UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE MASTER EN INGENIERÍA QUIMICA

ANÁLISIS VOLTAMÉTRICO DE ORINAS DE PACIENTES DE CÁNCER DE VEJIGA MEDIANTE UNA LENGUA ELECTRÓNICA. ELABORACIÓN DE UN MODELO DE DETECCIÓN

AUTOR: IGNACIO TALAVERA PRIETO

TUTOR: JOSÉ LUÍS VIVANCOS BONO

COTUTORA: Mª CARMEN MARTÍNEZ BISBAL

Curso Académico: 2016-17

RESUMEN

Este trabajo de final de máster recoge la elaboración de un modelo de predicción de cáncer de vejiga elaborado a partir de muestras de orina de pacientes antes y después de padecer la enfermedad. Las muestras de orina son medidas mediante una lengua electrónica voltamétrica y los resultados traspasados a formato digital. Estas medidas son utilizadas en el software informático Matlab para entrenar sistemas de análisis multivariante que permitan predecir si muestras de futuros pacientes padecen de cáncer de vejiga o no. Los sistemas entrenados fueron sistemas de análisis discriminantes, tanto lineal como cuadrático, además de un análisis de componentes principales. Tras las pruebas realizadas se descubre en este trabajo el mejor número de electrodos para trabajar con la lengua electrónica y el método de análisis que mejor se ajusta a estos datos para la predicción. En último lugar se concluyen cuales son las mejores combinaciones de electrodos para los sistemas entrenados y se encuentran dos sistema capaces de predecir con una precisión aproximada del 75%. Con estos sistemas se predicen muestras de pacientes un tiempo después de su recuperación de la enfermedad.

Palabras clave: Cáncer de vejiga, lengua electrónica voltamétrica, Matlab, análisis multivariante, modelos de predicción, análisis discriminante, análisis de componentes principales, muestras de orina.

RESUM

Aquest treball de final de màster arreplega la elaboració d'un model de predicció de càncer de bufera elaborat a partir de mostres d'orina de pacients abans i després de patir la malaltia. Les mostres d'orina son mesurades per mitjà d'una llengua electrónica voltamétrica i els resultats son utilizades en el software informàtic Matlab per a entrenar sistemes d'anàlisis multivariant que permet predir si mostres de futurs pacients patixen càncer de bufeta o no. Els sistemes entrenatsvan ser sistemes d'anàlisis discriminants, tant lineal com cuadràtic, a mès d'un anàlisis de components principals. Tras les probes realitzades es descubreix en aquest treball el millor número d'electrodes per a treballar amb la llengua electrónica i el mètode d'anàlisis que millor s'ajusta a aquest dades para la predicció. En ultim lloc es conciouen quais son les millors combinacions d'electrodes per als sistemas entrenats i es troben dos sistemes capaços de predir amb una precissió aproximada del 75%. Amb aquest sistemes es prediuer mostres de pacients un temps després de la seua recuperació de la malaltia.

Paraules clau: càncer de bufeta, llengua electrónica voltamétrica, Matlab, anàlisis multivariant, model de predicciò, anàlisis discriminant, anàlisis de components principals, mostres d'orina.

ABSTRACT

This work includes the development of abladder cancer prediction model made from patients urine samples before and after suffering the disease. Urine samples are measured using a voltammetric electronic tongue and the results are transferred to digital format. These measures are applied with Matlab computer software for train multivariate analysis systems to predict whether samples of future patients are suffering of cancer. The trained systems were discriminant analysis, both linear and quadratic, in addition to a principal component analysis. After the tests were performed, the best number of electrodes to work with the electronic tongue and the method of analysis that get best fits for these data to make the prediction is discovered in this work. Finally, the best combinations of electrodes for the trained systems are concluded, and were found two systems to predict with an approximate accuracy of 75%. These systems were used to predict patient samples a time after they recovery from the disease.

Keywords: Bladder cancer, electronic voltammetric tongue, Matlab, multivariate analysis, predictive models, discriminant analysis, principal component analysis, urine samples.

ÍNDICE

Tabla de contenido

ÍNDICE DE LA MEMORIA

1. INTRODUCCIÓN

1.1 OBJETIVO

El objetivo de este trabajo final de máster es la elaboración de un modelo de predicción de cáncer de vejiga a partir de muestras de orina de pacientes diagnosticados con este tipo de cáncer antes y después de su operación.

1.2 INTRODUCCIÓN DEL TRABAJO

Todo el mundo conoce el cáncer, ya sea por familiares, amigos o conocidos, y es bien sabido que uno de los factores clave a la hora de afrontar una recuperación es identificarlo antes de que se comiencen a producir los primeros síntomas, ya que este tiempo puede significar que se consiga una curación o no.

Actualmente la forma más común de detectar un cáncer de vejiga es mediante una cistoscopia, proceso que conlleva una intervención, por lo que la finalidad de este trabajo es encontrar un modelo de detección a partir de las muestras de orina de los pacientes.

Para la elaboración de este TFM se ha contado con la ayuda del Hospital Universitario y Politécnico de la Fe de Valencia, el cual ha proporcionado muestras de orina de pacientes enfermos de cáncer de vejiga antes y después de ser tratados. Estas muestras fueron medidas con una lengua electrónica voltamétrica para recabar los datos necesarios para su estudio. Con estos datos se ha trabajado, mediante el programa informático Matlab, para conseguir crear un modelo de predicción que pueda detectar con la mayor confianza posible muestras de pacientes posteriores, para determinar si padecen o no cáncer de vejiga sin la necesidad de recurrir a la cistoscopia.

2. ASPECTOS TEÓRICOS

2.1 INTRODUCCIÓN

La vejiga es un órgano hueco situado en la parte inferior del abdomen. Tiene forma de globo pequeño y una pared muscular que le permite agrandarse o encogerse para almacenar la orina que producen los riñones. Hay dos riñones, uno en cada lado de la columna vertebral, encima de la cintura. Túbulos diminutos en los riñones filtran y limpian la sangre. Estos vacían los productos de desecho y producen la orina. La orina discurre por cada riñón hacia la vejiga a través de un tubo largo que se llama uréter. La vejiga almacena la orina hasta que esta pasa a través de la uretra y al exterior del cuerpo. En la figura 1 se representa un diagrama de la situación de la vejiga en el cuerpo humano.

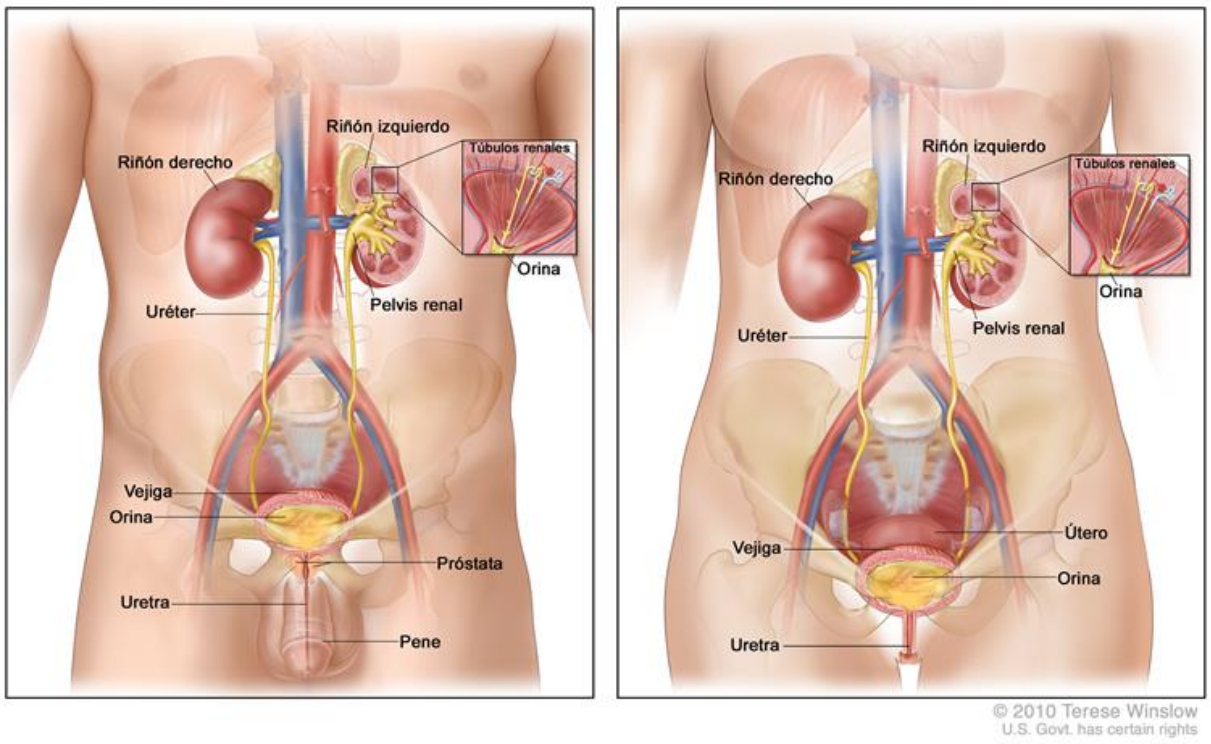

Figura 1: Anatomía del sistema urinario masculino y femenino [1]

Existen tres tipos de cáncer de vejiga, que comienzan en las células que la revisten. A estos cánceres se les da el nombre según el tipo de células que se vuelven malignas (cancerosas):

- Carcinoma de células de transición: cáncer que comienza en las células de la capa de tejido más interna de la vejiga. Estas células se pueden estirar cuando la vejiga está llena y encogerse cuando se vacía. La mayor parte de los cánceres de vejiga comienzan en las células de transición. El carcinoma de células de transición puede ser de grado bajo o de grado alto:

- El carcinoma de células de transición de grado bajo vuelve con frecuencia después del tratamiento, pero pocas veces se disemina a la capa muscular de la vejiga u otras partes del cuerpo
- El carcinoma de células de transición de grado alto, con frecuencia vuelve tras el tratamiento y con frecuencia a la capa muscular de la vejiga, hacia otras partes del cuerpo y a los ganglios linfáticos. Casi todas las muertes por cáncer de vejiga se deben a enfermedad de grado alto.

❖ Carcinoma de células escamosas: cáncer que comienza en las células escamosas, que son células delgadas y planas, que se pueden formar en la vejiga después de una infección o irritación prolongadas.

❖ Adenocarcinoma: cáncer que comienza en las células glandulares (de secreción) que están en el revestimiento de la vejiga. Este es un tipo poco común de cáncer de vejiga.

El cáncer que está en el revestimiento de la vejiga se llama cáncer de vejiga superficial. El cáncer que se disemina a través del revestimiento de la vejiga y que invade la pared muscular de la vejiga o que se disemina a órganos cercanos y ganglios linfáticos se llama cáncer de vejiga invasivo. [2]

2.2 DIAGNOSTICO

Cuando se sospecha que existe un cáncer de vejiga, la prueba de diagnóstico más útil es la cistoscopia. Los estudios radiológicos, como las exploraciones con tomografía computarizada (TC) o las ecografías, no tienen la sensibilidad suficiente para ayudar a detectar cánceres de vejiga. La cistoscopia se puede realizar en un consultorio de urología.

Si se observa un cáncer de grado alto en la cistoscopia, se suele citar al paciente para una exploración bimanual con anestesia y una repetición de la cistoscopia en un quirófano, de modo que se pueda realizar la biopsia o la resección transuretal del tumor. Si se observa un cáncer de grado alto o cáncer invasivo, el paciente se estadifica con una TC del abdomen y la pelvis, o bien una radiografía o una TC del tórax. Los pacientes con una elevación de la fosfatasa alcalina no hepática o síntomas que indican metástasis óseas se someten a una exploración ósea. [3]

2.3 DIAGNÓSTICO ALTERNATIVO: LENGUA ELECTRÓNICA

Los sistemas gustativos biológicos han servido de inspiración estos últimos años a la comunidad científica para crear dispositivos que puedan simular de manera similar el funcionamiento de estos sistemas. A estos dispositivos se les conoce por el nombre de lenguas electrónicas, y tienen un gran número de aplicaciones, cada vez más elevado, como son el sector alimenticio o el sector médico. En este caso, el más importante es el sector médico, ya que se utilizará una lengua electrónica para intentar predecir la existencia de cáncer de vejiga. La mayoría de los trabajos realizados hasta este momento con lenguas electrónicas están basados en sensores

electroquímicos. También pueden encontrarse trabajos sobre sistemas potenciométricos, los cuales utilizan electrodos de ión selectivo o electrodos de ión no selectivo construidos con diversos materiales. El problema de estos sistemas es que se encuentran limitados a la detección de especias químicas de carácter iónico, además son muy sensibles al ruido electrónico, por lo que requieren un alto coste en instrumentación y adquisición de datos. El tipo de trabajos que nos incumbe en este caso, son los relacionados con las lenguas electrónicas basadas en voltametría, las cuales trabajan con un sistema de electrodos elaborados a base de metales nobles como platino, oro, iridio, rodio y paladio. En contraposición a los sensores potenciométricos, los sensores voltamétricos no están limitados únicamente a la detección de especies cargadas, además poseen propiedades como robustez, versatilidad y sencillez, las cuales mejoran sus características. Sin embargo, también se ha de indicar que este sistema ofrece señales más débiles y con poca información debido a la naturaleza de los sensores, que no presentan procesos electroquímicos definidos. [4]

En las lenguas electrónicas el sistema de procesado de datos combina las señales recibidas de varios sensores de distinto tipo para determinar una propiedad de la muestra o cuantificar alguno de sus compuestos. Para cada aplicación se escogen aquellos sensores que proporcionen la respuesta combinada adecuada para clasificar las muestras. En los últimos años, en relación a las lenguas electrónicas basadas en voltametría, el autor con mayor repercusión ha sido Winquist, el cual detalla las lenguas voltamétricas y sus aplicaciones más importantes en los últimos años.

Las lenguas electrónicas pueden definirse como sistemas basados en múltiples sensores de baja selectividad o sensibilidad cruzada conjugados con herramientas para el reconocimiento de patrones y de análisis multivariante. Estos permiten la clasificación de muestras o la cuantificación de sus parámetros fisicoquímicos. [5].

El principal aspecto que diferencia las lenguas electrónicas de los métodos de análisis convencionales es el uso de sensores de baja sensibilidad y sensores de sensibilidad cruzada, por lo que no se encuentran diseñados para detectar una única especie en concreto. Este aspecto tiene que ver con una de las funciones principales de las lenguas electrónicas, como es la clasificación. En este caso se tomaron muestras de orina de diferentes pacientes, antes y después de superar la enfermedad, para poder clasificarlos después. Estas medidas serán luego tratadas mediante software de análisis multivariante para intentar llegar a una clasificación correcta que permita predecir muestras futuras.

Para la realización de este Trabajo Final de Máster se ha utilizado una lengua electrónica voltamétrica creada por la Universidad Politécnica de Valencia. Esta lengua puede diferenciarse de otras anteriores por su forma de trabajo, ya que en este caso se ha conseguido trabajar con hasta ocho electrodos de forma simultánea, cuando otras lenguas únicamente son capaces de trabajar con dos o tres electrodos. Estos electrodos son de sensibilidad cruzada y baja selectividad, lo que permiten diferenciar entre varias especies diferentes. El funcionamiento de esta lengua está basado, como bien se ha dicho, en voltametría, que consiste en la aplicación de un potencial eléctrico que produzca reacciones de oxidación y reducción debidas a la transferencia de iones desde el seno de la disolución hasta la superficie del electrodo. Esta respuesta electroquímica queda almacenada en un archivo como intensidad de corriente para su posterior procesado y análisis. La intensidad de corriente puede ser representada en voltagramas para conocer el comportamiento de la corriente eléctrica en los diferentes electrodos según el potencial aplicado para cada uno de ellos. Una vez los

datos han sido medidos y almacenados, el siguiente paso es el procesado de estos, mediante herramientas de análisis multivariante, con la finalidad de conseguir clasificar las muestras medidas y poder discernir entre dos o más tipos de estas.

La estructura de la lengua electrónica puede observarse en la figura 2. Está compuesta por tres elementos: los sensores, el sistema de medida y el sistema de procesado de datos. Por lo general, los sensores tienen una sensibilidad cruzada a distintos compuestos de la muestra y su número depende de la aplicación y de la técnica que se haya empleado. Por otro lado, el sistema de medida es el encargado de captar las señales que generan los sensores, digitalizarlas y transmitirlas al sistema de procesado de datos, donde son guardados. Este sistema de procesado de datos tendrá como misión analizar las medidas recibidas y clasificarlas correctamente, basándose en reconocimiento de patrones y herramientas de análisis multivariante. Para que el sistema consiga realizar una clasificación correcta, antes debe ser programado, mediante un proceso de aprendizaje con muestras patrón, de las cuales se conocen sus características y clasificación.[6] y [7].

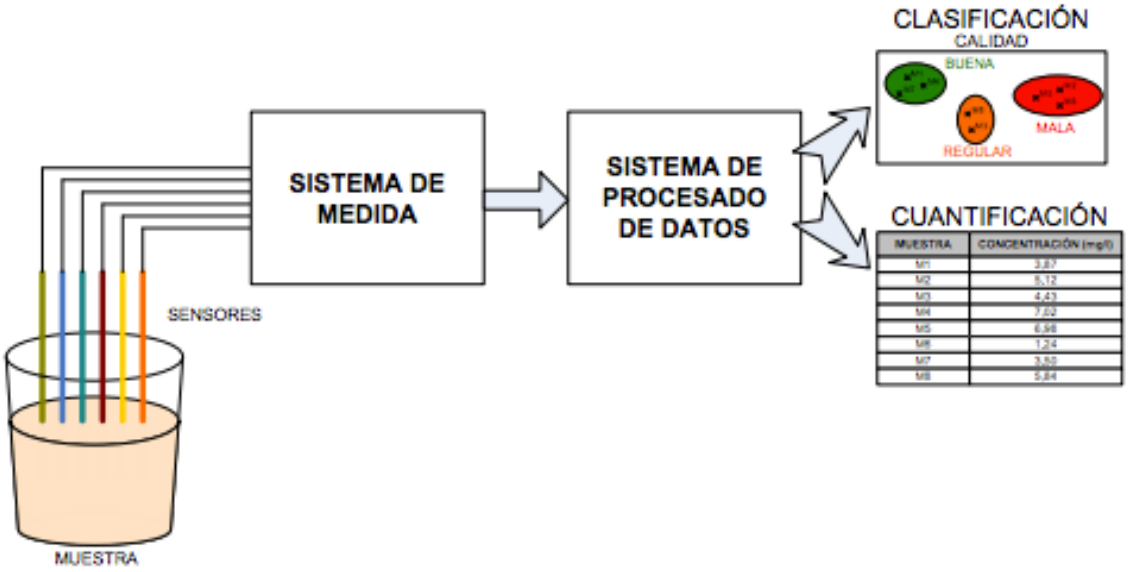

Figura 2: Elementos constituyentes de la lengua electrónica [7]

El equipo utilizado para la realización de este trabajo consta de los siguientes elementos:

- Software informático: Se trata de la aplicación que se utilizará para realizar los ensayos y almacenar los datos obtenidos.
- Equipo electrónico, el cual se compone, entre otras cosas, de un potenciostato que permite realizar al mismo tiempo medidas a ocho electrodos de trabajo.
- Sensores electroquímicos, los cuales están integrados por tres electrodos:
  - Electrodo de trabajo (WE): Este electrodo, también conocido como electrodo indicador, es en el que se producen los procesos electroquímicos de interés. El diseño de esta lengua se basa en un diseño creado por Winquist, que trata de emplear electrodos de trabajo de cuatro metales diferentes en una estructura cilíndrica de acero inoxidable. Los electrodos de trabajo están conformados por hilos de distintos materiales metálicos de 1,5 cm de longitud.

Estos cuatro hilos metálicos son encapsulados en la estructura cilíndrica a través de un epoxi. En su totalidad se compone de ocho metales agrupados en dos lenguas diferentes. La primera formada por metales nobles (Au, Pt, Rh, Ir) y la segunda por metales no nobles (Cu, Co, Ni, Ag). En este caso, la ordenación de los electrodos en la lengua electrónica es la siguiente:

- Electrodo 1: Electrodo de Iridio (Ir).
- Electrodo 2: Electrodo de Rodio (Rh)
- Electrodo 3: Electrodo de Platino (Pt).
- Electrodo 4: Electrodo de Oro (Au).
- Electrodo 5: Electrodo de Plata (Ag).
- Electrodo 6: Electrodo de Cobalto (Co)
- Electrodo 7: Electrodo de Cobre (Cu).
- Electrodo 8: Electrodo de Níquel (Ni).

