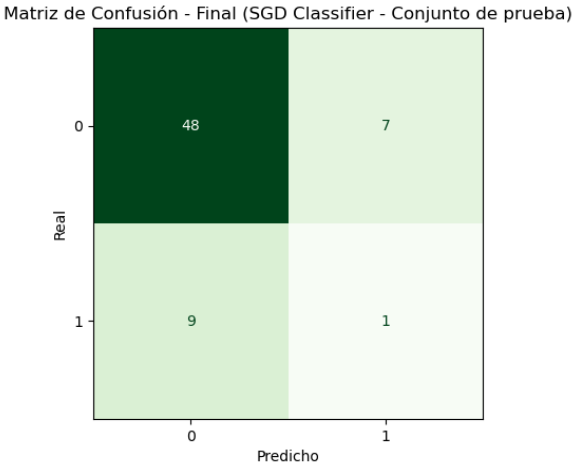

Figura D.336: SDG, Final, C4, SMOTE

Decision Tree

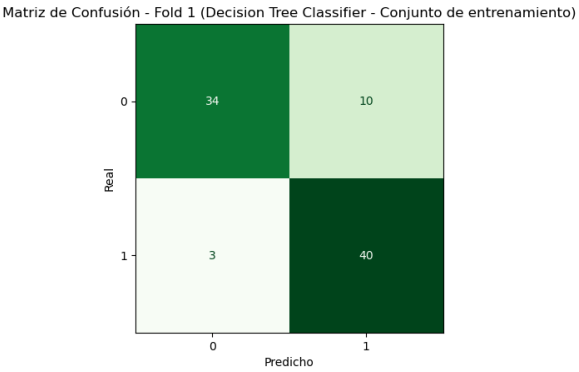

Figura D.337: Decision Tree, Fold 1, C4, SMOTE

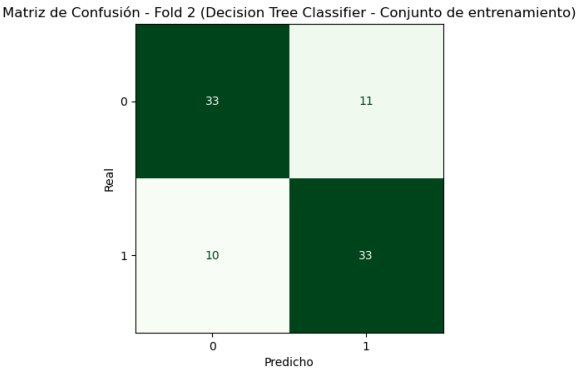

Figura D.338: Decision Tree, Fold 2, C4, SMOTE

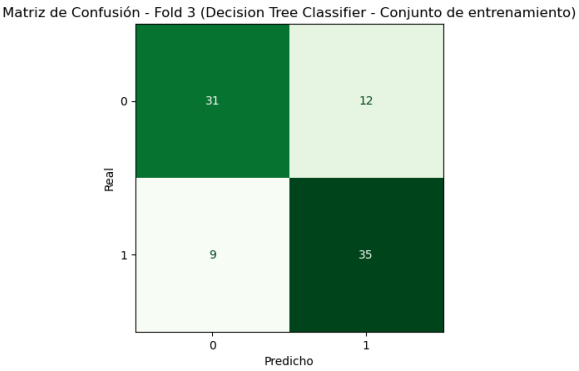

Figura D.338: Decision Tree, Fold 3, C4, SMOTE

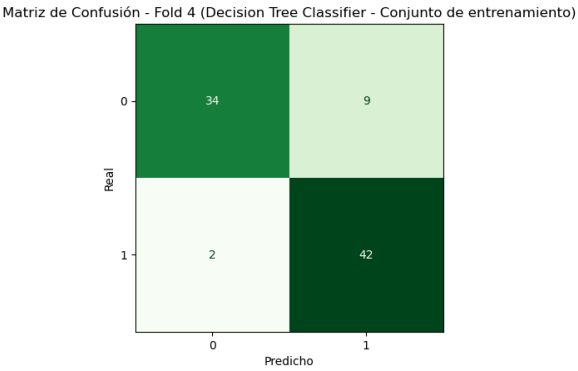

Figura D.339: Decision Tree, Fold 4, C4, SMOTE

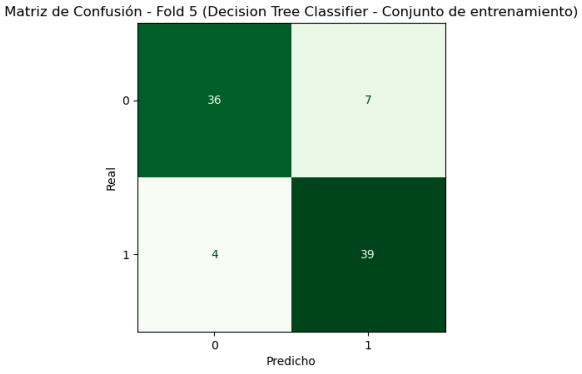

Figura D.339: Decision Tree, Fold 5, C4, SMOTE