- Electrodo de referencia o electrodo de calomelanos (RE): Su función es proporcionar un potencial estable al electrodo de trabajo. Está compuesto por un tubo interior que contiene una pasta con Hg2Cl2, KCl saturado y Hg. Este tubo a su vez se halla dentro de otro tubo en el que hay una disolución de KCl saturada. Existe un orificio que conecta ambos tubos. En el extremo del tubo exterior se encuentra el puente salino implementado mediante vidrio poroso o fibra de amianto. Actualmente, este electrodo está siendo reemplazado por el electrodo de referencia de plata y cloruro de plata (Ag/AgCl).

- Electrodo auxiliar o contraelectrodo (CE): Está formado por una estructura de acero inoxidable cilíndrica, la cual consta de 10 cm de longitud y 16 y 18 cm de diámetro interno y externo. El grado del acero inoxidable corresponde a un 316L. Su función en la lengua es conducir la electricidad desde la fuente de la señal hasta el electrodo de trabajo a través de la disolución. En la figura 3 se muestra una fotografía de la lengua electrónica utilizada, mientras que en la figura 4 puede encontrarse un diagrama explicativo de la situación de los electrodos.

Figura 3: Lengua electrónica voltamétrica

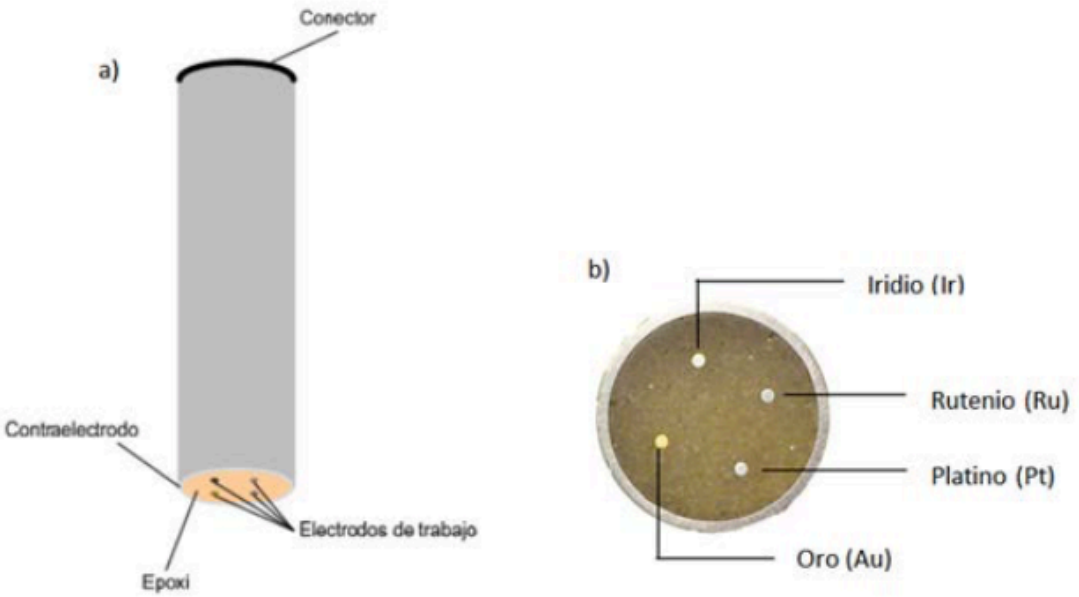

Figura 4: Diagrama electrodo

Tras realizarse las medidas, la lengua electrónica queda saturada de cargas iónicas, sobre todo la zona donde se encuentran los metales nobles, que por su alta reactividad se saturan con mayor facilidad. Esto hace que sea imposible realizar la siguiente medición sin antes proceder a la

limpieza de la lengua, ya que las señales eléctricas disminuyen y no se obtendrían buenos resultados. Para minimizar este efecto, la superficie de la lengua se lija con una lija fina en medio acuoso, para ser pulida posteriormente con alúmina fina, y obtener así una superficie homogénea. Si la superficie fuese irregular supondría una variación de la superficie de exposición, modificando la respuesta voltamétrica e induciendo a errores, por eso es importante un buen acabado. [6] [7].

2.4 ANÁLISIS DE RESULTADOS

Como se ha comentado anteriormente, este trabajo se realiza para la obtención de un modelo predictivo de cáncer de vejiga, consiguiéndose este modelo a raíz de muestras de pacientes enfermos y sanos. Estas muestras son analizadas mediante la lengua electrónica voltamétrica. A partir de estas muestras medidas, se trata de conseguir un modelo estadístico basado en técnicas de análisis multivariante, que permita estimar con un alto porcentaje de acierto las muestras de pacientes venideros. El problema de estas muestras viene en el sistema de medida, ya que al estar basado en técnicas voltamétricas genera una gran cantidad de información por cada medida. Exactamente para cada muestra de paciente se generan 7728 datos , siendo el número de muestras inicial cercano a 200. Esto supone una matriz de datos voluminosa, que hace necesaria la utilización de herramientas de análisis estadístico potentes, por lo que en este caso se ha decidido utilizar el software informático Matlab.

Los métodos de análisis multivariante utilizan técnicas de extracción para los principales parámetros de los datos medidos. Con esto se busca reducir las dimensiones de los datos. Para conseguirlo, se establecen combinaciones lineales de las variables originales, generándose nuevas variables ortogonales e independientes entre sí, que integran la máxima variabilidad de datos. En función del método utilizado, puede distinguirse entre variables principales, funciones discriminantes o variables latentes.

Las técnicas de análisis multivariante fueron creadas como herramienta para procesar la cantidad de información captada por los sistemas de medida electrónicos. El objetivo de estos es extraer la información más relevante de todo el conjunto de datos, minimizando o eliminando los valores que sean menos significativos o despreciables, para obtener un modelo lo más acertado posible. Tratan de reducir el tamaño de la muestra medida, quedándose únicamente con los valores más importantes para la clasificación.

Los principales métodos que han sido utilizados para la realización de este trabajo son:

- Análisis de componentes principales (PCA; principal component analysis).
- Análisis por mínimos cuadrados parciales (PLS; partial least squares).
- Análisis discriminante (DA: discriminant analysis).
- Análisis discriminante por mínimos cuadrados parciales (PLS-DA).

2.4.1 ANÁLISIS DE COMPONENTES PRINCIPALES (PCA)

El análisis de componentes principales es una técnica estadística empleada para disminuir las dimensiones de un conjunto de datos. Un PCA busca la proyección de los datos según la cual estos queden mejor representados en términos de mínimos cuadrados. Así se consigue convertir un conjunto de observaciones de variables posiblemente correlacionadas en un conjunto de valores de

variables sin correlación lineal llamadas componentes principales. De esta manera se resume la información repetida y se simplifica la gráfica de datos.

El análisis de componentes principales se basa en la conversión de los ejes de coordenadas que describen el sistema en un nuevo sistema en el cual los nuevos ejes son llamados componentes principales son ortogonales, ordenados de manera que el primer componente explique la mayor varianza posible, el segundo componente sea el siguiente que más explique, y así sucesivamente.

Los resultados de un análisis de componentes se expresan como los parámetros de observación (scores) y los pesos de las variables (loads). Para conseguir una correcta interpretación de los resultados, estos se representan en dos dimensiones, pero dependiendo de la aplicación, pueden llegar a representarse tridimensionalmente. En el gráfico de scores (figura 5) pueden observarse las diferentes agrupaciones, que son denominadas "clusters" (clases), de modo que los puntos que se encuentren próximos en el plano tienen características comunes, mientras que si están alejados no las tienen. Una representación gráfica simple podría propinar un error, por lo que se debe asumir qué porcentaje de la información total contiene cada uno de los parámetros. De este modo, el componente principal uno corresponde al porcentaje más alto de varianza, por lo que las variaciones dentro del componente principal uno tienen mayor relevancia que las variaciones en los demás componentes posteriores, perdiéndose importancia sucesivamente en cada componente. [6].

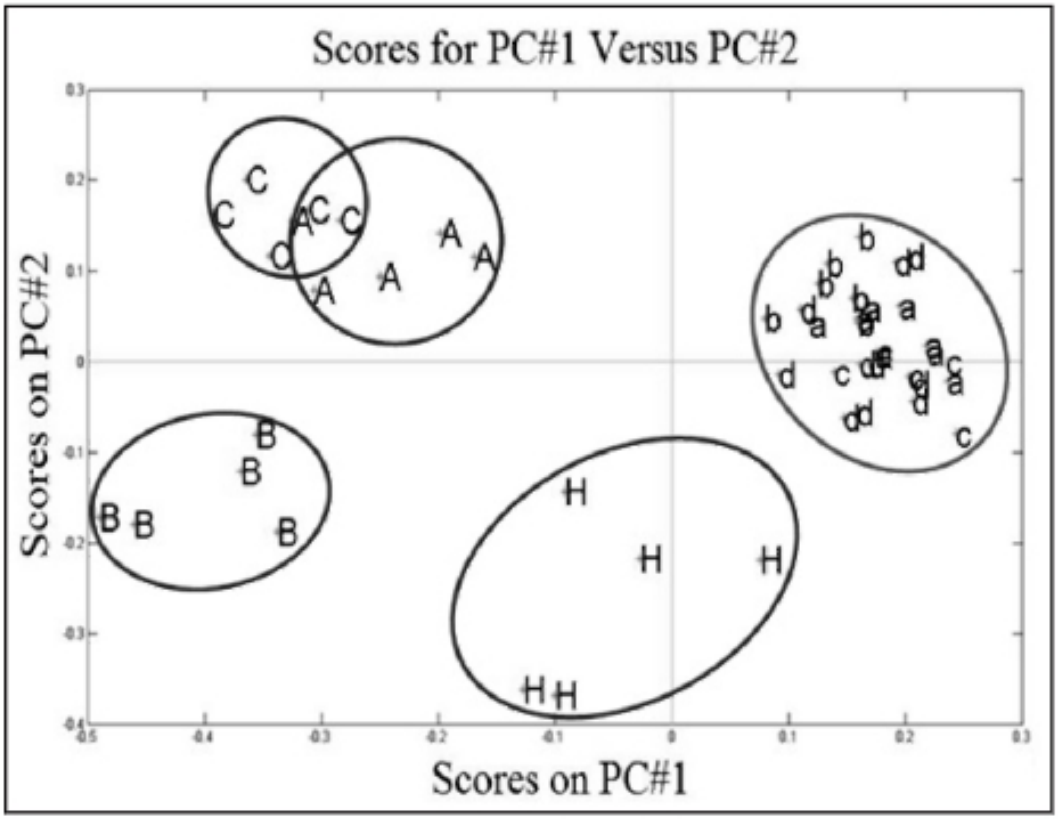

Figura 5: Gráfica de observaciones para PCA [11]

2.4.2 REGRESIÓN POR MÍNIMOS CUADRÁDOS PARCIALES (PLS)

La regresión por mínimos cuadrados es una técnica que generaliza y combina las características del análisis de componentes principales y la regresión múltiple. Su objetivo es principal es conseguir predecir o analizar un conjunto de variables dependientes a partir de un conjunto de variables independientes u observables. La predicción se consigue extrayendo de las variables observadas un conjunto de factores ortogonales denominados como variables latentes, que representan el mejor poder predictivo sobre las variables dependientes. Este tipo de análisis es realmente útil cuando se necesita predecir un conjunto de variables dependientes a partir de un conjunto de variables independientes. La principal diferencia en relación a los métodos de regresión de componentes principales (PCR) es que procura que los componentes principales contengan la mayor información para la matriz de predicción "Y". Para lograrlo, durante la etapa de calibración, el algoritmo PLS utiliza la información que contiene la matriz a determinar (Y), obteniéndose variables latentes que actúan como coeficientes. Paso previo a realizar la descomposición factorial, las matrices "X" e "Y" se auto escalan o centran al igual que para el análisis PCA. Cada una de estas matrices se descompone simultáneamente en una suma factorial. La descomposición de estas no es independiente, sino que se realiza simultáneamente, estableciéndose una relación interna entre los scores de los bloques de ambas matrices. [6].

El análisis de regresión por mínimos cuadrados parciales es una técnica comúnmente utilizada en los estudios basados en lengua electrónica, con la finalidad de obtener modelos de calibración y predicción a partir de una matriz generada por las variables independientes (respuestas eléctricas) y las variables dependientes (concentraciones), previamente comprimidas, y que estos modelos de predicción sean capaces de predecir muestras nuevas. Para este trabajo se utiliza un modelo de regresión tanto lineal como cuadrático.

2.5 MATLAB

Todos los cálculos realizados en este trabajo se han llevado a cabo a través de Matlab, por lo que es necesario explicar qué es y cual es su finalidad en este estudio. Matlab es la abreviatura de "Matrix Laboratory" (laboratorio de matrices), y se trata de una herramienta de software matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M). Entre sus prestaciones básicas se hallan la manipulación de matrices, la representación de datos y funciones, la implementación de algoritmos, la creación de interfaces de usuario y la comunicación con programas en otros lenguajes y con otros dispositivos hardware. El paquete Matlab dispone de diferentes aplicaciones, las cuales vienen ya programadas y facilitan la realización de numerosos procedimientos. En este estudio se ha utilizado la aplicación llamada "Classification Learner". [9]

2.5.1 CLASSIFICATION LEARNER

La aplicación de Matlab Classification Learner permite entrenar modelos para la clasificación de datos utilizando una maquina de aprendizaje supervisado (supervised machine learning). Por aprendizaje supervisado o aprendizaje automático se entiende el subcampo de las ciencias de computación e inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a los ordenadores aprender. De una forma más sencilla, se trata de generar programas capaces de

generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Por lo tanto, puede calificarse como un proceso de inducción del conocimiento.

Utilizando el Classification Learner pueden realizarse tareas comunes de aprendizaje con el ordenador, como la exploración de datos, selección de funciones, especificar esquemas de validación de datos, entrenar modelos y asesorar resultados. Puede decidirse entre varios tipos de modelos de clasificación, incluyéndose árboles de decisión, máquina de vectores de apoyo (SVM), o k-nearest neighbors (vecinos más cercanos), y seleccionar métodos de conjunto como "bagging, boosting y random subspace". La aplicación ayuda a escoger el mejor modelo para los datos introducidos permitiendo realizar la evaluación del modelo y las comparaciones del modelo utilizando las matrices de confusión y la curva ROC. Además te permite exportar los modelos de clasificación creados a la zona de trabajo de Matlab para poder realizar predicciones a nuevos datos.[9]

2.5.2 MATRIZ DE CONFUSIÓN

La matriz de confusión es una herramienta que permite la visualización de los resultados de un algoritmo que se emplea en aprendizaje supervisado. Se compone de una matriz "n x n" donde "n" es el número de clases. Cada columna de la matriz representa el número de predicciones de esa clase, mientras que cada fila representa a la clase real en sí. Su principal función es que puede detectarse si el sistema está confundiendo clases entre sí. Si en los datos de entrada el número de muestras de clases diferentes no varía mucho, la tasa de error del clasificador no puede ser representativa de lo bien o mal que está analizando el clasificador. Es decir, si el clasificador tiene una gran cantidad de muestras de clase 1 y muy pocas muestras de clase 2, este puede establecer un sesgo hacia la clase 1. Si esto ocurriese, su precisión según la matriz de confusión sería muy alta, pero esto no implicaría que fuese un buen clasificador, pues se obtendrá un 100% de error de clasificación en las muestras de segunda clase. Para visualizar esto mejor, en las siguientes figuras se expondrán un ejemplos de matriz de confusión normal y matriz de confusión afectada por este sesgo.

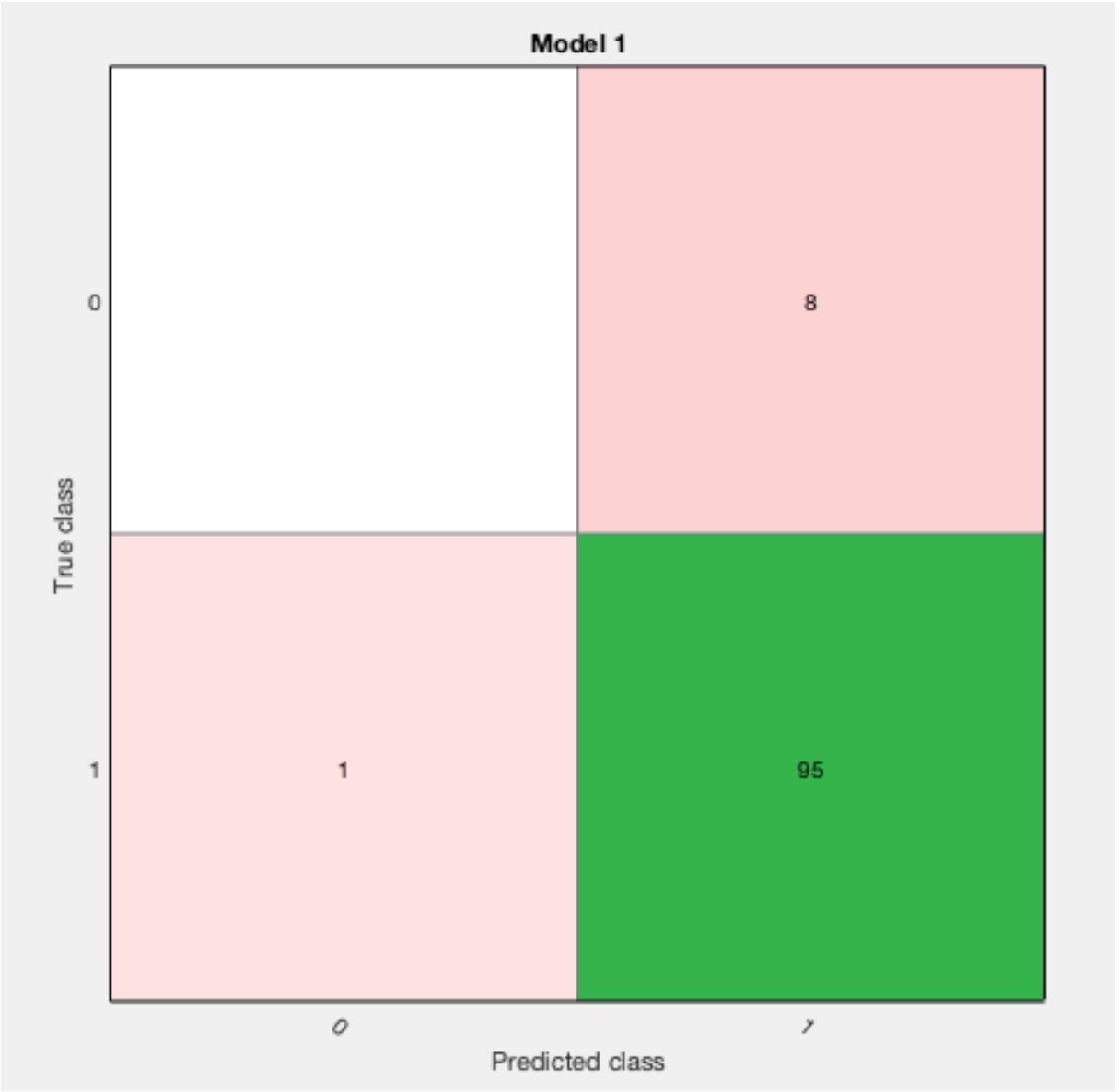

Figura 6: Ejemplo de matriz de confusión con sesgo

En esta matriz se observa como se ha aplicado el sesgo hacia la clase 1, ya que predice todas las muestras medidas menos una como de clase 1. El resultado podría parecer bueno si se atiende al número total de aciertos, ya que se obtiene un 90% aproximadamente de éxito. Para evitar esto, la matriz puede desglosarse en dos submatrices más:

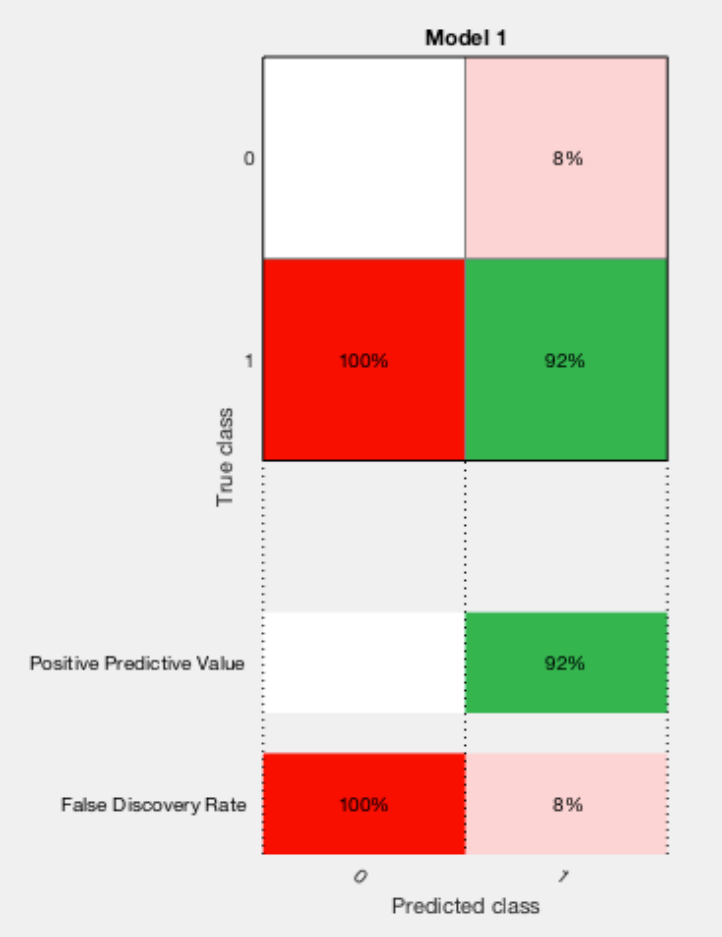

Figura 7: Matriz de confusión, ratios de predicción

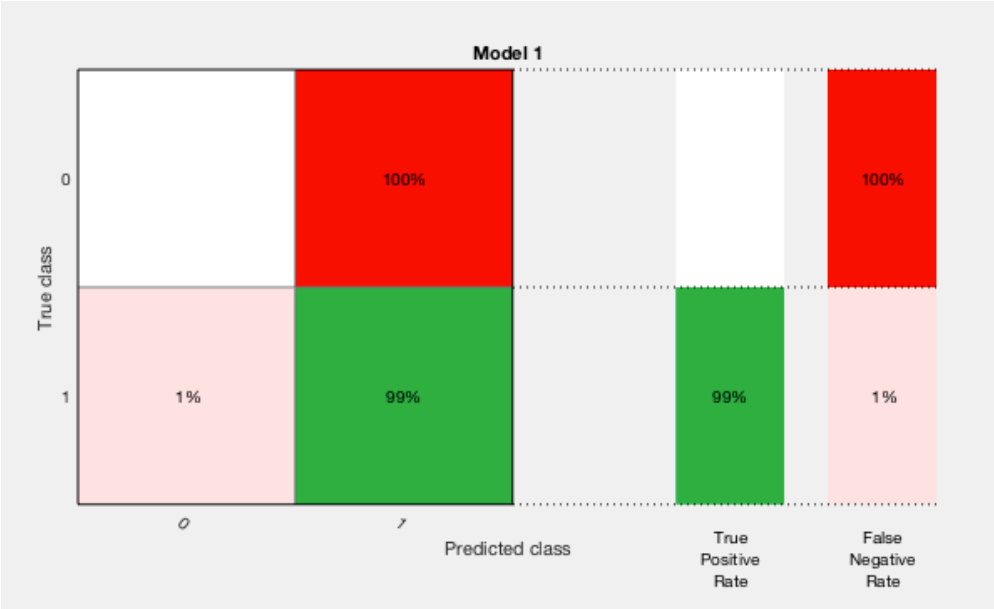

Figura 8: Matriz de confusión, ratios reales

Si se observan estas dos matrices, pueden comprobarse la tasa verdadera positiva, tasa de falsos negativo, tasa de descubrimiento falsa y valor predictivo positivo. Aquí ya puede observarse como para la clase 1 los ratios salen muy elevados, ya que se han predicho con acierto la mayoría, pero cuando se trata de la clase 2, los ratios positivos son bajísimos o ni siquiera existen, mientras que los ratios de falso positivo o negativos son muy elevados. Con esto se descubre que a pesar de que el programa en un principio determine que el clasificador tiene una confianza del 90%, a la hora de la verdad esto no es cierto.

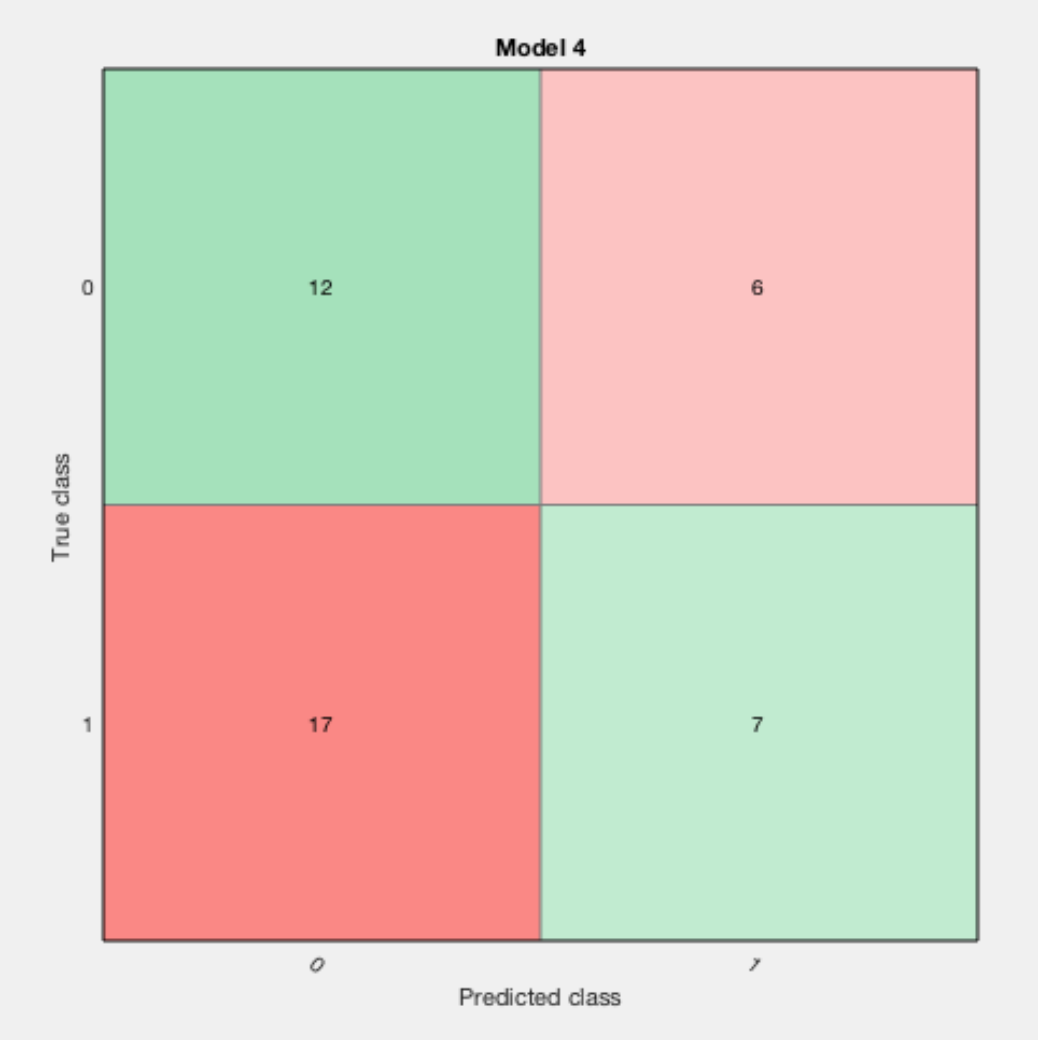

Figura 9: Ejemplo matriz de confusión

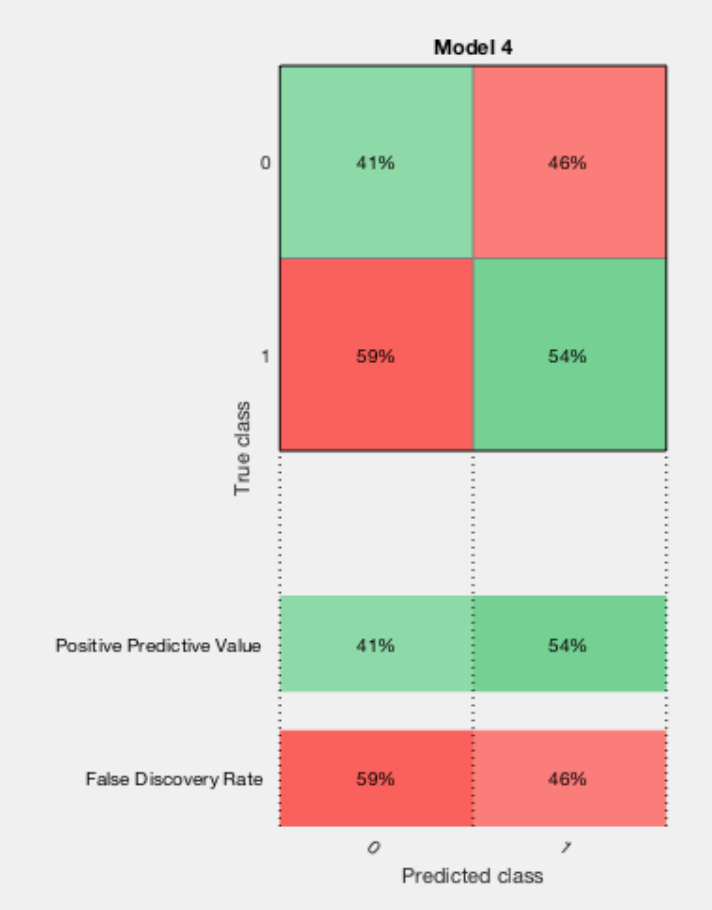

Figura 10: Ejemplo matriz de confusión, ratios de predicción sin sesgo

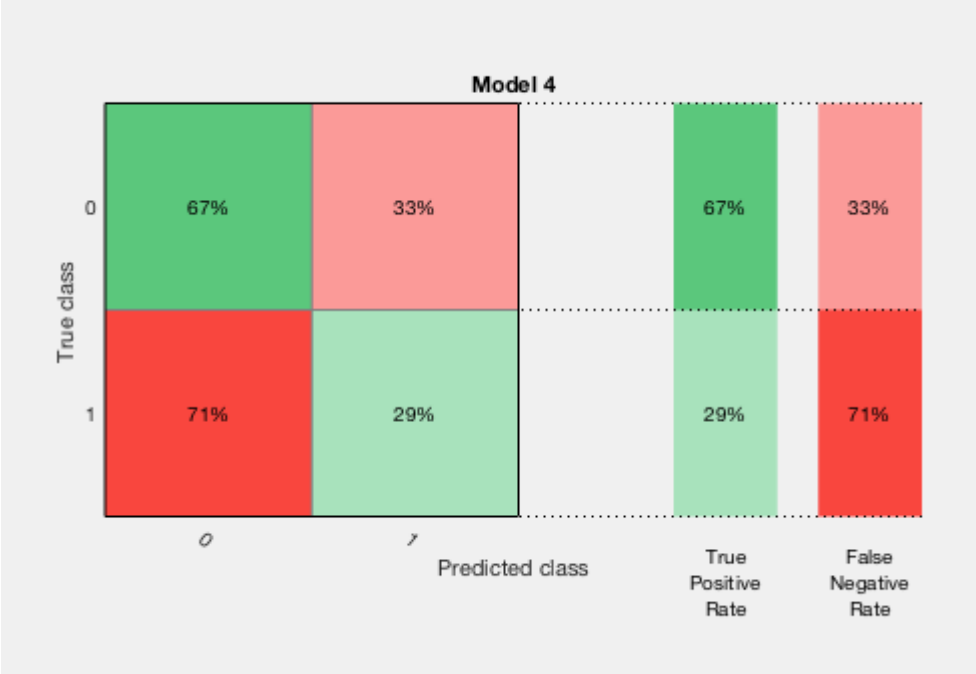

Figura 11: Ejemplo matriz de confusión, ratios reales sin sesgo

En estas matrices puede comprobarse como no se ha producido el sesgo hacia ninguna clase, ya que el clasificador predice tanto para la clase 1 como para la clase 2. También puede verse reflejado en los ratios comentados anteriormente, los cuales para este caso están equilibrados entre las clases. Para este ejemplo los resultados del clasificador no han sido buenos, y en los valores de los ratios puede observarse, ya que los valores positivos de predicción son muy bajos mientras que los ratios de error son altos.

2.5.3 CURVA ROC

Una curva ROC (Receiver Operating Characteristic) es una representación gráfica de la sensibilidad frente a la especificidad de un sistema clasificador según se varía el umbral de discriminación. Una interpretación diferente de este gráfico es la representación del ratio de verdaderos positivos (VPR) frente al ratio de falsos positivos (FPR) también según el umbral de discriminación, que es el valor a partir del cual se decide que un caso es un positivo. El término ROC también puede significar Relative Operating Characteristic porque se trata de una comparación entre dos características operativas (VPR y FPR) según se modifique el umbral de decisión.

El análisis de esta curva ROC proporciona una herramienta para seleccionar los modelos que puedan ser óptimos y descartar los modelos que no sean aceptables independientemente del coste de la distribución de las dos clases sobre las que se decide. La curva es independiente de la distribución de las clases en la población, se relaciona directa y naturalmente con el análisis de coste/beneficio en toma de decisiones diagnósticas.

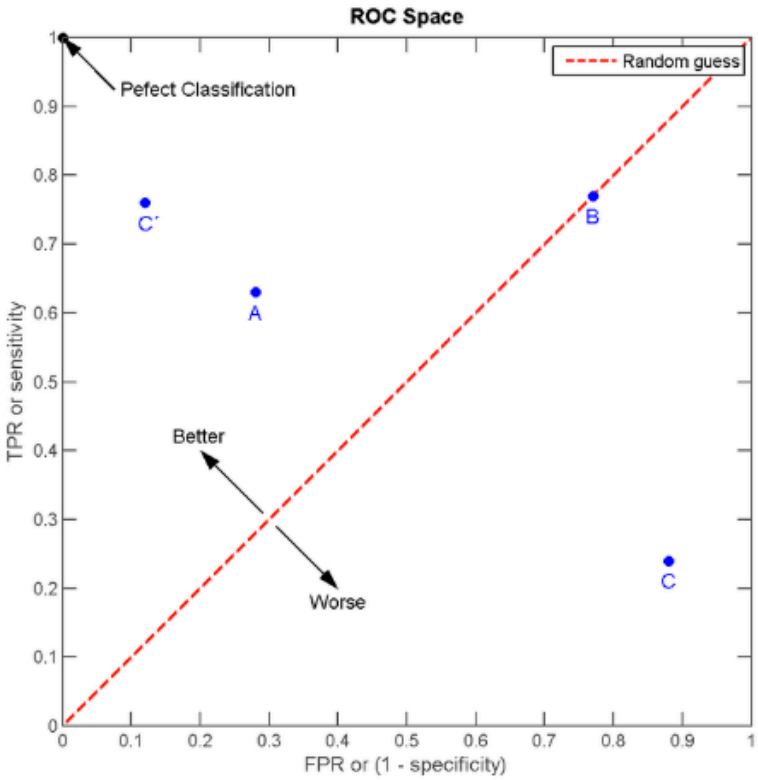

Figura 12: Ejemplo de curva ROC [10]

Un espacio ROC se define por FPR y VPR como ejes "x" e "y", y representa los intercambios entre verdaderos positivos y falsos positivos. Dado que estos términos son conocidos por sensibilidad y 1-especificidad, este gráfico también es conocido como la representación de la sensibilidad frente a (1-especifidad). Cada resultado de predicción de la matriz de confusión representa un punto en el espacio ROC.

El método ideal de predicción se situaría entonces en la esquina superior izquierda, representando un 100% de sensibilidad (ningún falso negativo) y 100% de especificad (ningún falso positivo). Este punto se le conoce también con el nombre de clasificación perfecta. Por otro lado, de una clasificación que fuese aleatoria se obtendría un punto situado a lo largo de la diagonal punteada, conocida como línea de no discriminación, desde el extremo inferior izquierdo hasta el extremo superior derecho.

La curva de cada sistema de clasificación dividirá el espacio ROC. Los puntos que queden por encima de la diagonal se consideran mejores que el azar, mientras que los resultados que queden por debajo serán considerados como peores que el azar.

Para poder interpretar las curvas ROC obtenidas en este trabajo, se utilizará lo que se conoce como el área bajo la curva ROC, comúnmente conocida también como AUC. Este índice puede interpretarse como la probabilidad de que un clasificador ordenará o puntuará una instancia positiva elegida aleatoriamente más alta que una negativa. Puede comprobarse que el área bajo la curva ROC es equivalente a la prueba de Mann-Whitney, que es otra prueba no paramétrica aplicada a dos muestras independientes, cuyos datos hayan sido medidos en una escala de nivel ordinal como mínimo. El índice AUC posee un valor comprendido entre 0,5 y la unidad, donde 1 representa la clasificación perfecta y 0,5 la imposibilidad de discriminación por parte del clasificador. Es decir, si para una AUC se obtiene un valor de 0,7 significaría que existe un 70% de probabilidad de que el resultado predicho sea más correcto que el predicho aleatoriamente. Por esto, cuanto más alto sea el valor mejor será la predicción.[10]

A modo de guía se podrían clasificar los valores de AUC de la siguiente manera:

- Entre 0,5 y 0,6: Clasificador malo
- Entre 0,6 y 0,75: Clasificador regular
- Entre 0,75 y 0,9: Clasificador bueno
- Entre 0,9 y 0,97: Clasificador muy bueno
- Entre 0,97 y 1: Clasificador excelente

2.5.4 TIPOS DE ANÁLISIS

Una vez conocidos los indicadores de la calidad del sistema de clasificación, se introducen brevemente los aspectos teóricos de los tipos de análisis utilizados a lo largo del trabajo para conseguir los modelos de clasificación. Los tipos de análisis que se han utilizado han sido análisis discriminante lineal y análisis discriminante cuadrático.

Un análisis discriminante es una técnica estadística multivariante que se utiliza con el fin de describir las diferencias significativas entre varios grupos o clases, sobre los que se observan un número de variables (variables discriminantes). En concreto, se describen y comparan las medidas de las variables clasificadoras a través de los grupos o clases.

Si se da la situación de que haya diferencias, se explicará en que sentido se dan y proporciona procedimientos de asignación sistemática de nuevas observaciones con un grupo o clase desconocida a una de las clases analizadas, usando para esto sus valores en las variables clasificadoras. Este método será el utilizado como modelo de predicción de clases para cada muestra a partir de las variables clasificatorias. Una vez conocido el fundamento de este tipo de análisis se utilizará tanto análisis discriminante lineal como cuadrático.

El análisis discriminante lineal es una generalización del discriminante lineal de Fisher, un método utilizado en estadística de reconocimiento de patrones y aprendizaje de máquinas para encontrar una combinación lineal de rasgos que caracterizan dos o más clases. El resultado obtenido de este análisis puede ser utilizado como un clasificador lineal. Este análisis esta basado en un análisis de la regresión y un análisis de la varianza, intentando expresar una variable dependiente como la combinación lineal de otras variables. Este tipo de análisis puede estar complementado con un análisis de componentes principales (PCA).

Para el caso del análisis discriminante cuadrático, al contrario que para el análisis lineal, no puede suponerse que los grupos tienen matrices de covarianza idénticas. En el análisis lineal, las observaciones serán clasificadas en el grupo que posea la distancia cuadrática más pequeña, pero en el caso de análisis cuadrático, no puede simplificarse a una función lineal. Esto es debido a que, para un análisis cuadrático, la distancia no es simétrica.

Como norma general, debería utilizarse un análisis discriminante lineal en el caso que las matrices de covarianza se presupongan idénticas para todos los grupos. Por otro lado, se debería utilizar un análisis discriminante cuadrático cuando las matrices de covarianzas no se presupongan similares para todos los grupos.

3. DATOS INICIALES

Para comenzar el trabajo, se partirá de 207 muestras de pacientes, recogidas tanto en el momento de la enfermedad, como justo después de la operación y un tiempo después a modo de control. Por esto, se pueden diferenciar dos o tres clases diferentes. Por un lado, pueden entenderse como clases enfermo de cáncer de vejiga (BC) o no enfermo de cáncer de vejiga. Estas dos clases pueden aumentarse a tres si dentro de la clase no enfermo se extraen las clases de post operatorio (CTRL) y la clase de control un tiempo después (CTRLpost).

Cada una de estas 207 muestras cuentan con 7728 variables cada una, por lo que se parte de una matriz de 207x7728. Con la finalidad de poder reconocer las muestras durante el trabajo, se inserta una columna más a la matriz. Esta columna contiene la numeración del análisis que se realizó RMN, por lo que cada muestra quedó desde este momento marcada con una numeración única que permite identificarla de las demás muestras. Todos los datos de cada muestra de cada paciente (género, clasificación, clase, etc.) se encuentran de igual modo marcados en una hoja Excel.