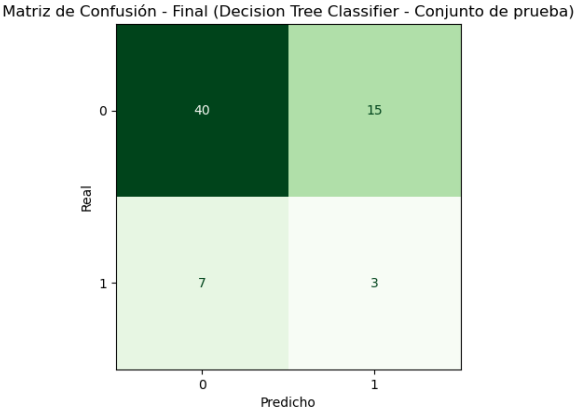

Figura D.340: Decision Tree, Final, C4, SMOTE

Random Forest

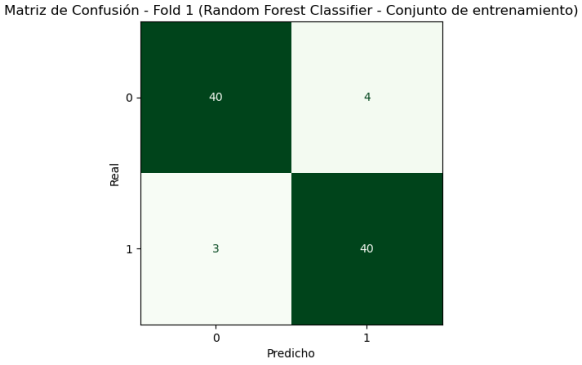

Figura D.341: Random Forest, Fold 1, C4, SMOTE

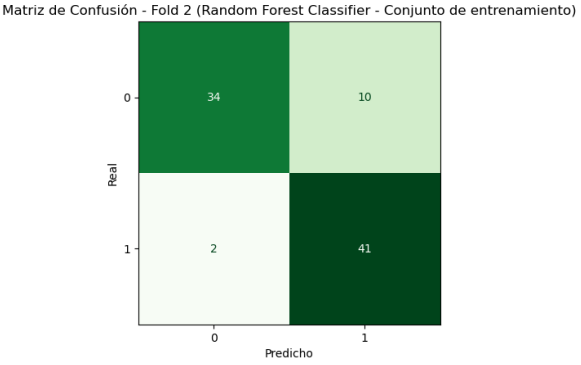

Figura D.342: Random Forest, Fold 2, C4, SMOTE

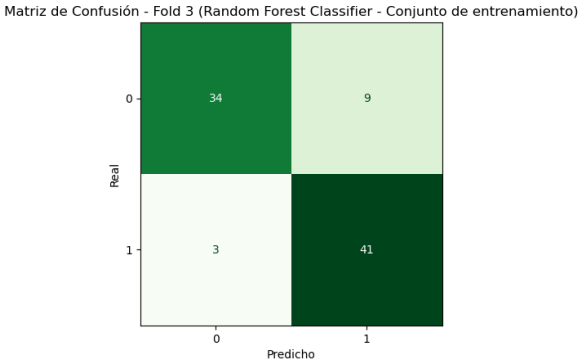

Figura D.342: Random Forest, Fold 3, C4, SMOTE

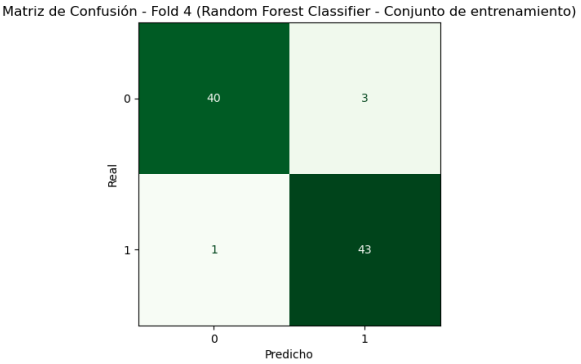

Figura D.343: Random Forest, Fold 4, C4, SMOTE

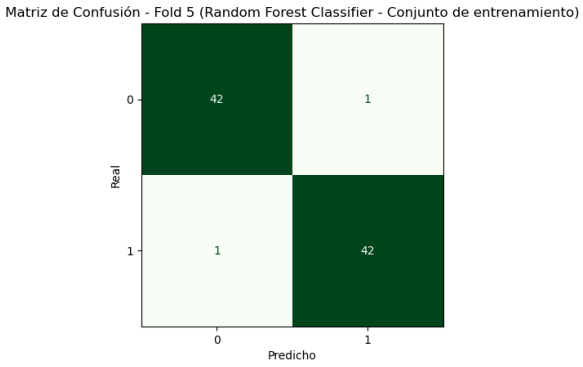

Figura D.343: Random Forest, Fold 5, C4, SMOTE