Previo al comienzo de este trabajo, estudios realizados a estas muestras por parte de la UPV y del Hospital Universitario y Policlínico de La Fe de Valencia encontraron que algunas muestras de las 207 no eran aptas para la clasificación, bien fuese por contaminación o por errores en la medida. Estas muestras fueron consideradas como outliers y eliminadas, siendo marcado en la hoja Excel las muestras que habían sido detectadas como outliers y algunas más que podrían serlo.

Una vez se comenzó la realización del trabajo se descubrió que las muestras detectadas como outliers y eliminadas de la matriz de datos de Matlab podría ser que no coincidiesen con las muestras detectadas y marcadas como outliers en la hoja Excel. Debido a esto, se decidió utilizar las 207 muestras iniciales en Matlab, y eliminar de esta matriz las muestras detectadas como outliers en la hoja Excel, con la finalidad de comparar la nueva matriz con la matriz de Matlab proporcionada por la universidad con los outliers ya eliminados. Tras realizar esta adecuación a la hoja Excel y comparar las dos matrices se comprueba que estas difieren en 3 muestras. Al no existir la posibilidad de cerciorarse si el error proviene de la hoja Excel o viene en la eliminación de muestras en la matriz de Matlab se decide trabajar con las dos matrices paralelamente, con el fin de discernir cual de las dos matrices se trata de la correcta, y si las diferencias entre las muestras eliminadas son significativas o realmente no influyen en los resultados y puede trabajarse con ambas considerando el error despreciable. Por esto se trabajará con dos variables de datos: "Xupv" y Xnew", que hacen referencia a la proveniencia de los datos, "Xupv" proviene de los datos iniciales de la universidad y el hospital y "Xnew" proviene de los datos al eliminar los outliers de acuerdo con la hoja de cálculo.

3.1 ADECUACIÓN DE LAS MUESTRAS A LA APLICACIÓN

Como ya se ha comentado anteriormente, para poder reconocer las muestras en todo momento y buscar la información necesaria sobre ellas en la hoja de cálculo, estas se numeran según la numeración RMN. Para numerarlas, se ha realizado una lectura de la numeración de la hoja mediante la función de Matlab "Xlsread" que permite adaptar a Matlab los datos de una hoja de cálculo Excel. Cuando los datos se encuentran ya importados en Matlab, se eliminan aquellos números que corresponden a las muestras eliminadas. Una vez el vector de la numeración y la matriz de datos tienen las dimensiones correctas, se unen mediante la función de Matlab "horzcat", poniendo en la primera columna la numeración y en las demás los valores medidos para cada muestra.

Este proceso se realiza tanto para las muestras de "Xupv" como para las muestras de "Xnew". De este modo, se obtuvieron dos matrices, de 193x7729 y de 194x7729 respectivamente. Al realizar las primeras pruebas a estas matrices de datos, se comprobó que todavía quedaban muestras que, anteriormente no habían sido reconocidas como outliers, pero realmente lo eran. En este proceso se eliminaron 24 muestras, las mismas para ambas matrices de datos. Estos nuevos cambios finalizaron con dos matrices de 169x7729 para "Xupv" y de 170x7729 para "Xnew".

Con todos los datos de las muestras preparados y los correspondientes outliers eliminados, el último paso antes de pasar a la aplicación de clasificación es la adición de la clase a cada muestra de la matriz. Para esto se importan las clases desde la hoja Excel con el comando "xlsread" y se eliminan los outliers al igual que en el caso de la numeración RMN y de las muestras de datos. Una vez se han adecuado las dimensiones de las clases se unen a la matriz de datos de dos formas diferentes:

1. Mediante la función "table" se unen las clases a la última columna de la matriz de datos en forma de texto, por lo que existirán tres clases a la hora de realizar el entrenamiento: "BC", "CTRL" y "CTRLpost".

2. Convirtiendo previamente las clases a modo lógico, mediante el comando "strcmpi", que convierte en lógico los valores de las clases, siendo 1 la clase que anteriormente era "BC" (enfermo) y como 0 las demás clases existentes (sanos), por lo que se reduce el número de clases a 2. De este modo solo se diferenciará entre pacientes sanos o enfermos.

De este modo, se han preparado 4 matrices de datos diferentes: "Xfinal_upv", que contiene las muestras iniciales del trabajo con los tres tipos de clases. "Xlogica_upv", que contiene las muestras iniciales del trabajo con los dos tipos de clase lógica. "Xnew_clases", que contiene las muestras modificadas del trabajo con los tres tipos de clases y "Xnew_claseslogica", que contiene las muestras modificadas del trabajo con los dos tipos de clases lógicas. Cada una de estas matrices será analizada mediante Classification Learner, la aplicación de análisis y entrenamiento de Matlab. Esta aplicación nos permitirá decidir que columnas incluir para el aprendizaje y cuales no incluir y, además, nos permite decir que columnas componen las clases, por lo que simplemente hay que indicar la matriz que se quiere analizar, eliminar del análisis la primera columna, la cual corresponde a la numeración RMN y no es de interés para la clasificación de las muestras y marcar como columna de clases la última columna de la matriz.

En esta primera prueba se realizará una separación de las muestras de la matriz entre calibración y validación mediante el método de "hold out validation". Esto significa que un porcentaje de las muestras de la matriz será utilizado para calibrar el modelo y el porcentaje restante será utilizado para la validación de este. Para cada matriz de datos se realizan dos pruebas, una al 25% y otra al 50%, esto quiere decir que en un caso se utilizarán el 25% de las muestras para validar y el 75% para calibrar mientras que en el caso de 50% se utilizarán la mitad de muestras para la calibración y la mitad restante para la validación.

Con todo esto seleccionado, el último paso es decidir cual será el tipo de análisis que se realizará. Se decide que los modelos que más pueden acercarse a la clasificación de estas muestras es el modelo de análisis discriminante. Por esto se decide realizar tanto un análisis discriminante lineal y un análisis discriminante cuadrático a cada matriz. Además, para cada uno de los tipos de análisis, se entrena un modelo aplicando también un PCA y sin su aplicación, para comprobar como afecta este a la clasificación y al entrenamiento del programa. Sabido todo esto, en las figuras 13 y 14 se muestra un esquema que simplifica las pruebas que se han realizado, a fin de exponerlo de una forma más visual y sencilla.

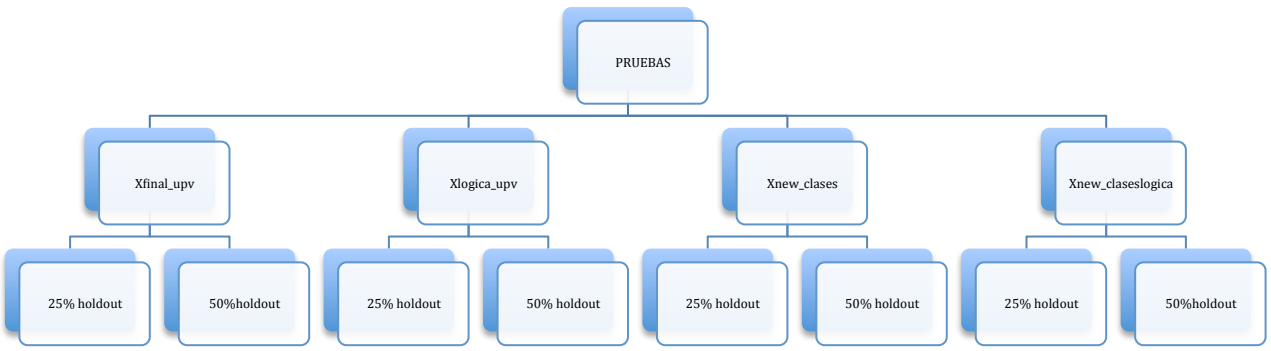

Figura 13: Diagrama de las pruebas realizadas (1)

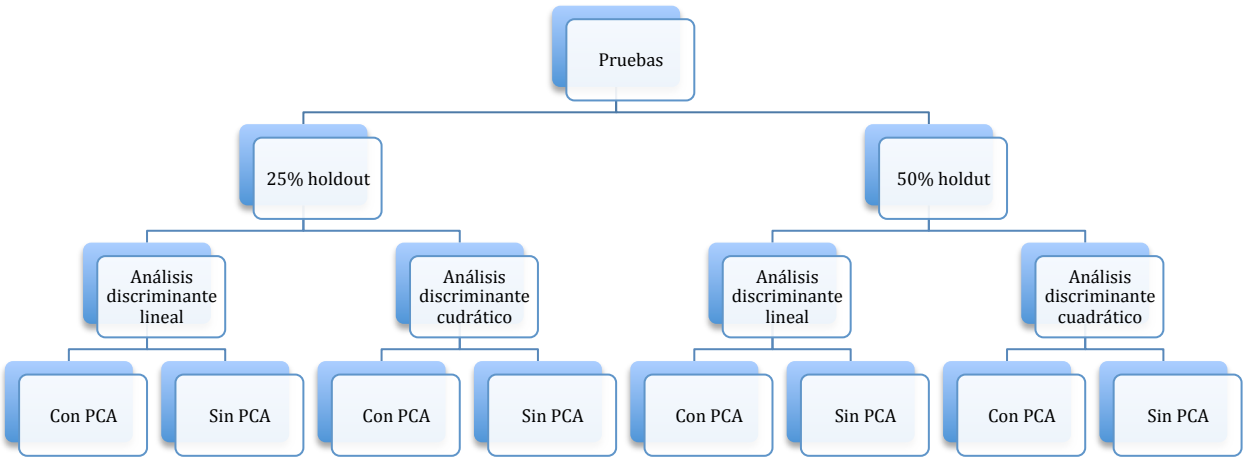

Figura 14: Diagrama de las pruebas realizadas (2)

4. Resultados

A continuación se muestran los resultados más importantes obtenidos en estos análisis. Debido a la gran cantidad de pruebas realizadas y resultados obtenidos solo se mostrarán los que se consideren mejores o más significativos en cada caso. Para cada matriz de muestras se escogen los dos resultados más complacientes, uno con un 25% de hold out y otro con un 50% de hold out. En el caso de que existan dos sistemas interesantes para un mismo porcentaje de calibración y validación se mostrarán los dos. Para cada sistema de clasificación se muestra la precisión en la predicción obtenida, además de las matrices de confianza y la curva ROC.

4.1 RESULTADOS DE LA MATRIZ INICIAL "Xfinal_upv"

4.1.1 25% HOLD OUT VALIDATION

Tabla 1: Resultados para "Xfinal_upv" con 25% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	43,2%
Discriminante lineal (Con PCA)	59,5%
Discriminante cuadrático (Sin PCA)	52,3%
Discriminante cuadrático (Con PCA)	49,7%

Para este primer caso se escogen los sistemas de análisis discriminante lineal con PCA y análisis discriminante cuadrático sin PCA, ya que son los que han obtenido una mejor precisión y es de interés analizar sus resultados.

Análisis discriminante lineal con PCA

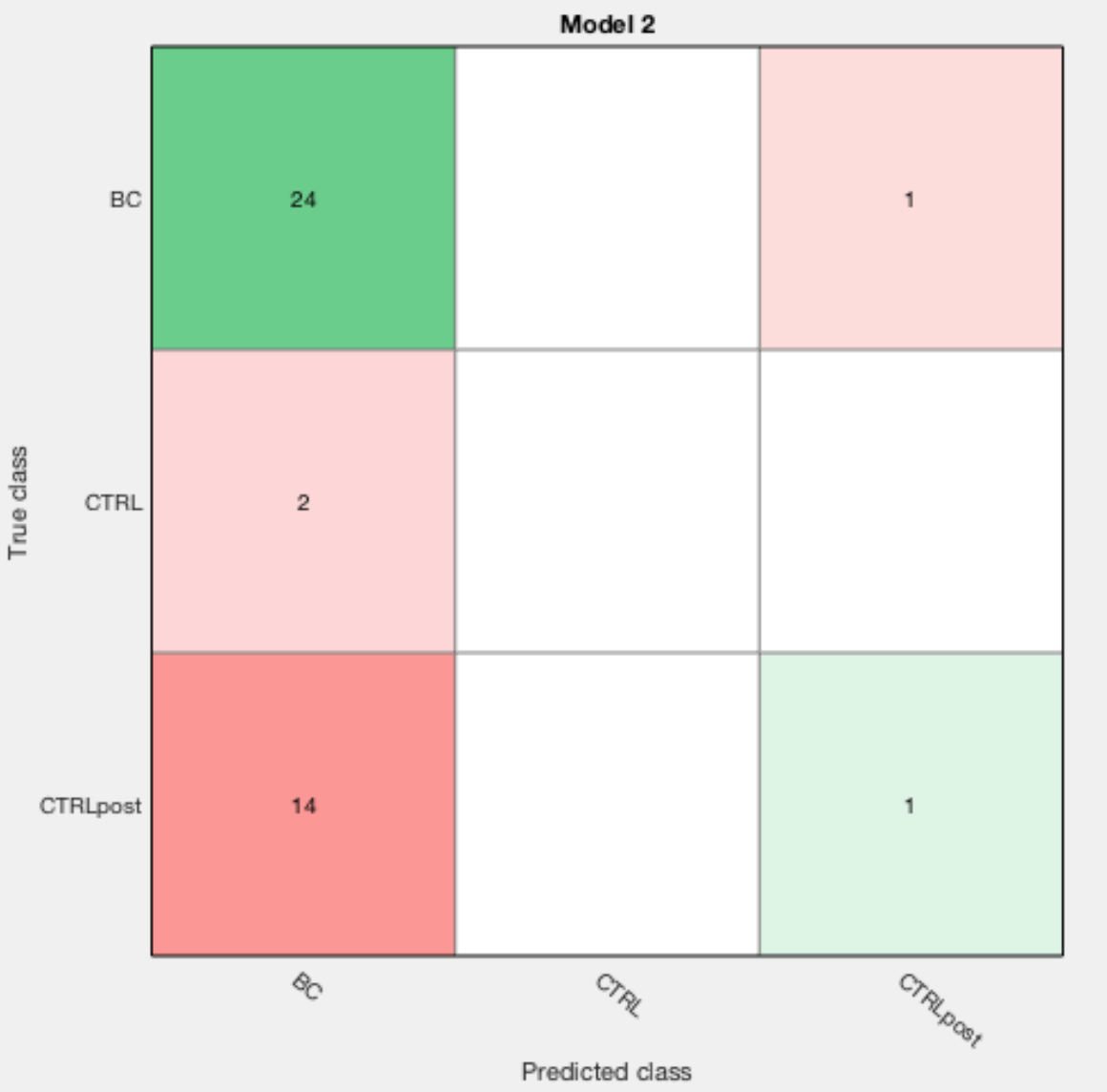

Figura 15 : Matriz de confusión para "Xupv_final" , análisis discriminante lineal con PCA (1)

Si se analiza esta matriz de confusión se comprueba como se ha producido el sesgo comentado en el apartado teórico. En este caso se ha producido hacia la clase "BC", para la cual predice todas las muestras menos 2. En ningún momento predice que ninguna muestra sea de clase "CTRL" y únicamente en dos ocasiones predice una muestra como clase "CTRLpost". Con estos resultados se confirma que el valor de la precisión obtenido es falso. Por lo tanto a priori este sistema no es válido para realizar predicciones.

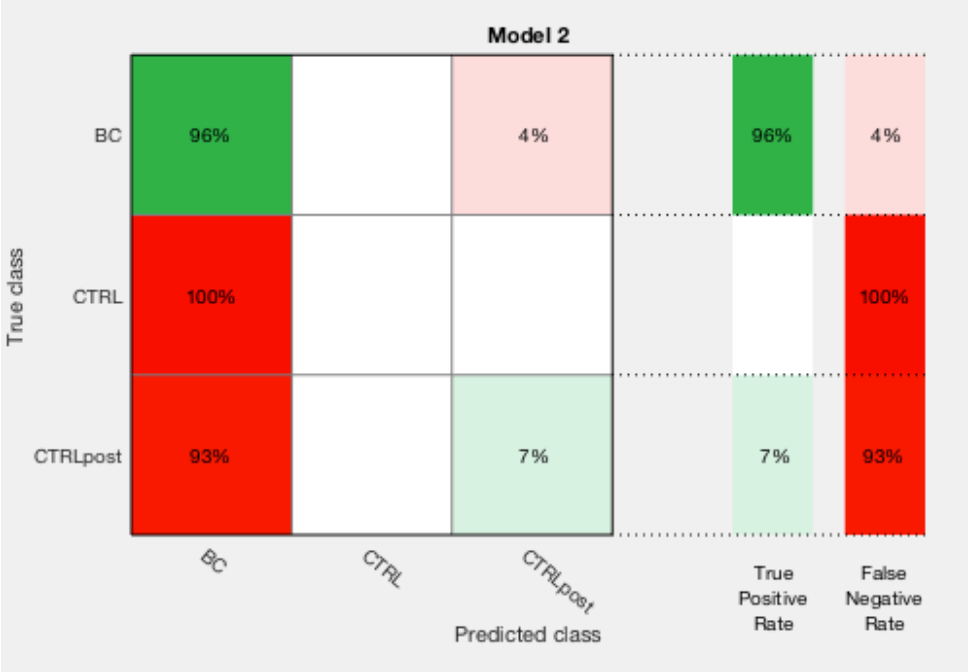

Figura 16: Matriz de confusión para "Xupv_final", análisis discriminante lineal con PCA (2)

Figura 17: Matriz de confusión para "Xupv_final", análisis discriminante lineal con PCA (3)

Si se atiende a estas dos matrices de confusión se comprueba que de las muestras reales de clase "BC" el sistema acierta en un 96% de las ocasiones, este resultado es engañoso, ya que al predecir casi la totalidad de muestras de esta clase, es normal que acierte. Este error se ve en la segunda matriz de confusión, donde se comprueba que de la totalidad de clases predichas como "BC" solo se está acertando el 60%. Este porcentaje dista mucho del 96% anterior y viene a demostrar que el sistema esta sesgado, por lo que no puede ser utilizado.

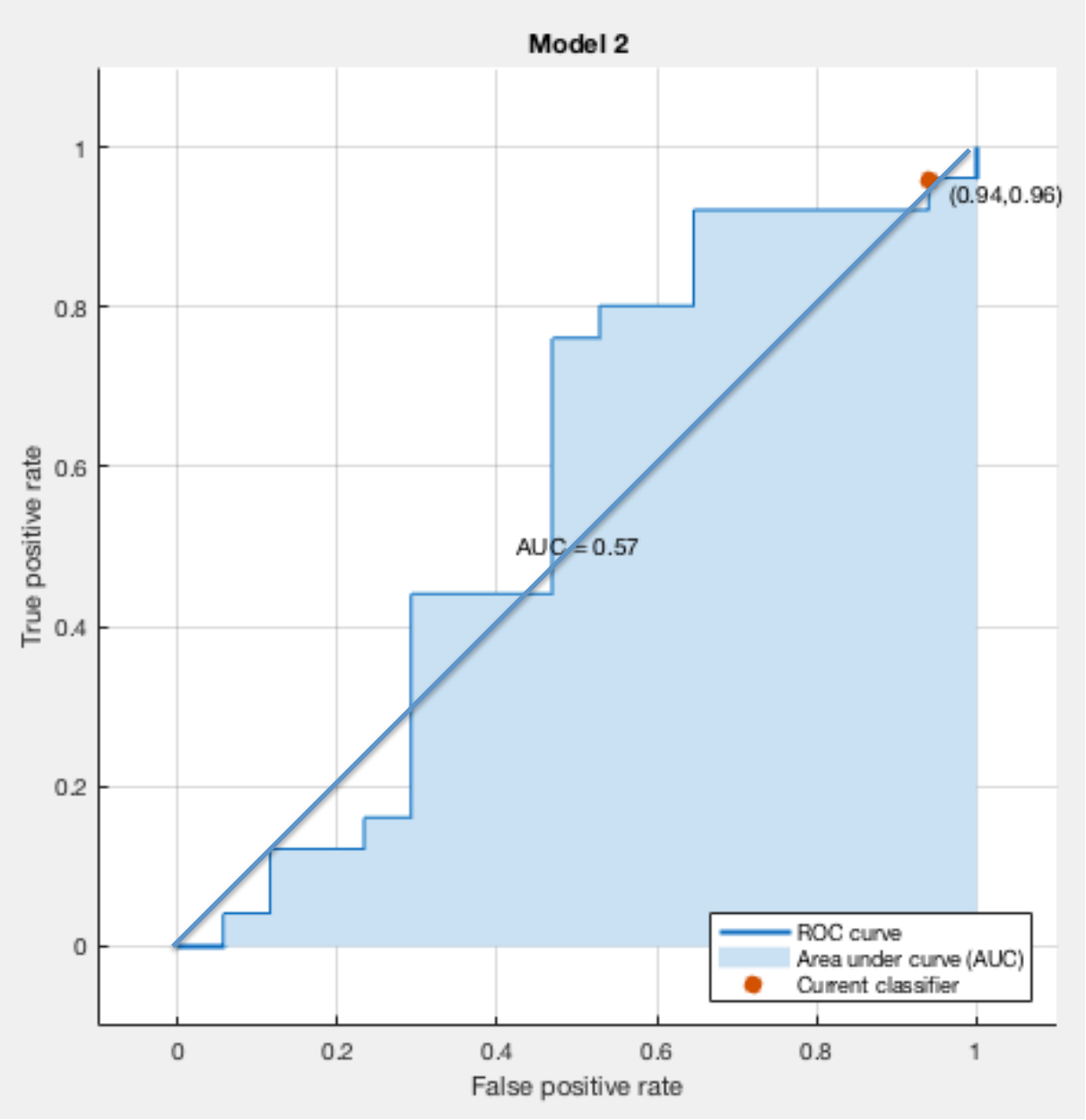

Figura 18: Curva ROC para "Xupv_final", análisis discriminante lineal con PCA

Si se atiende a los resultados obtenidos en la curva ROC para este sistema se comprueba como el valor del área bajo la curva (AUC) es de 0,57, por lo que solo predice un 7% mejor que el

azar. Si comparamos este valor con la tabla expuesta en el apartado de introducción teórica corresponde a un sistema de clasificación malo. Además. gráficamente puede observarse que la porción de área que supera la línea diagonal es muy pequeña, confirmándose que el sistema de clasificación debe ser descartado.

Análisis discriminante cuadrático sin PCA

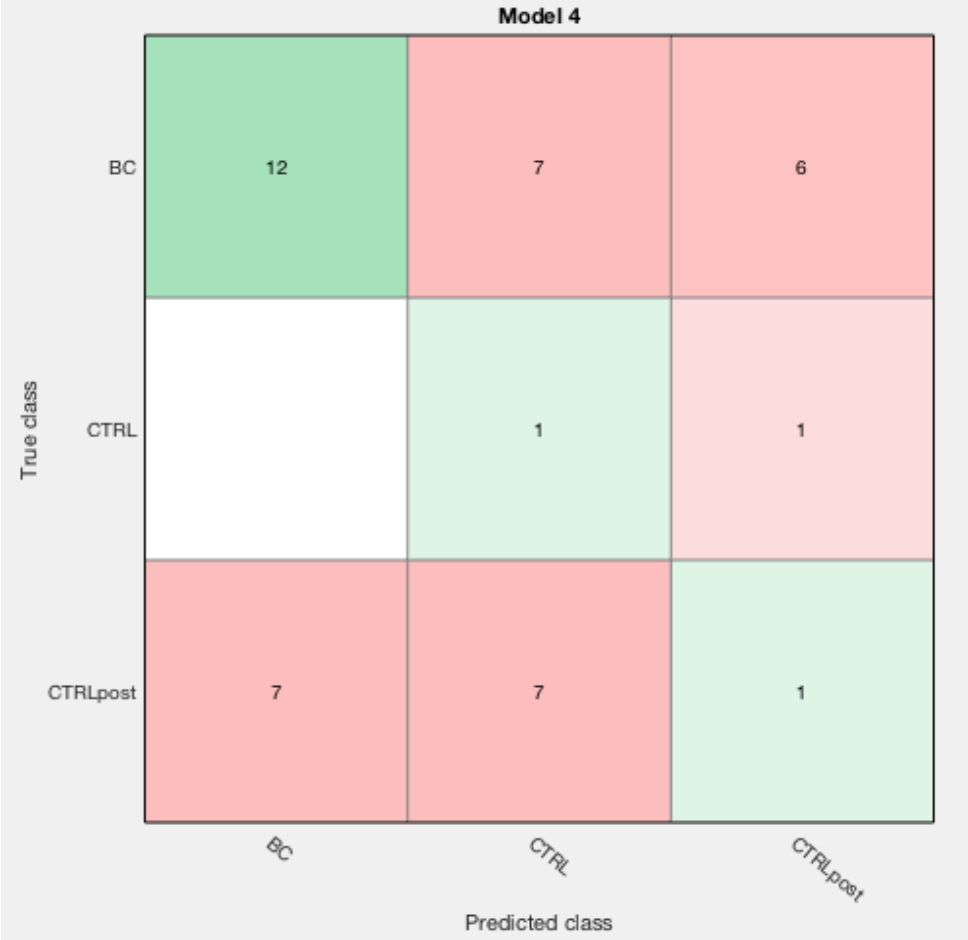

Figura 19: Matriz de confusión para "Xpuv_final" Análisis discriminante cuadrático sin PCA (1)

En este caso se observa como el sesgo anterior no se ha producido al no incluirse el análisis de componentes principales. Se han predicho aproximadamente el mismo número de clases para las muestras de pacientes sanos, mientras que se ha predicho un número mayor para pacientes con cáncer de vejiga. Aun con esto el valor de la precisión es bastante bajo todavía, con poco más de un 50% es prácticamente como dejar la decisión al azar, por lo que a falta de analizar los ratios y la curva ROC los resultados en principios no deberían ser aceptables.

Figura 20: Matriz de confusión para "Xupv_final", análisis discriminante cuadrático sin PCA (2)

En esta matriz se comprueba lo expuesto anteriormente, los resultados son bajos. Se puede extraer como de la totalidad de las muestras que realmente eran de enfermos de cáncer se acierta el 48%, que en este caso es menos de la mitad, pero el peor caso es para las muestras de control un tiempo después de la operación, ya que este sistema es incapaz de detectarlas, fallando el 93% de las ocasiones. Para las muestras de control tras la operación, el ratio de verdaderos positivos y de falsos negativos se encuentra en 50%, este valor indica lo mismo que los anteriores, este sistema de clasificación no es mejor que el azar, por lo que debe ser descartado.

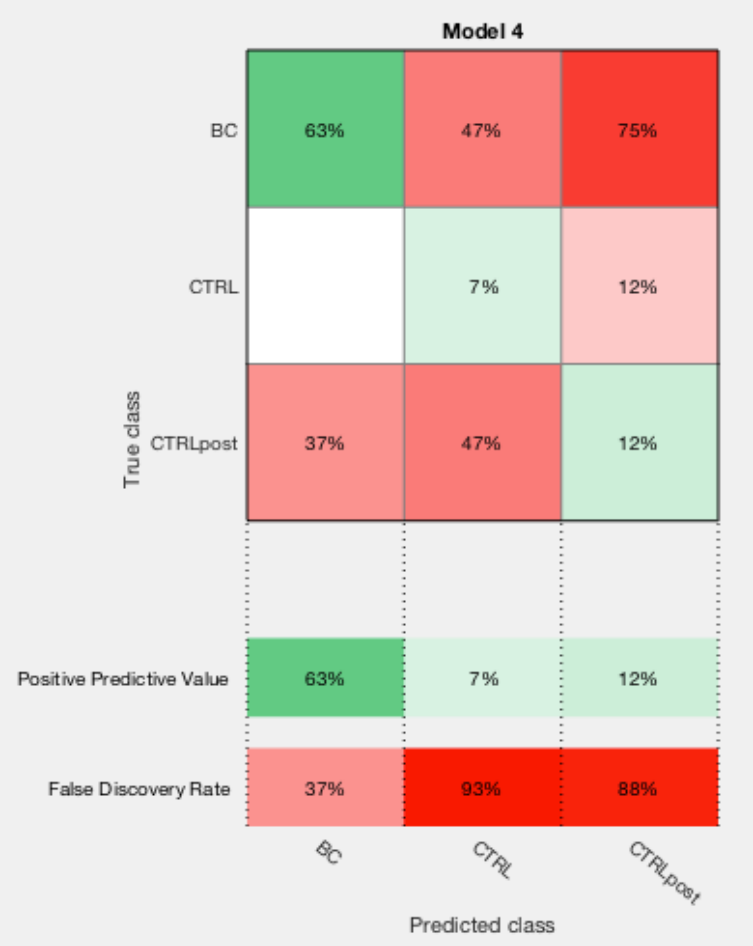

Figura 21: Matriz de confusión para "Xupv_final", análisis discriminante cuadrático sin PCA (3)

De esta matriz se puede extraer la precisión de las predicciones realizadas por el sistema de clasificación. Para el caso de enfermos de cáncer ("BC") el sistema acierta el 63% de sus predicciones. Este valor no es muy elevado, pero si se atiende a los obtenidos anteriormente puede parecer una mejora. Esta esperanza se diluye al observar los ratios para las muestras de pacientes sanos ("CTRL y CTRLpost") para las cuales de las predicciones que hace solamente consigue un 7 y un 12% de acierto, es decir, de cada 10 muestras que predice como paciente sano 8 de estas estarán equivocadas. Estos valores son muy bajos, por lo que se comprueba lo mismo que con las matrices anteriores, el sistema no es válido.

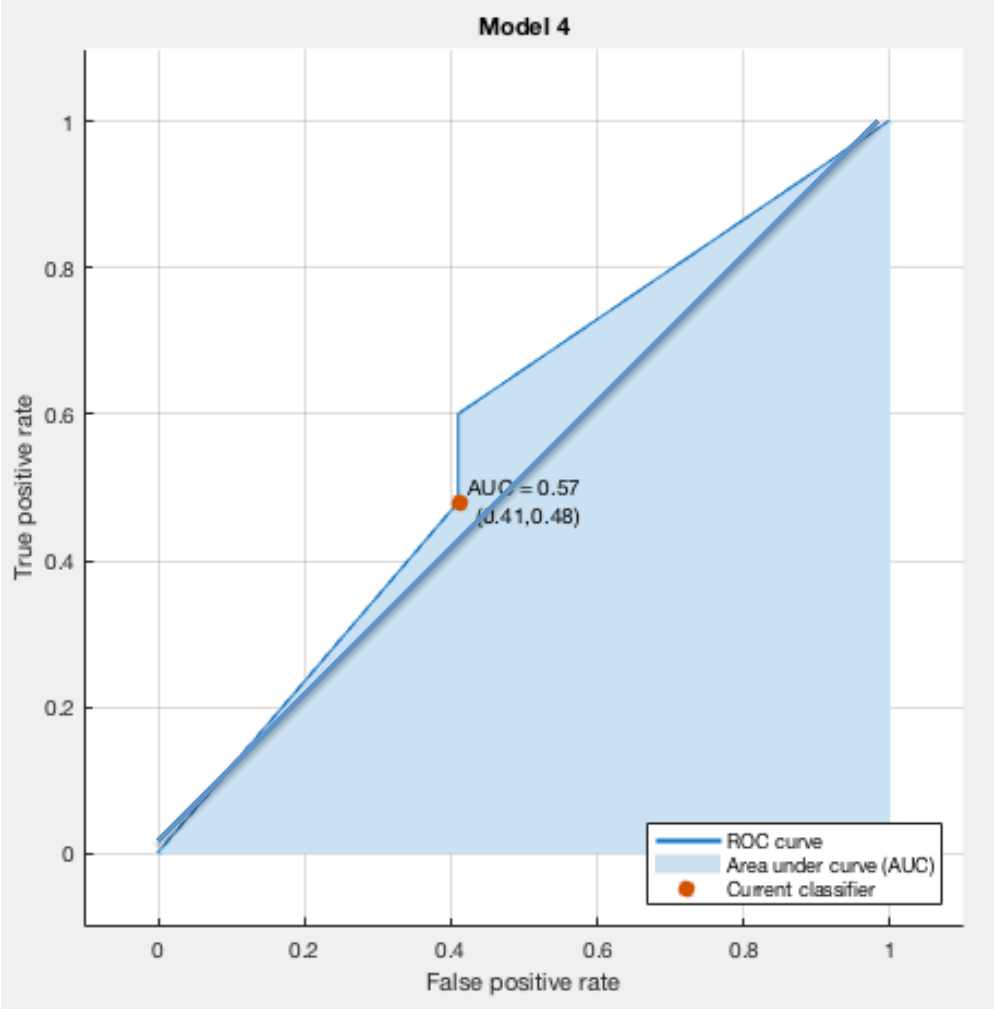

Figura 22: Curva ROC para "Xupv_final", análisis discriminante cuadrático sin PCA

Si se atiende a la curva ROC se comprueba que el valor del área bajo la curva es el mismo que para el caso anterior (0,57). Como se ha comentado anteriormente, este valor corresponde a un sistema de clasificación malo, resultado que ya se esperaba por el dato de la precisión y la matriz de confusión. Sin embargo, esta curva ROC, a pesar de tener el mismo valor AUC que la del caso anterior, es mejor que la anterior, ya que si se compara con el área que supera la diagonal se ha mejorado un poco respecto al caso anterior. Cierto es que esta mejoría es simplemente anecdótica ya que los resultados obtenidos siguen siendo nefastos, por lo que definitivamente se debe descartar este sistema de predicción.

4.1.2 50% HOLD OUT VALIDATION

Tabla 2: Resultados para "Xupv_final" con 50% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	46,7%
Discriminante lineal (Con PCA)	56,8%
Discriminante cuadrático (Sin PCA)	54,1%
Discriminante cuadrático (Con PCA)	45,1%

Al igual que en el caso anterior, separando un 50% para calibración y otro 50% para validación los resultados obtenidos son similares. El mejor resultado parece ser el sistema de análisis discriminante lineal con PCA, pero se ha vuelto a producir la situación de sesgo, por lo que el mejor sistema para estos datos es el sistema de análisis discriminante cuadrático sin PCA. Los demás sistemas se encuentran por debajo del 50% de precisión, es decir, no superan ni la probabilidad de lanzar una moneda al aire.

Análisis discriminante cuadrático sin PCA

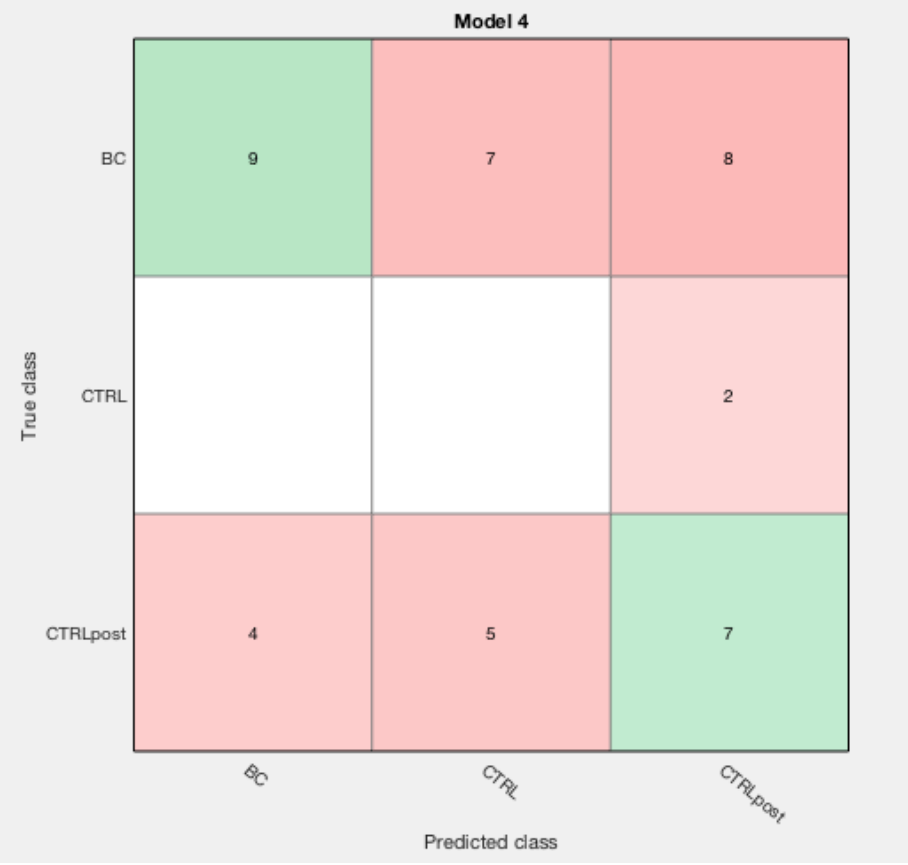

Figura 23: Matriz de confusión para "Xupv_final" con 50% hold out (1)

En este caso se comprueba que se han equiparado las predicciones, en los casos anteriores la mayoría de las muestras eran clasificadas como "BC", mientras que para este caso ya ni siquiera es la predicción mayoritaria, que ha pasado a ser "CTRLpost", la cual anteriormente era la menos predicha. Esto parece indicar que con un mayor número de muestras para validar se mejoran los resultados obtenidos. Aún con esta posible mejora, los resultados siguen siendo muy pobres todavía, parece ser que hay algo que impide una correcta clasificación.

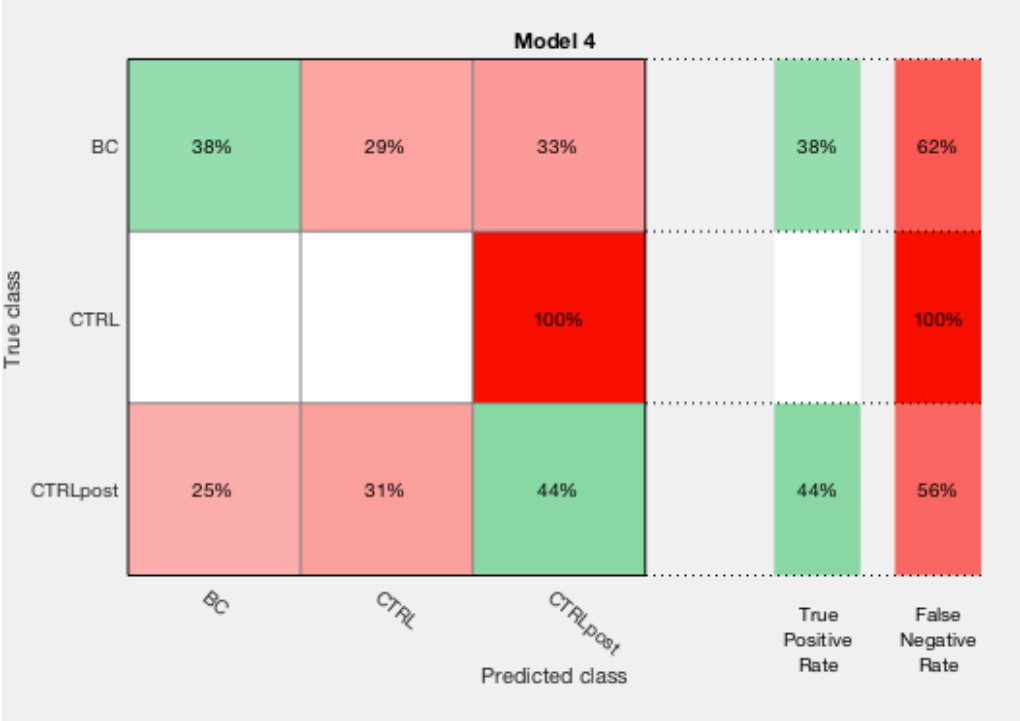

Figura 24: Matriz de confusión para "Xupv_final" con 50% hold out (2)

Las perspectivas de mejora parecen desvanecerse al llegar al análisis de esta matriz, ya que aunque se han predicho más equitativamente las clases, y se han predicho mejor un número mayor de clases que en otros casos, los ratios indican que realmente, basándose en la clase real de las muestras y en las predicciones, los resultados generales bajan. En ningún caso se alcanza el 50% de acierto para las clases reales. Si bien es cierto que los porcentajes son bajos pero se han equiparado, aún así, son demasiado bajos. El caso más preocupante es para la clase "CTRL", para la cual se no se ha predicho bien ninguna de sus muestras, que han sido clasificadas como "CTRLpost". El método de clasificación puede considerarse aleatorio con estos resultados.