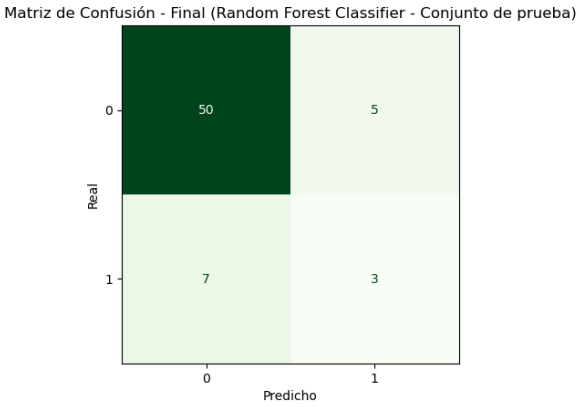

Figura D.344: Random Forest, Final, C4, SMOTE

AdaBoost

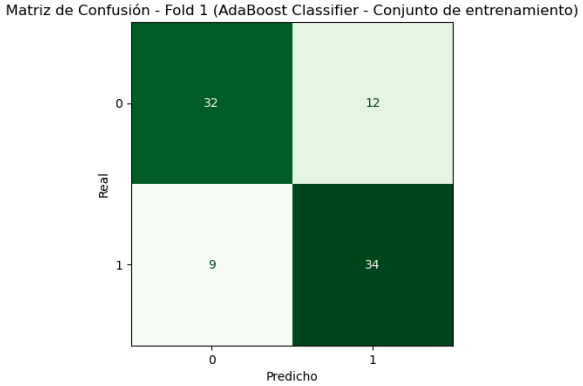

Figura D.345: AdaBoost, Fold 1, C4, SMOTE

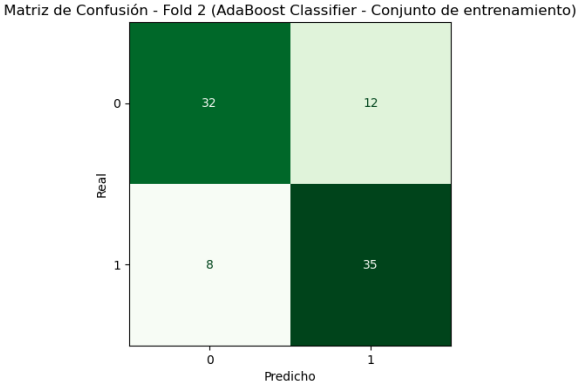

Figura D.346: AdaBoost, Fold 2, C4, SMOTE

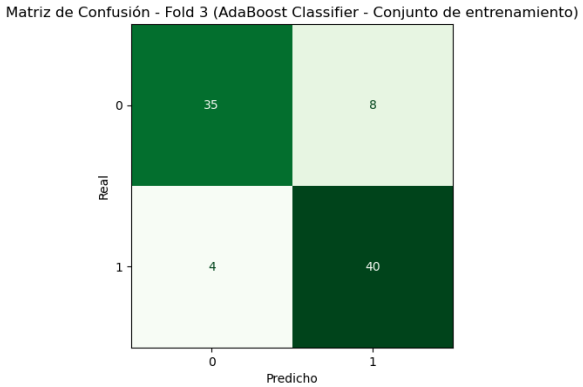

Figura D.346: AdaBoost, Fold 3, C4, SMOTE

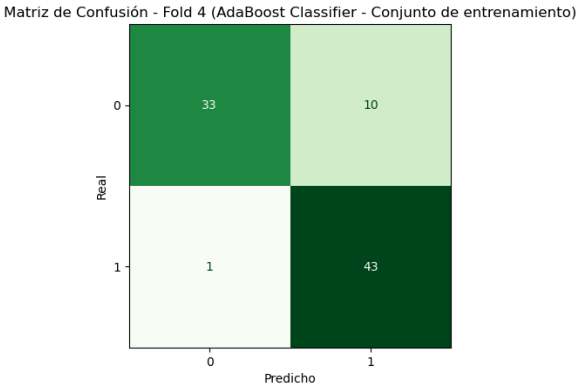

Figura D.347: AdaBoost, Fold 4, C4, SMOTE

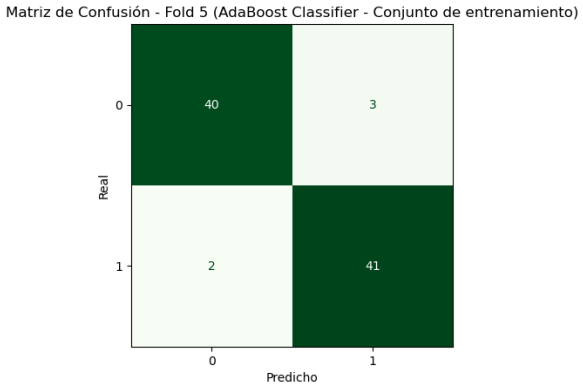

Figura D.347: AdaBoost, Fold 5, C4, SMOTE

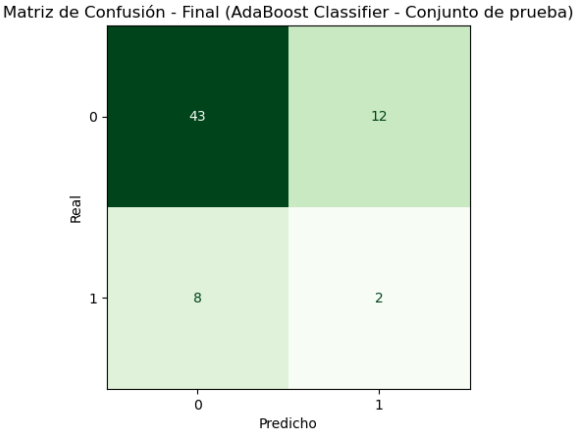