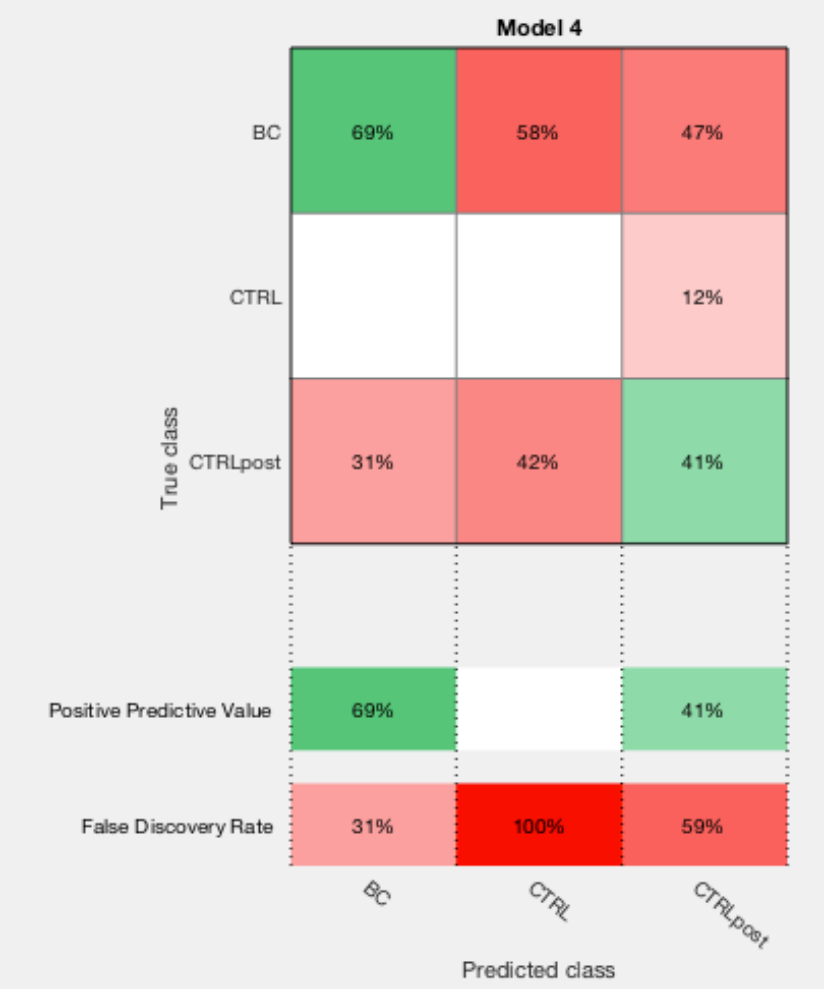

Figura 25: Matriz de confusión para "Xupv_final" con 50% hold out (3)

En esta matriz los resultados mejoran, obteniéndose que cuando se predice que una muestra es "BC" se acierta el 69% de las veces. Este valor muestra una tendencia a aproximarse a los valores deseados. El problema vuelve cuando se examinan las demás clases, para las cuales se han obtenido valores del 0% y del 41% de valor de predicción positivo. Como en el caso anterior, el mayor problema radica en la clase "CTRL", la cual es sistema de análisis es incapaz de predecir correctamente.

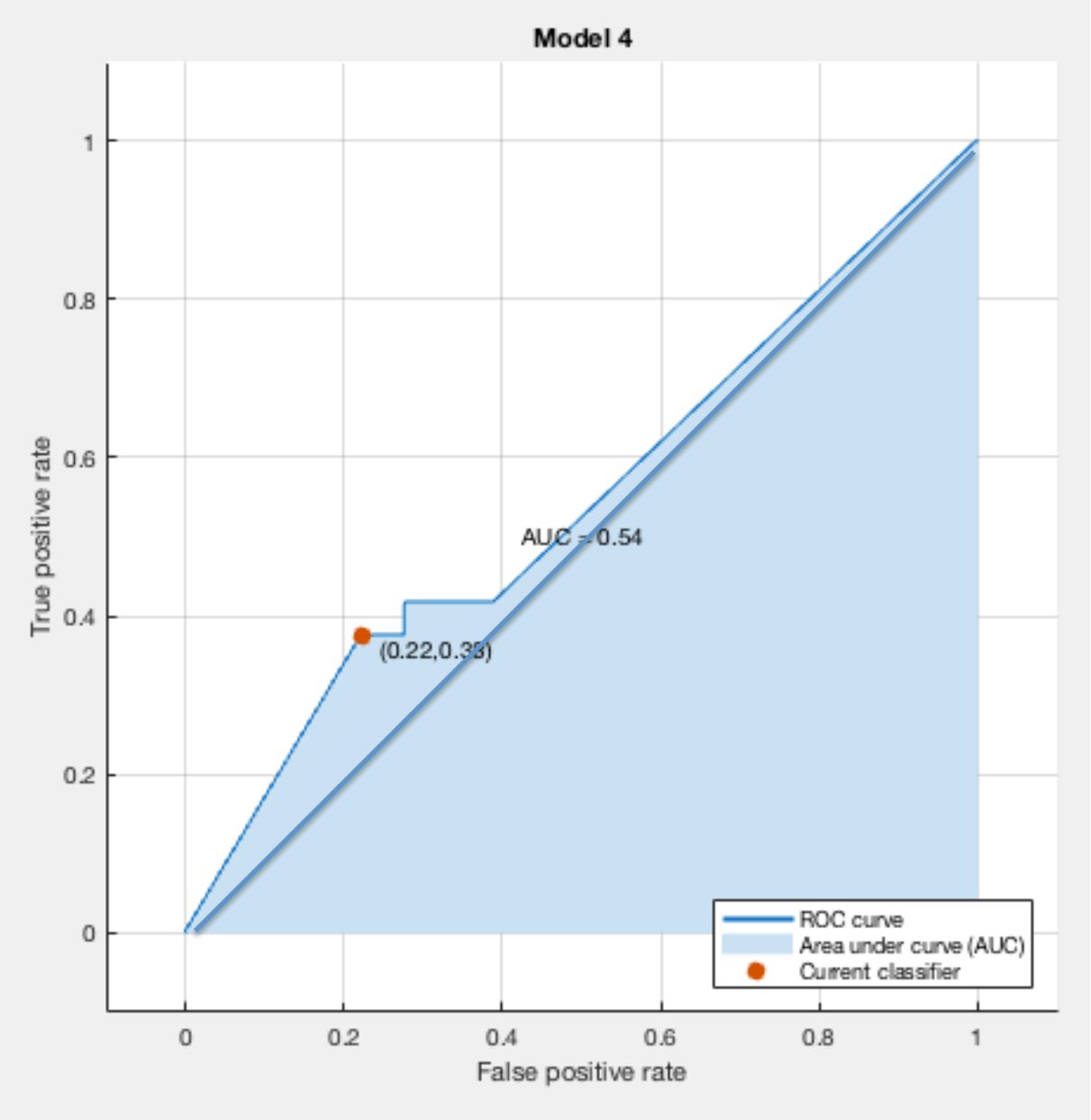

Figura 26: Curva ROC para "Xupv_final" con 50% hold out

En la curva ROC se demuestra lo expuesto, el valor de AUC ha incluso disminuido, por lo que el sistema se considera malo. La curva que se ha obtenido es parecida a la del caso anterior pero invertida, en este caso el pico que sobresale se encuentra en la parte inferior de la diagonal, mientras que en el caso anterior se encontraba en la parte superior. Tras estos resultados, el sistema de análisis debe ser descartado.

4.2 RESULTADOS DE LA MATRIZ INICIAL CON CLASES LÓGICAS "Xlogica_upv"

4.2.1 25% HOLD OUT VALIDATION

Tabla 3: Resultados para "Xlogica_upv" con 25% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	59,5%
Discriminante lineal (Con PCA)	54,8%
Discriminante cuadrático (Sin PCA)	52,4%
Discriminante cuadrático (Con PCA)	47,6%

En este caso los dos mejores resultados corresponden a los análisis discriminantes lineales, tanto con PCA como sin PCA. Se expondrán solo los resultados para el análisis sin PCA, debido a que, como en los casos anteriores, al aplicarlo en análisis lineal, se ha producido un sesgo, clasificando la mayoría de las muestras como clase 1 ("BC").

Análisis discriminante lineal sin PCA

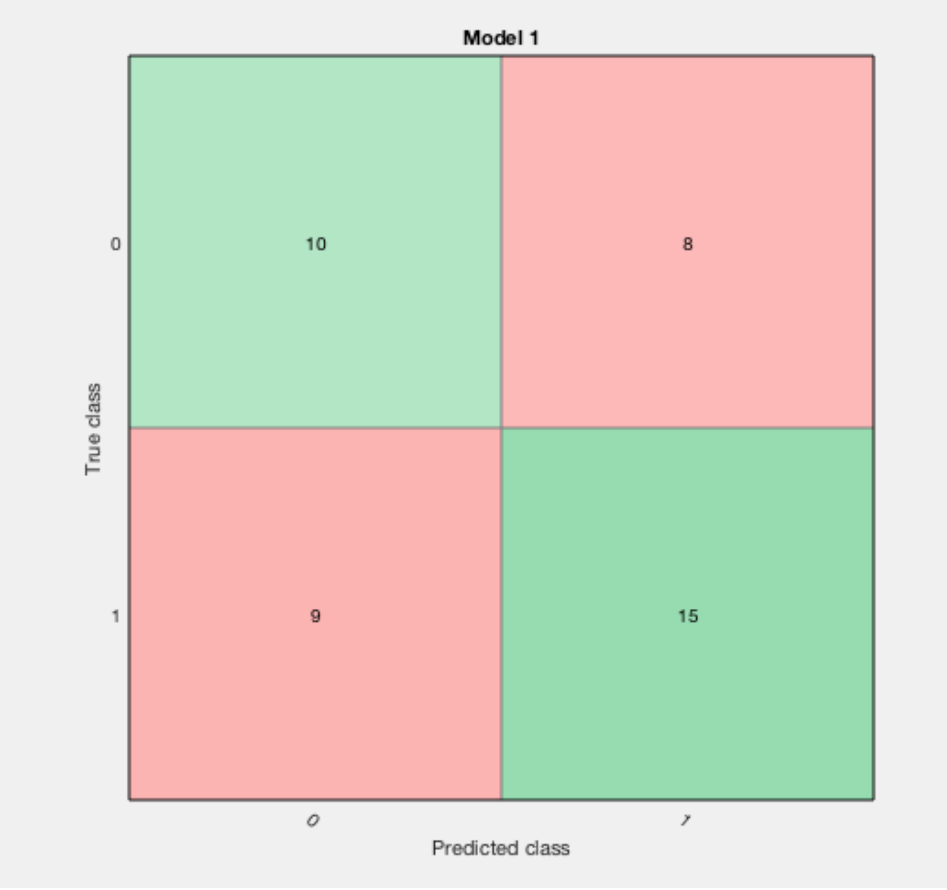

Figura 27: Matriz de confusión para "Xlogica_upv" con 25% hold out (1)

En esta ocasión no se ha producido sesgo hacia ninguna clase. Puede observarse como la predicción para la clase 1 que para la clase 0, pero estos resultados se compararán en las siguientes matrices, lo importante de esta figura es comprobar que el sistema predice ambas clases, con mayor o menor acierto. A priori este sistema sigue teniendo unos resultados de precisión muy bajos, pero pueden ser útiles para valorar posibles mejoras posteriores.

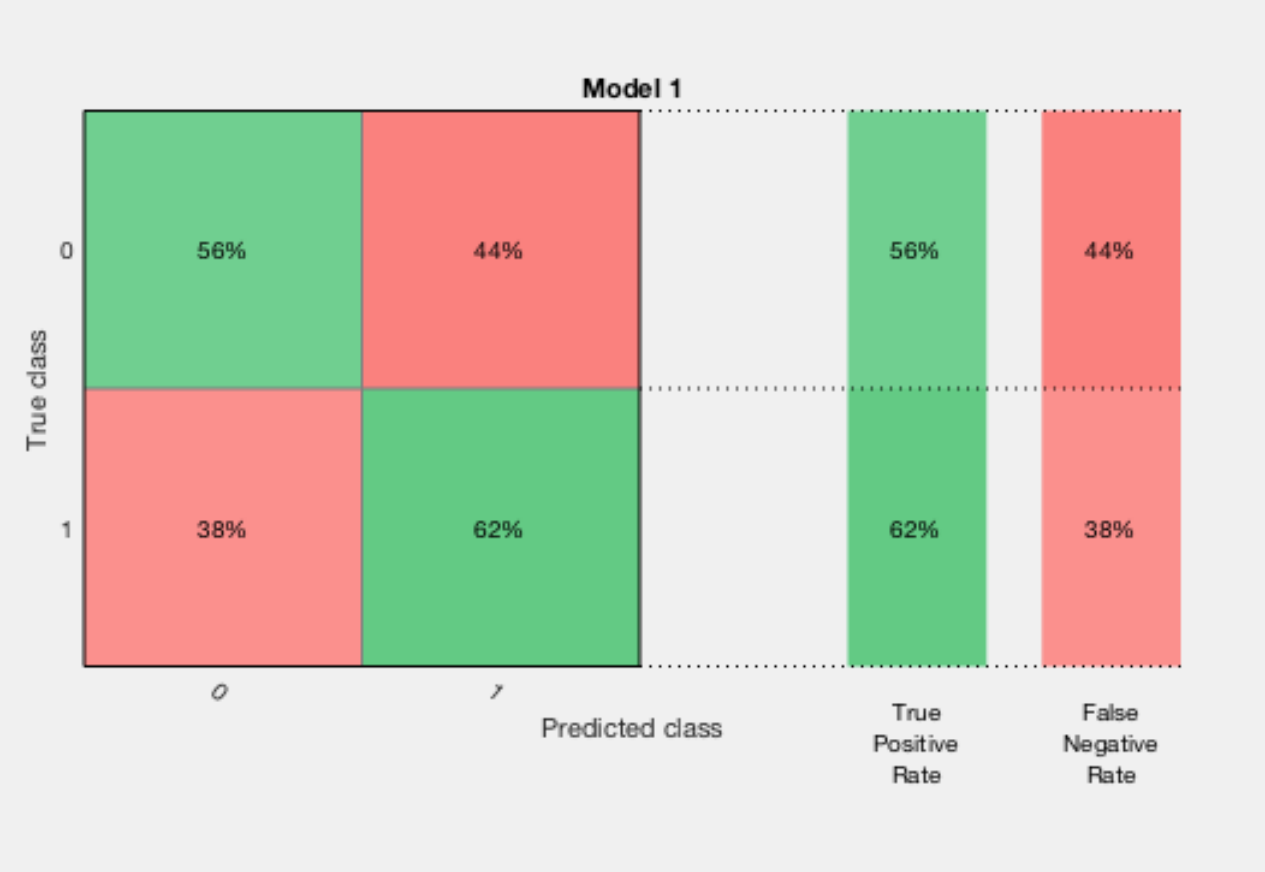

Figura 28: Matriz de confusión para "Xlogica_upv" con 25% hold out (2)

De acuerdo con esta matriz el sistema predice con éxito el 62% de las clases 1 reales y el 56% de las clases 0 reales. Estos resultados son mejores que para los apartados anteriores, pero esta mejora sigue siendo insuficiente para considerar los resultados como buenos. Debe comenzar a valorarse la posibilidad de la existencia de un error, bien sea en el entrenamiento del sistema o en las propias muestras en si.

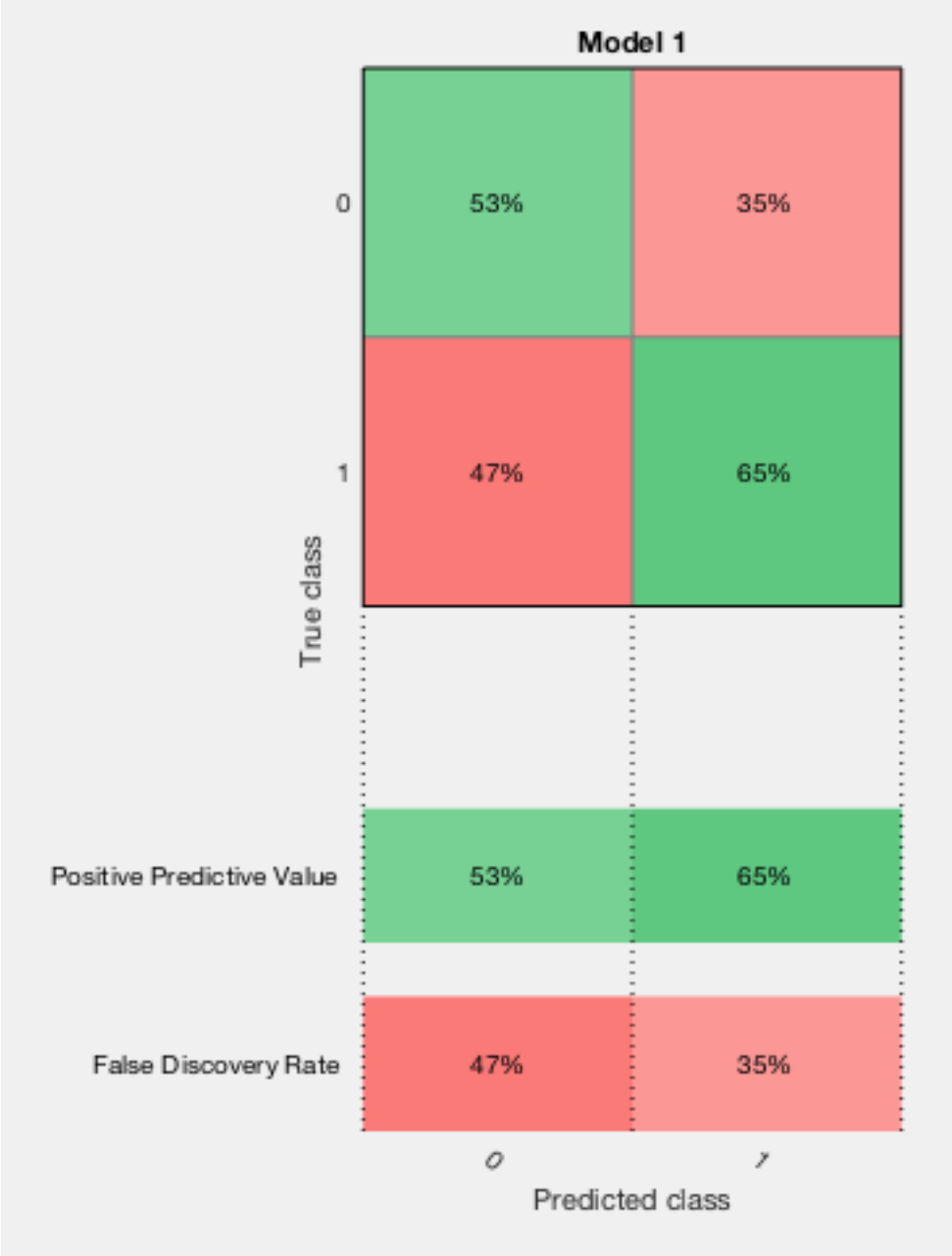

Figura 29: Matriz de confusión para "Xlogica_upv" con 25% hold out (3)

En esta matriz se confirma la mejoría producida al reducir el número de clases de 3 a 2, englobando las clases de pacientes posteriores a la operación en una sola clase. Este sistema acierta el 65% de clases que predice como clase 1 y el 53% de clases que predice como clase 0. Estos resultados pueden indicar que el problema se encuentre en las clases de las muestras de pacientes después de su operación, ya que es en estas clases en las que se produce el mayor error.