Figura D.348: AdaBoost, Final, C4, SMOTE

Gradient Boosting

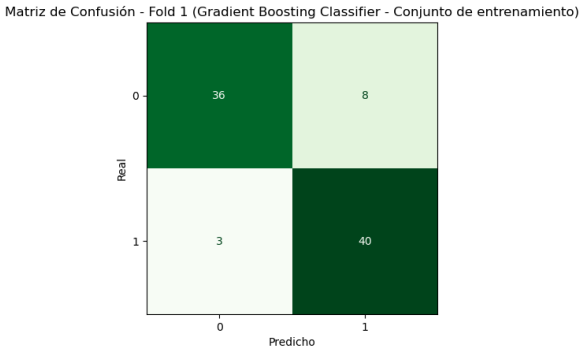

Figura D.349: Gradient Boosting, Fold 1, C4, SMOTE

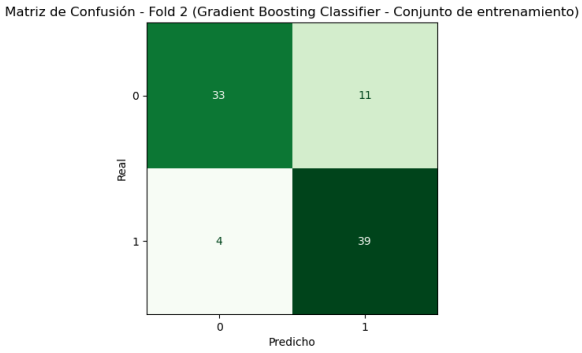

Figura D.350: Gradient Boosting, Fold 2, C4, SMOTE

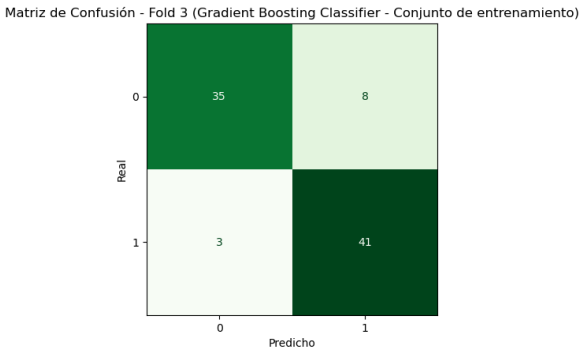

Figura D.350: Gradient Boosting, Fold 3, C4, SMOTE

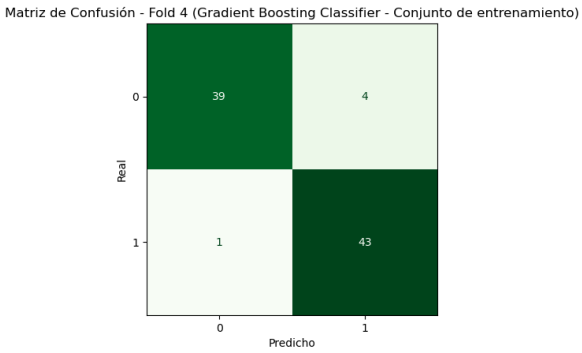

Figura D.351: Gradient Boosting, Fold 4, C4, SMOTE

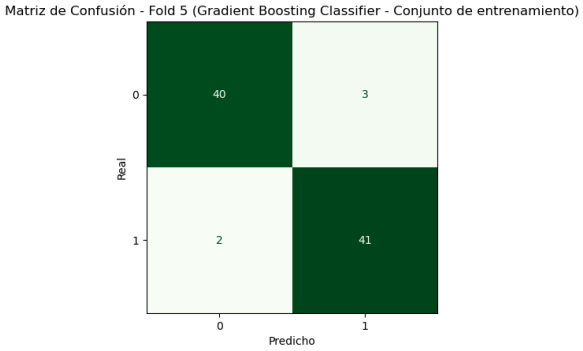

Figura D.351: Gradient Boosting, Fold 5, C4, SMOTE

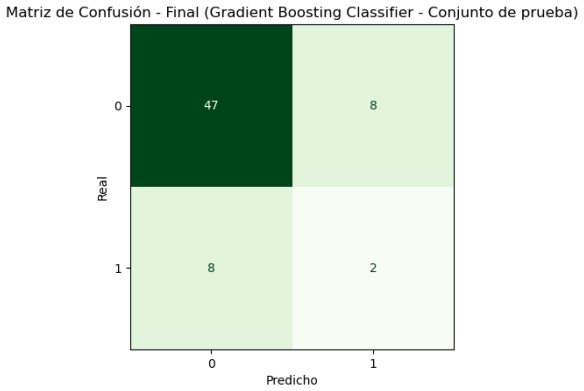