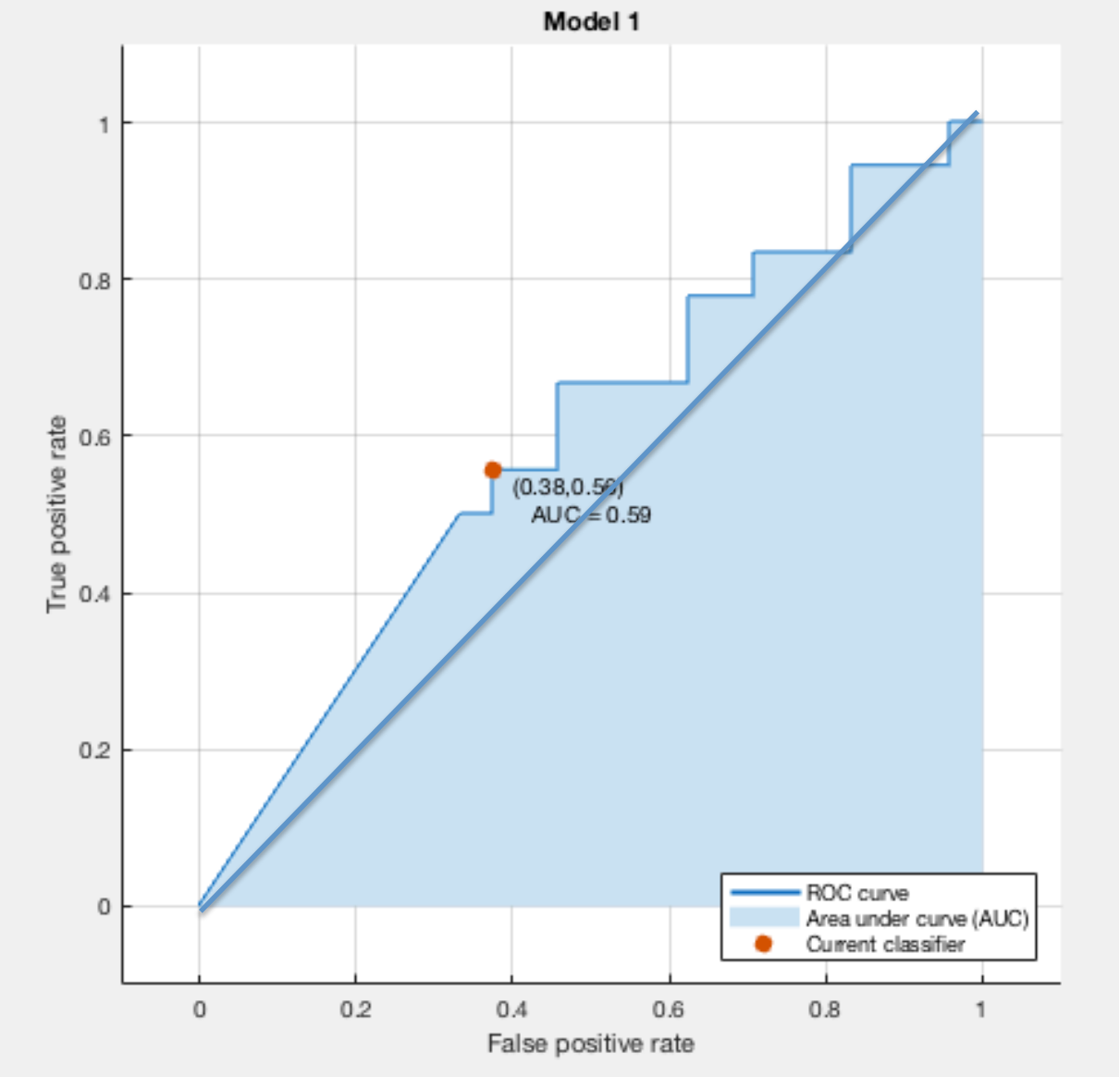

Figura 30: Curva ROC para "Xlogica_upv" con 25% hold out

Si se atiende a la curva ROC, se ha obtenido un valor de AUC de 0,59, con este valor todavía sería considerado como un clasificador malo, pero se encuentra al límite de considerarse un clasificador regular. Aunque esto es insuficiente, se denota la mejoría que se ha producido con el hecho de la reducción de clases a 2, por lo que a falta de comprobar los resultados para un 50% de hold out, esta parece una tendencia a seguir mejor que la utilización de las 3 clases iniciales.

4.2.2 50% HOLD OUT VALIDATION

Tabla 4: Resultados para "Xlogica_upv" con 50% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	56,0%
Discriminante lineal (Con PCA)	53,6%
Discriminante cuadrático (Sin PCA)	44,0%
Discriminante cuadrático (Con PCA)	54,8%

Para un 50% de muestras para validar se produce la misma situación que en el caso de 25%, en el caso de habilitar el análisis de componentes principales se produce un sesgo hacia la clase mayoritaria, que es "BC". Con esto, el mejor resultado se ha vuelto a obtener para el análisis discriminante lineal sin PCA. A continuación se muestran los resultados.

Análisis discriminante lineal sin PCA

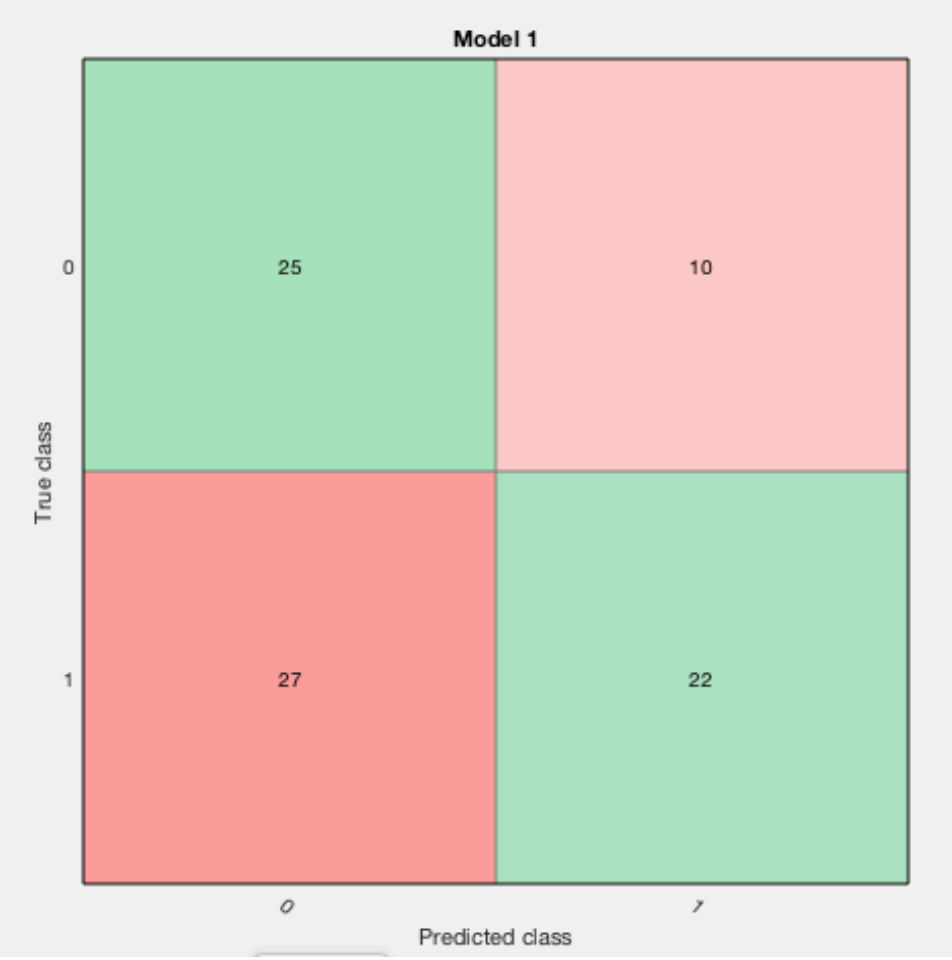

Figura 31: Matriz de confusión para "Xlogica_upv" con 50% hold out (1)

Para un 50% de muestras usadas para calibración y otro 50% de muestras usadas para validación los resultados apenas difieren, la clase 1 parece ser la que mejor precisión tiene mientras que la clase 0 sigue causando más problemas. Para una mejor valoración se deben analizar las siguientes matrices:

Figura 32: Matriz de confusión para "Xlogica_upv" con 50% hold out (2)

Al contrario de lo que se preveía por los datos de la matriz anterior, según estos resultados se han obtenido buenos resultados para la clase 0, mientras que para la clase 1 los resultados han sido peores. Se han predicho con acierto el 71% de las muestras de clase 0 y el 45% de las muestras de clase 1, revirtiéndose los resultados conseguidos para anteriores sistemas. En la siguiente matriz y en la curva ROC se comprobará si este resultado es consistente o no.

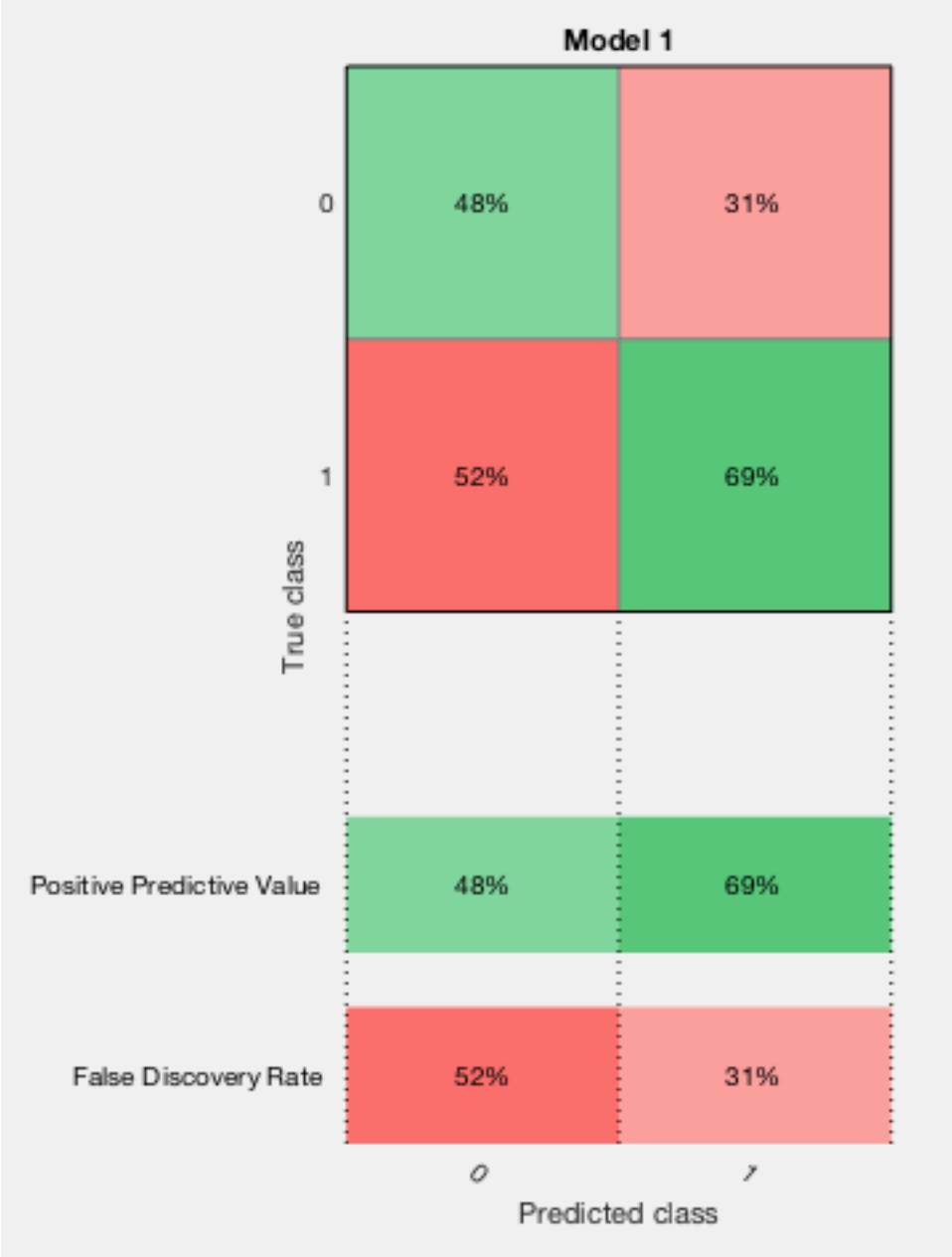

Figura 33: Matriz de confusión para "Xlogica_upv" con 50% hold out (3)

En esta matriz se vuelve a los resultados esperados, comprobándose que la mejoría anterior en la clase 0 era fruto de la casualidad. Se ha conseguido un 69% de acierto para las muestras de clase 1 predichas y un 48% para las clases 0 predichas. Estos valores son muy parecidos a los obtenidos para el sistema con 25% de hold out, por lo que a priori no parece que suponga una diferencia notable en los resultados la utilización o no de más muestras para la validación.

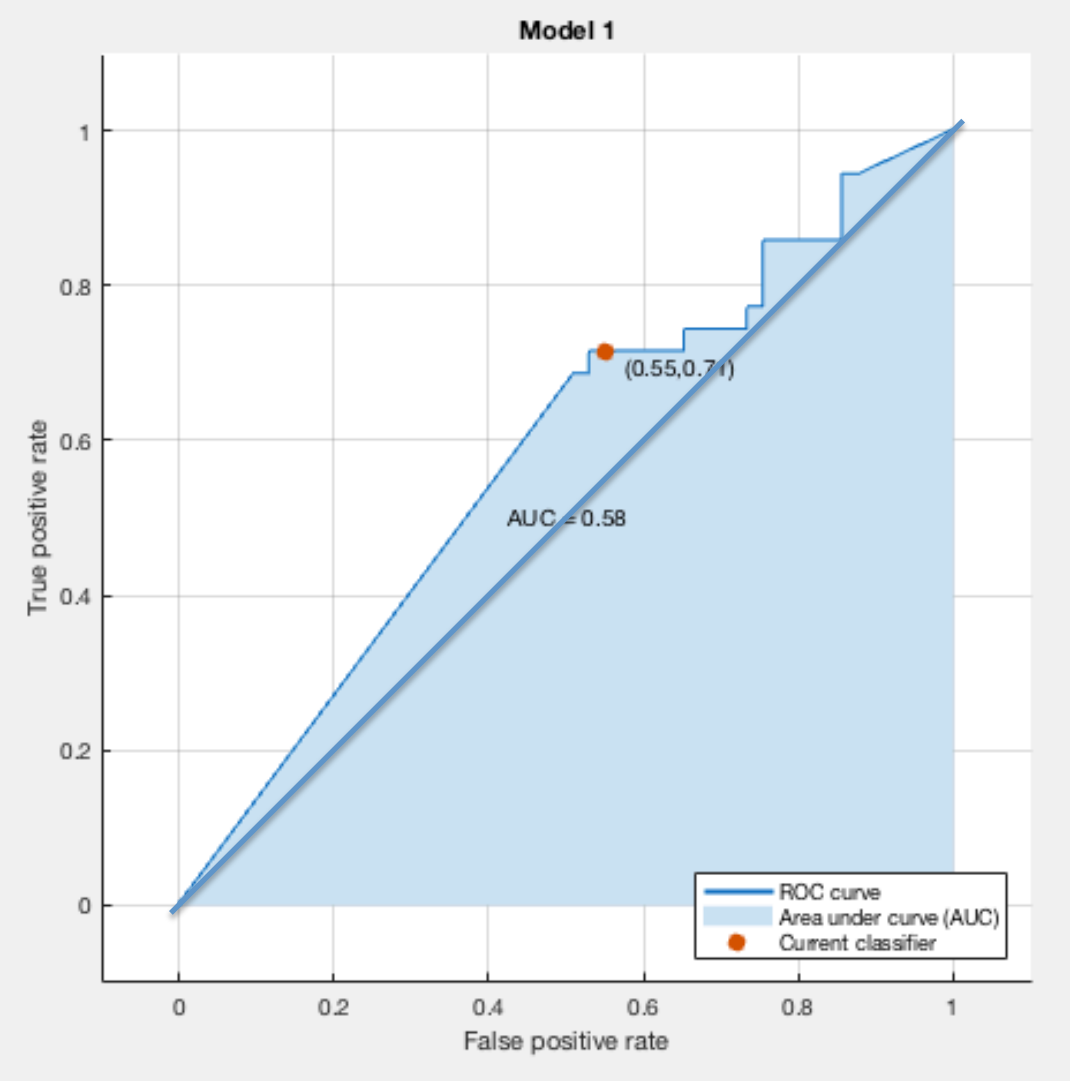

Figura 34: Curva ROC para "Xlogica_upv" con 50% hold out

En este caso el resultado de la curva ROC es muy similar a los resultados para un 25% de hold out, como en el caso de las matrices anteriores, el valor del área bajo la curva en este caso es 0,58 (anteriormente 0,59) por lo que prácticamente no se encuentran diferencias. Lo que si parece demostrarse, a falta de ver las pruebas para las matrices de "Xnew", es la mejoría al utilizar únicamente 2 clases en lugar de 3.

4.3 RESULTADOS DE LA MATRIZ NUEVA "Xnew_clases"

Debido a la similitud con los resultados obtenidos para las matrices de "Xupv", en este caso se omitirán las matrices de confusión que detallan los ratios, a fin de no recargar demasiado el documento con figuras que no aportan información relevante, ya que los resultados son prácticamente idénticos. Por eso, se representa únicamente la matriz de confusión principal y la curva ROC.

4.3.1 25% HOLD OUT VALIDATION

Tabla 5: Resultados para "Xnew_clases" con 25% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	35,7%
Discriminante lineal (Con PCA)	57,1%
Discriminante cuadrático (Sin PCA)	42,9%
Discriminante cuadrático (Con PCA)	42,9%

Con esta matriz de datos se ha vuelto a la situación de 3 clases, lo que parece resulta más complicado de analizar para el Classification Learner, ya que los resultados vuelven a empeorar. Para el caso del análisis lineal con PCA, como en los casos anteriores, se produce un sesgo hacia la clase "BC", pero en esta ocasión los mejores resultados no se han conseguido para el análisis lineal, sino para los análisis cuadráticos. Obteniéndose los mismos resultados tanto con la utilización del análisis principal de componentes como sin él.

Análisis discriminantes cuadráticos

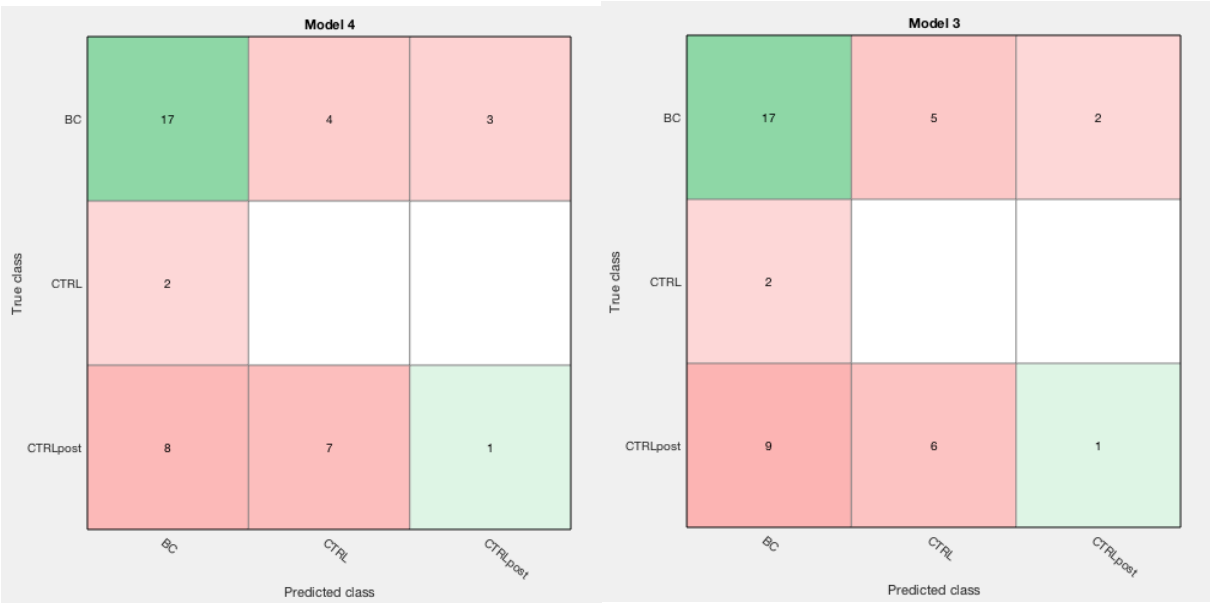

Figura 35: Comparación matrices de confusión para "Xnew_clases" para el análisis discriminante cuadrático (izquierda sin PCA, derecha con PCA)

De estas matrices puede observarse que ambas clasifican correctamente el mismo número de muestras, y se puede esperar que estas muestras hayan sido las mismas, las únicas diferencias se encuentran en las muestras que están mal clasificadas, que algunas de ellas difieren. De aquí se

puede extraer que hay un número de muestras que son correctamente clasificadas independientemente del método utilizado, por lo que su clasificación es más sencilla para el programa que las demás, mientras que hay otras muestras que el programa no reconoce como clasificarlas.