Figura D.352: Gradient Boosting, Final, C4, SMOTE

SVC

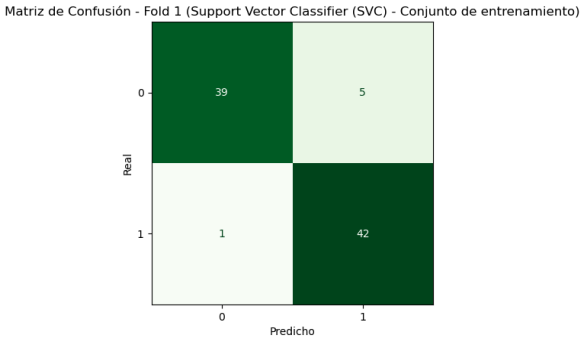

Figura D.353: SVC, Fold 1, C4, SMOTE

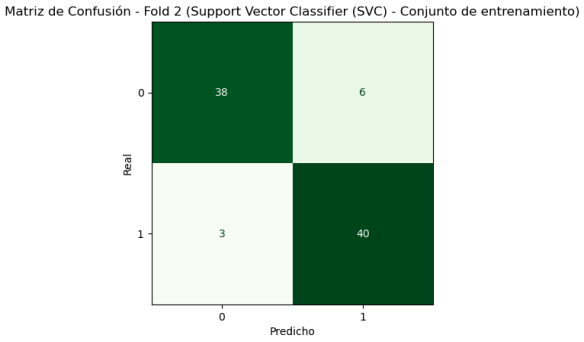

Figura D.354: SVC, Fold 2, C4, SMOTE

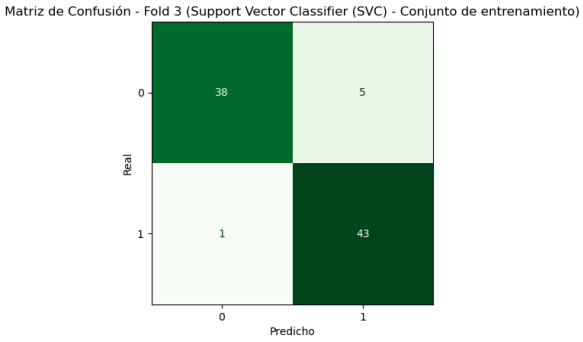

Figura D.354: SVC, Fold 3, C4, SMOTE

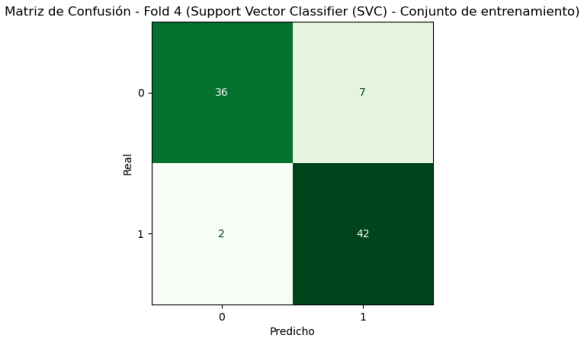

Figura D.355: SVC, Fold 4, C4, SMOTE

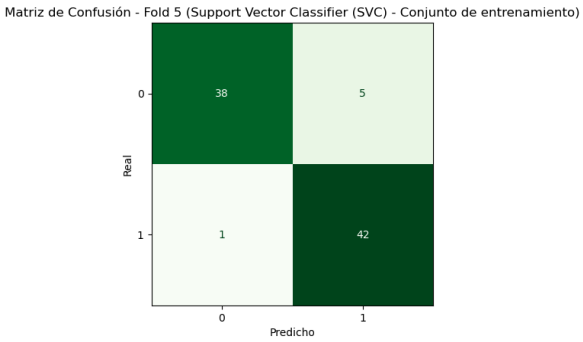

Figura D.355: SVC, Fold 5, C4, SMOTE

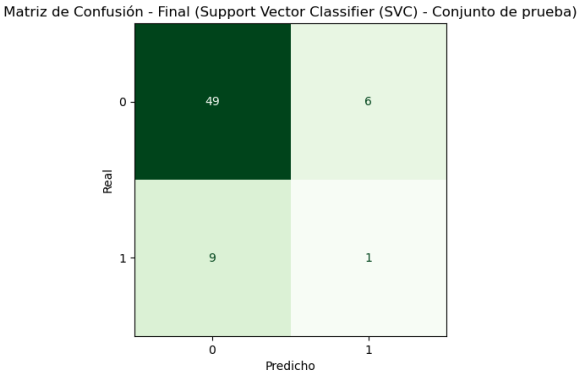

Figura D.356: SVC, Final, C4, SMOTE

k-Nearest Neighbors

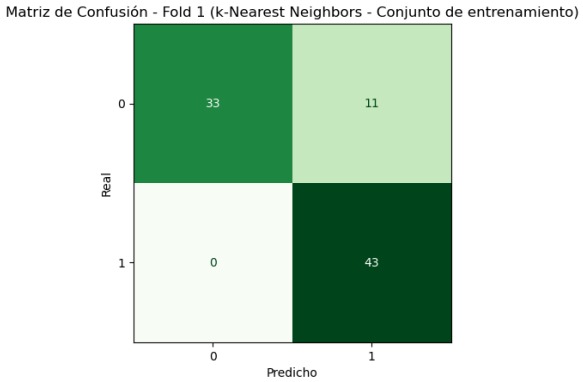

Figura D.357: k-Nearest Neighbors, Fold 1, C4, SMOTE

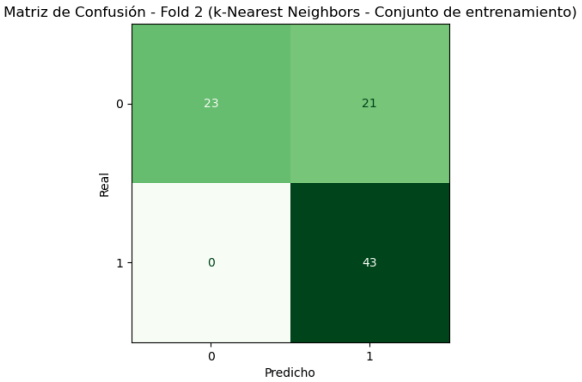

Figura D.358: k-Nearest Neighbors, Fold 2, C4, SMOTE

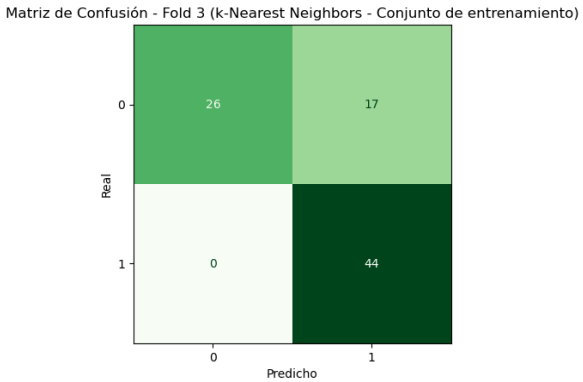

Figura D.358: k-Nearest Neighbors, Fold 3, C4, SMOTE

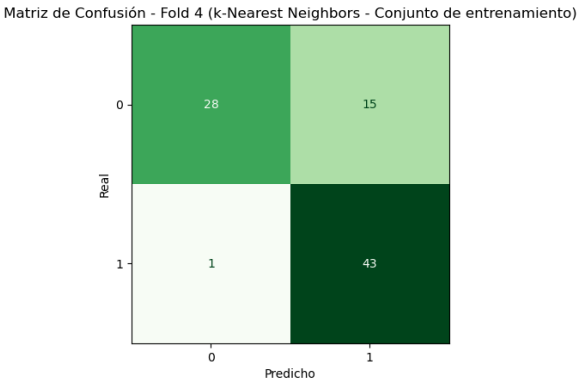

Figura D.359: k-Nearest Neighbors, Fold 4, C4, SMOTE

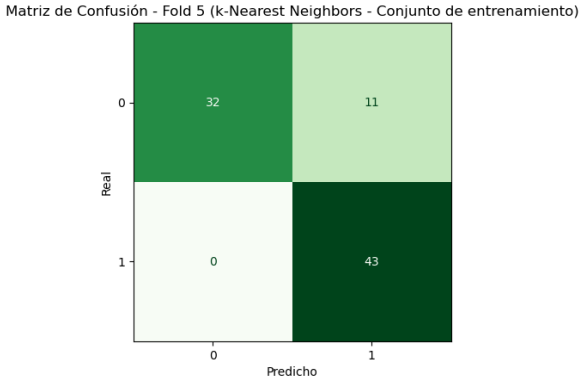

Figura D.359: k-Nearest Neighbors, Fold 5, C4, SMOTE

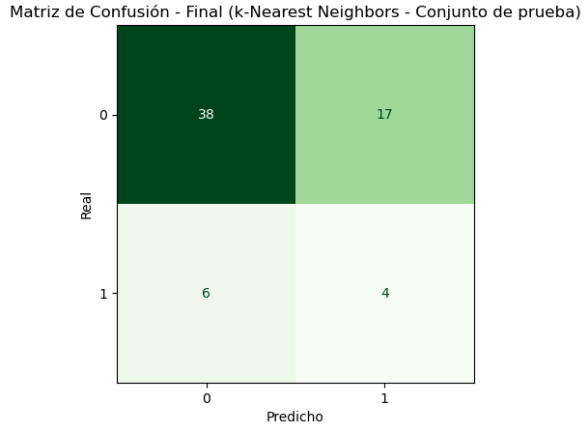

Figura D.360: k-Nearest Neighbors, Final, C4, SMOTE

Gaussian Naive Bayes

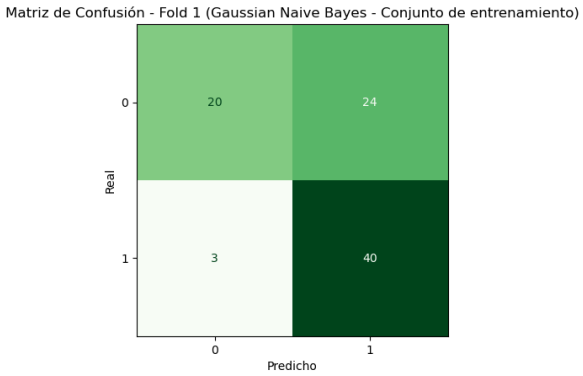

Figura D.361: Gaussian Naive Bayes, Fold 1, C4, SMOTE

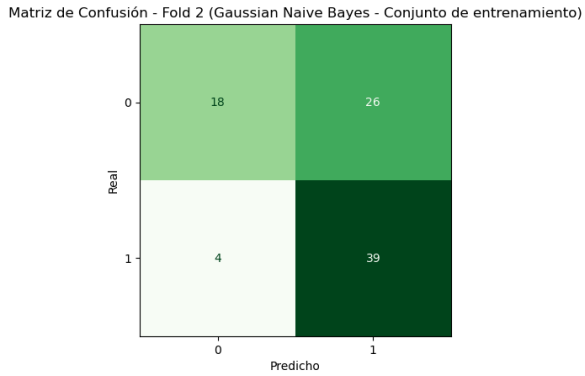

Figura D.362: Gaussian Naive Bayes, Fold 2, C4, SMOTE

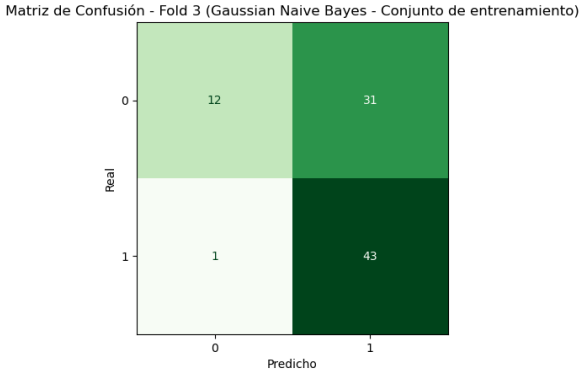

Figura D.362: Gaussian Naive Bayes, Fold 3, C4, SMOTE

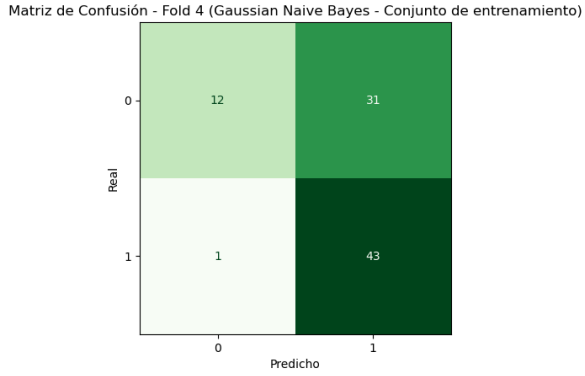

Figura D.363: Gaussian Naive Bayes, Fold 4, C4, SMOTE

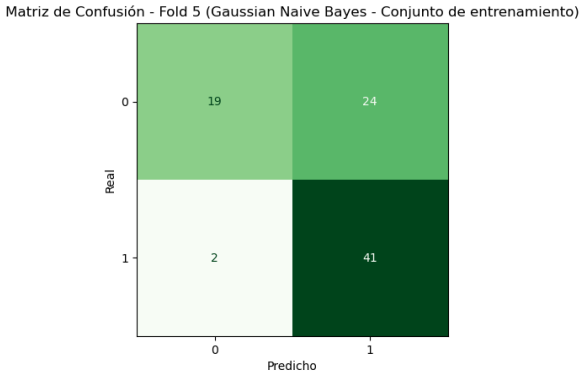

Figura D.363: Gaussian Naive Bayes, Fold 5, C4, SMOTE

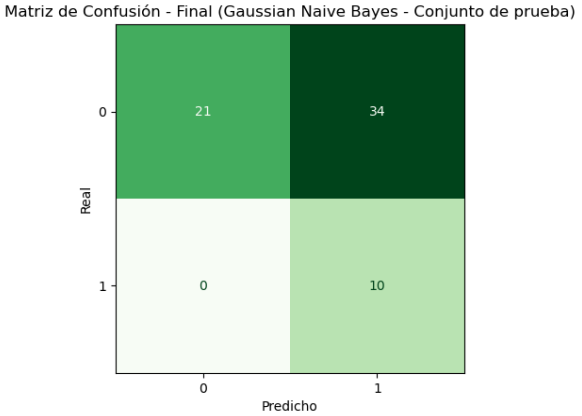

Figura D.364: Gaussian Naive Bayes, Final, C4, SMOTE

MLP

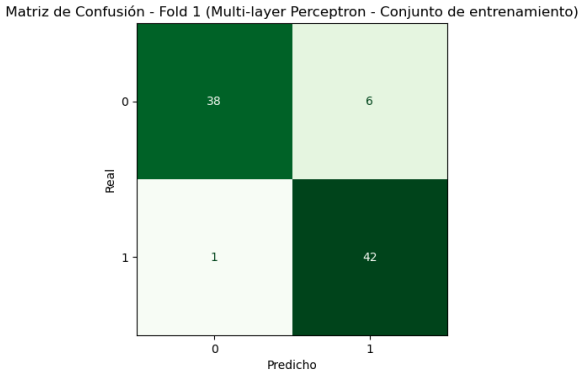

Figura D.365: MLP, Fold 1, C4, SMOTE

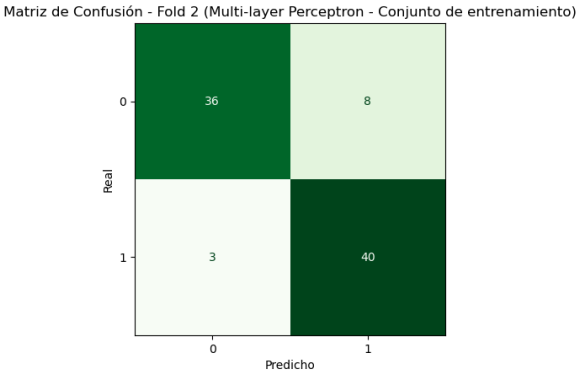

Figura D.366: MLP, Fold 2, C4, SMOTE

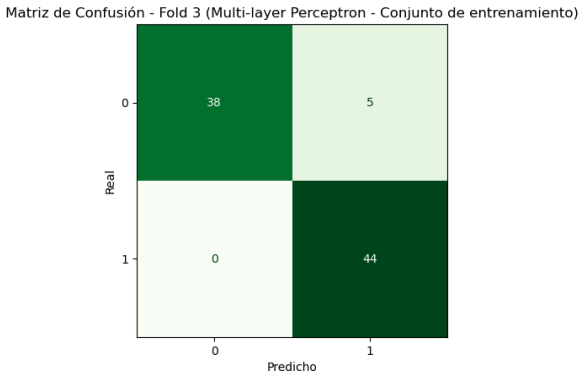

Figura D.366: MLP, Fold 3, C4, SMOTE

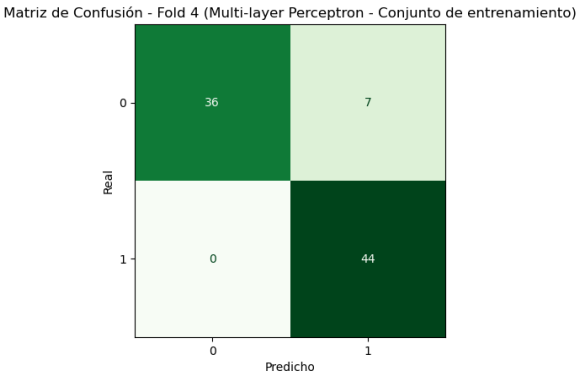

Figura D.367: MLP, Fold 4, C4, SMOTE

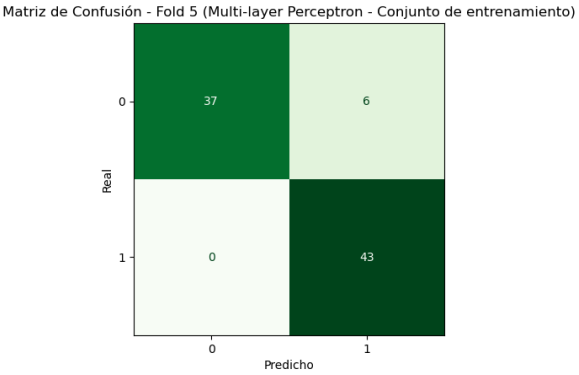

Figura D.367: MLP, Fold 5, C4, SMOTE

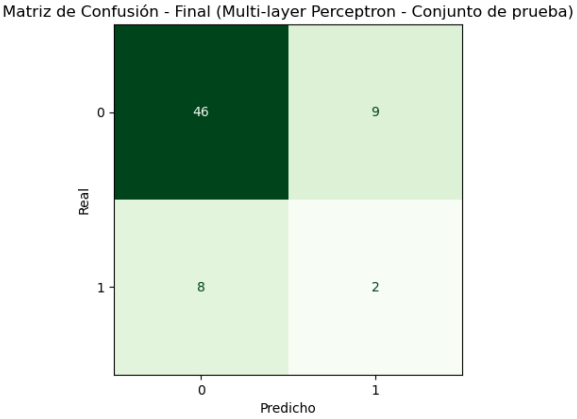

Figura D.368: MLP, Final, C4, SMOTE