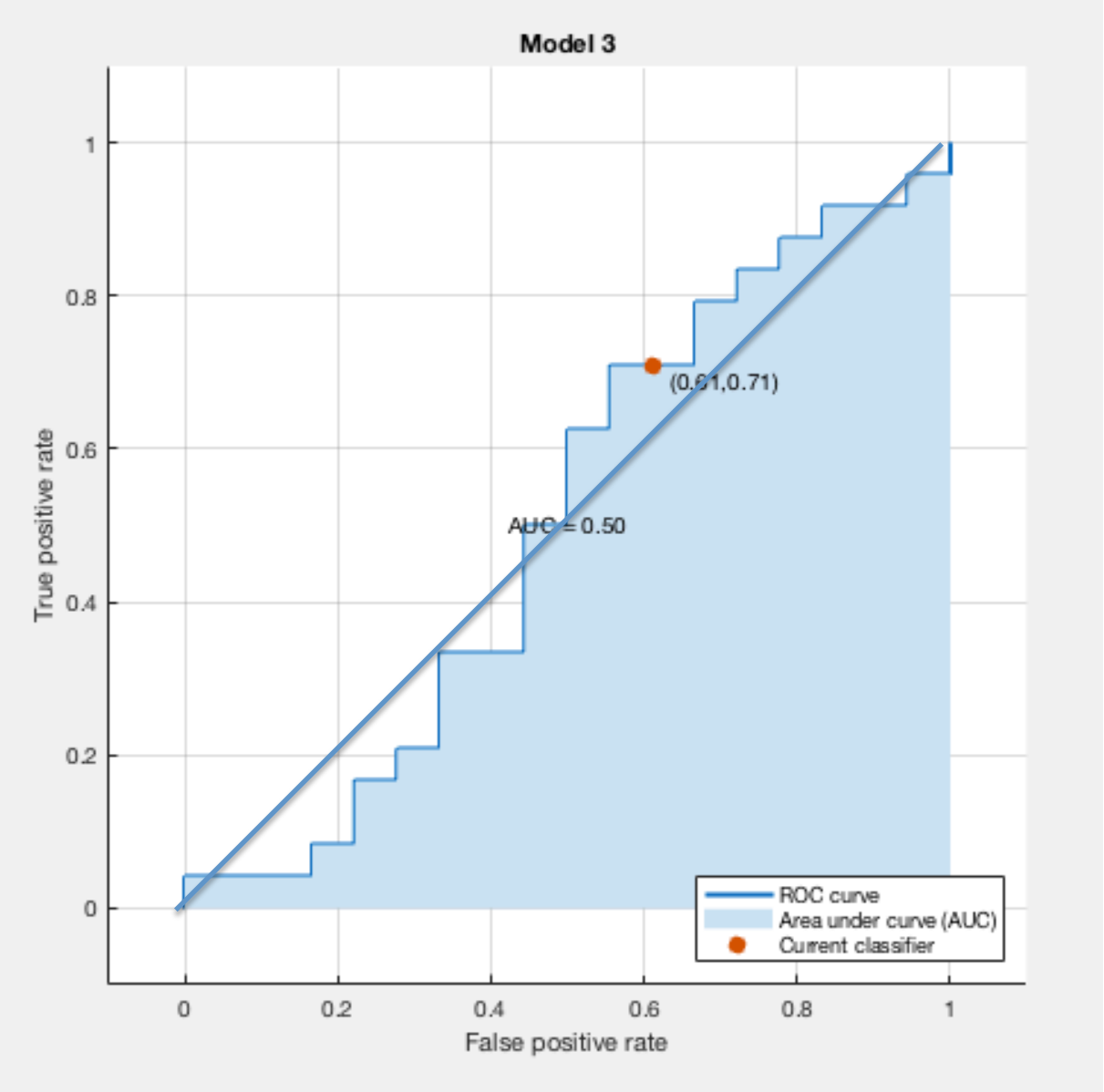

Figura 36: Curva ROC para "Xnew_clases" con 25% hold out

Esta curva ROC corresponde al sistema apoyado con análisis de componentes principales, que es muy similar a la otra. Su valor de área bajo la curva es 50%, un valor muy bajo, que indica que las predicciones tienen la misma fiabilidad que el azar, por lo que a pesar de que estos sistemas a priori calculaban una precisión del 54%, este valor es más bajo según la curva ROC, por lo que este clasificador es categorizado como malo. En las siguientes pruebas se comprobará si el aumento del

número de muestras para validación y la disminución del número de muestras para calibración tienen algún efecto en los resultados, y en el caso de ser así, analizar los resultados y por qué de estos cambios.

4.3.2 50% HOLD OUT VALIDATION

Tabla 6: Resultados para "Xnew_clases" con 50% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	45,9%
Discriminante lineal (Con PCA)	52,9%
Discriminante cuadrático (Sin PCA)	30,6%
Discriminante cuadrático (Con PCA)	38,8%

Al contrario que para las muestras de la variable "Xupv", en este caso el aumentar el número de muestras para la validación y disminuir el número utilizadas para la calibración no ha supuesto una mejora leve de los resultados, al contrario, han empeorado. Se observa que el único sistema entrenado que consigue superar la precisión del azar es el análisis discriminante lineal con PCA, para el cual, como era de esperar, se ha producido un sesgo hacia la clase "BC", por lo que el valor de precisión obtenido no es real. Con esto, los análisis cuadráticos, que con un 25% de validación habían sido los mejores, han pasado a tener unos porcentajes muy pobres, por lo que se analizarán los resultados del análisis lineal sin PCA, con la finalidad de intentar descifrar el por qué de estos resultados tan negativos.

Análisis discriminante lineal sin PCA

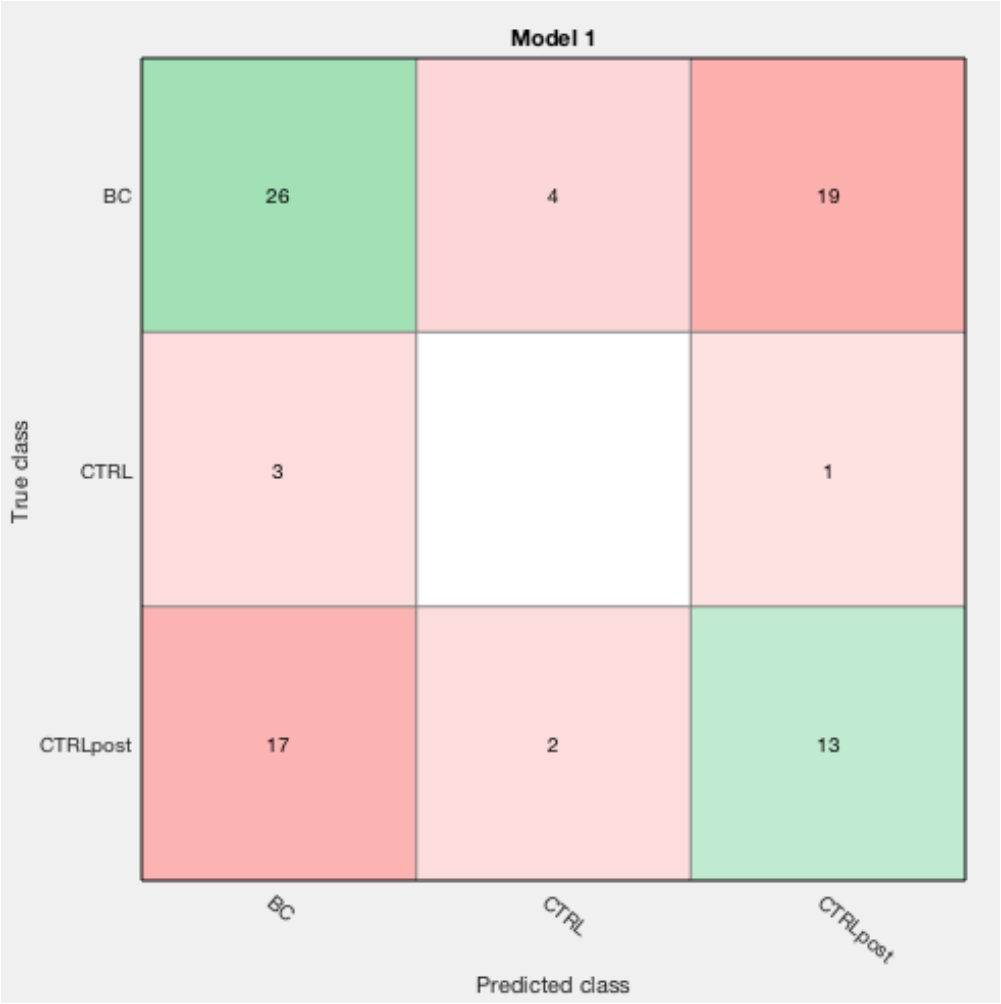

Figura 37: Matriz de confusión para "Xnew_clases" con 50% hold out

Si se analiza la matriz de confusión se comprueba que los resultados obtenidos no son tan diferentes a los obtenidos para un 25% de hold out como podría parecer a priori por los resultados del porcentaje de precisión. El sistema clasifica para las tres clases, encontrándose las mayores dificultades, como ya es costumbre, para las muestras de clase "CTRL" y "CTRLpost". El sistema sigue siendo incapaz de clasificar correctamente una muestra de la clase "CTRL". Para conocer si esta variación es real a efectos de la clasificación se decide valorar la curva ROC.

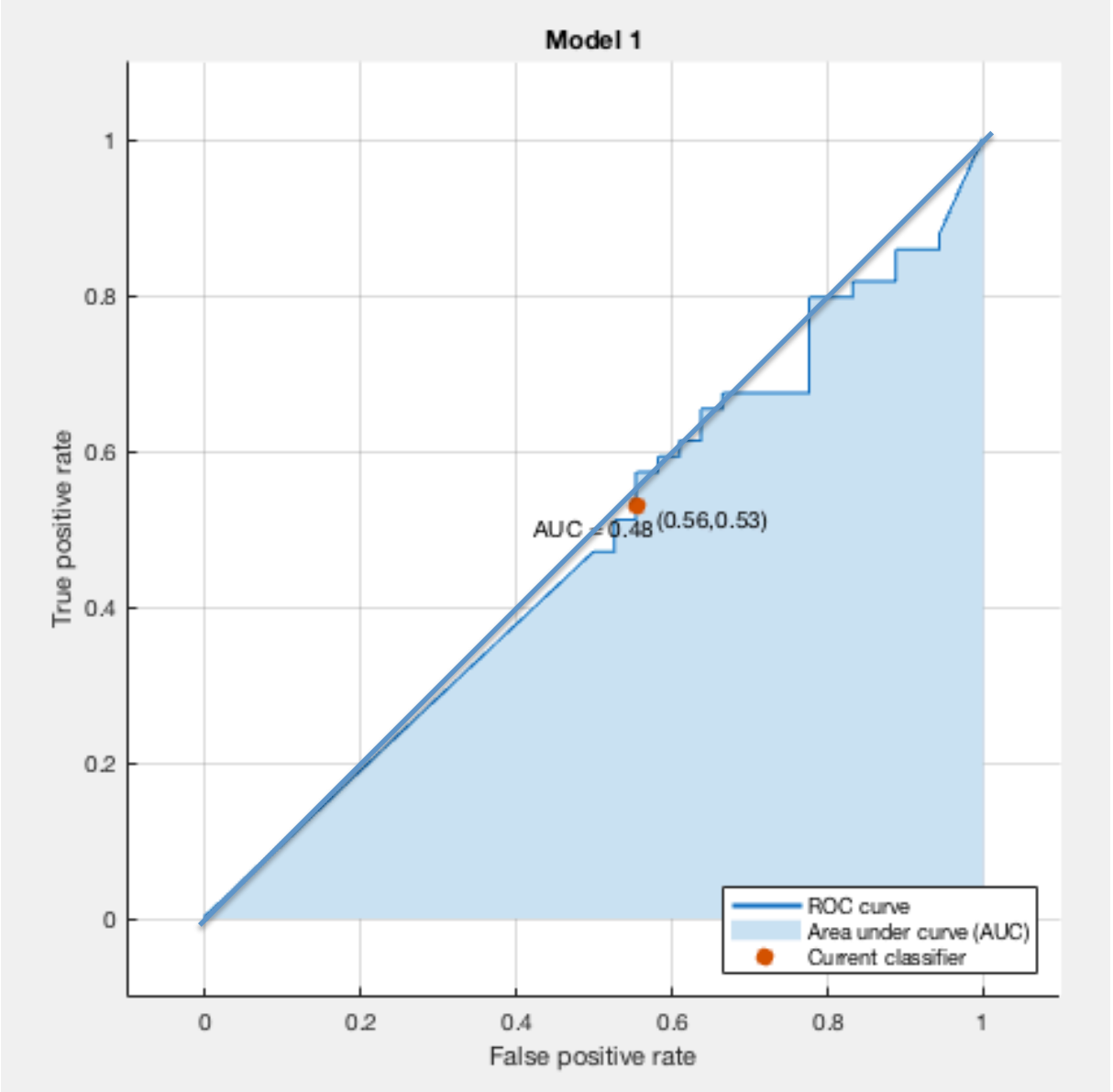

Figura 38: Curva ROC para "Xnew_clases" con 50% hold out

Analizando la curva ROC y comparándola con la curva anterior, el rendimiento ha disminuido desde 0,50 a 0,48. En este caso ya el sistema no es clasificado ni como malo, simplemente obtiene una peor clasificación que si se hiciese al azar, indicando que eliminando muestras de la calibración para utilizarlas en la validación no se encuentra la solución a los problemas de clasificación.

4.4 RESULTADOS DE LA MATRIZ NUEVA CON CLASES LÓGICAS "Xnew_Claseslogicas"

Al igual que para "Xupv", se reducen las tres clases a dos, para simplificar el análisis y comprobar si los resultados mejoran, empeoran o se mantienen. Igualmente, se realizaron pruebas para un 25% de validación y un 50% de muestras para validación.

4.4.1 25% HOLD OUT VALIDATION

Tabla 7: Resultados para "Xnew_Claseslogicas" con 25% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	47,6%
Discriminante lineal (Con PCA)	59,5%
Discriminante cuadrático (Sin PCA)	54,8%
Discriminante cuadrático (Con PCA)	45,2%

Como no es novedad, el mejor resultado se encuentra para el análisis discriminante lineal con PCA, e igualmente que en los demás casos, esto es debido a que se ha producido un sesgo hacia la clase "BC", por lo que para esta matriz en estas condiciones el mejor resultado obtenido ha sido para el sistema de análisis discriminante cuadrático sin análisis de componentes principales. A continuación se muestran los resultados para este sistema.

Análisis discriminante cuadrático sin PCA

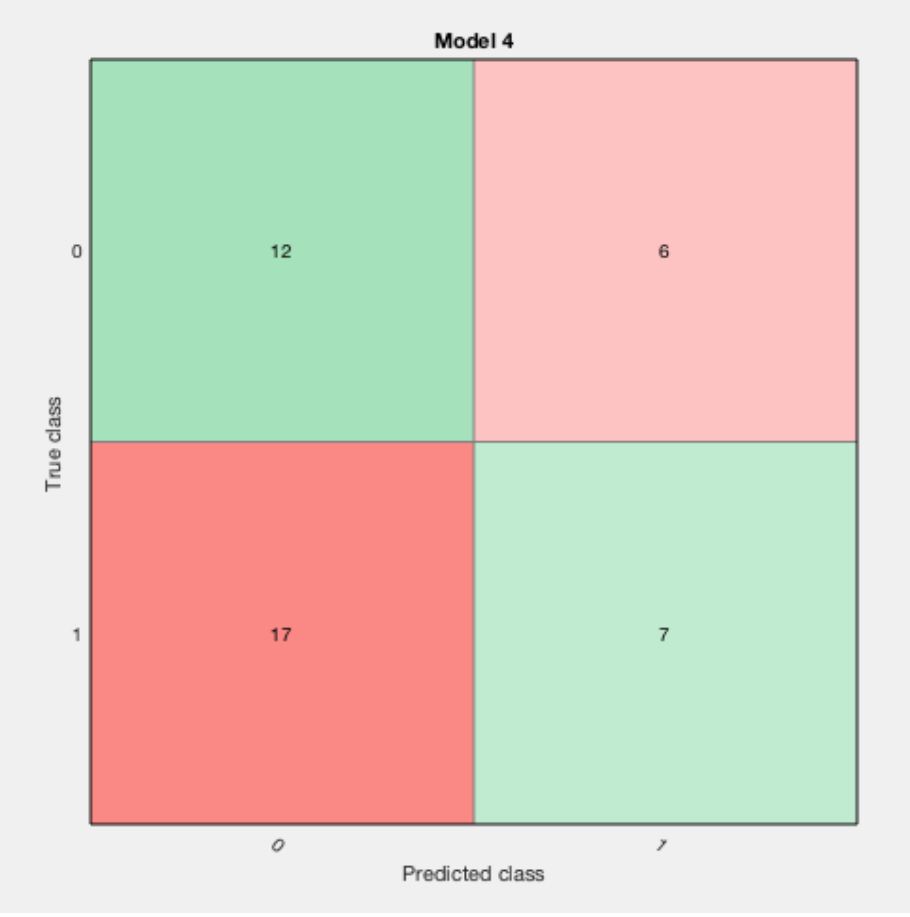

Figura 39: Matriz de confusión para "Xnew_Claseslogicas" 25% hold out

Valorando los resultados de esta matriz se comprueba como la clase predicha en mayor número de ocasiones ha sido la clase 0, en lugar de la clase 1 como en la mayoría de los casos. Este resultado invita a pensar que el sistema de análisis no es muy eficaz, ya que la clase mayoritaria suele ser 1. Esto se ve reflejado en los resultados, ya que el porcentaje de acierto cuando la clase es del 66% aproximadamente, pero en contraposición, el porcentaje acierto cuando la clase es 1 se encuentra alrededor del 30%. A falta de comprobar la curva ROC estos resultados son mucho peores que los obtenidos para la matriz "Xupv" en las mismas condiciones.

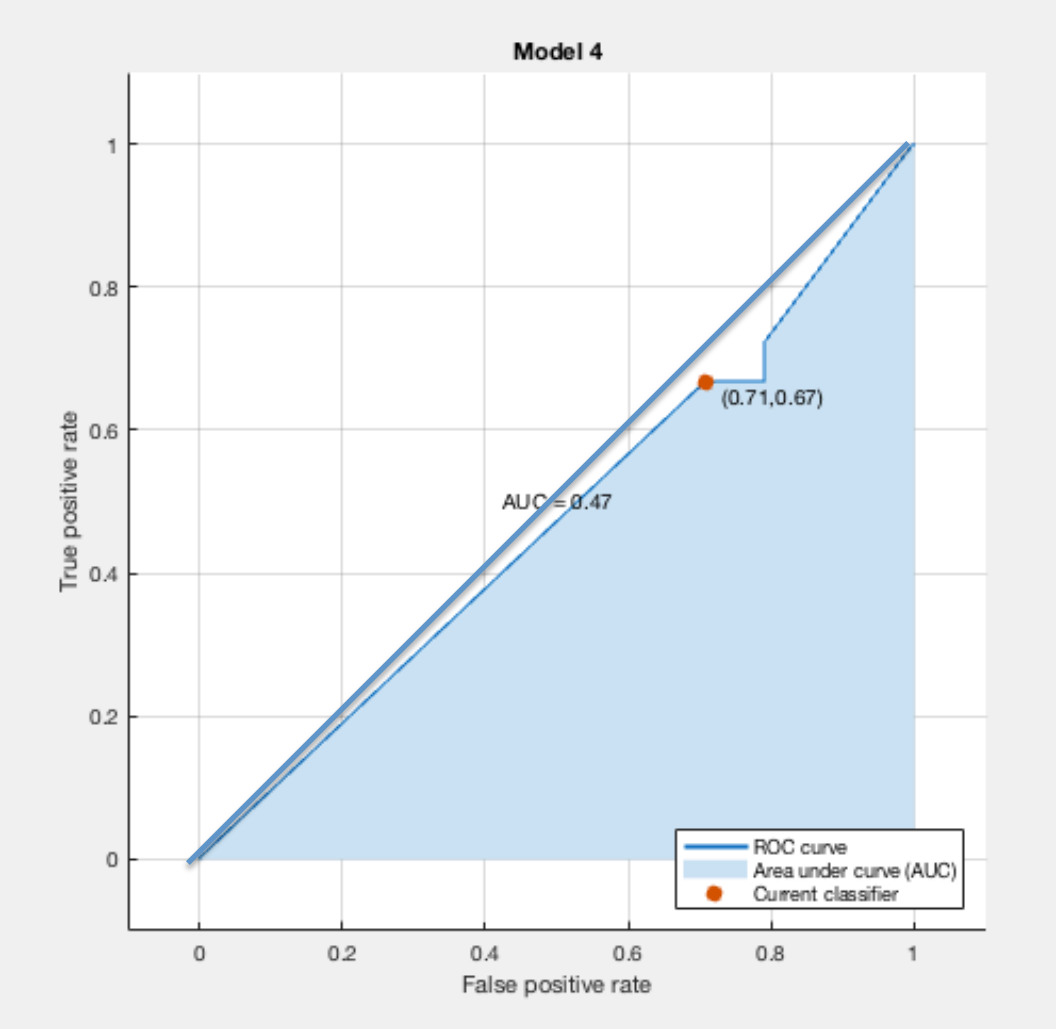

Figura 40: Curva ROC para "Xnew_Claseslogicas" con 25% hold out

La curva ROC confirma los indicios comentados anteriormente, el sistema de clasificación es muy malo. El valor del área bajo la curva es 0,47, encontrándose por debajo del medio punto. Estos resultados han empeorado en comparación a los obtenidos en las mismas condiciones para la matriz de "Xupv", lo que hace pensar que el error podría encontrarse en la hoja Excel y no en la matriz proporcionada en Matlab.

4.4.2 50% hold out validation

Tabla 8: Resultados para "Xnew_Claseslogicas" para 50% hold out

Tipo de análisis	Precisión
Discriminante lineal (Sin PCA)	55,3%
Discriminante lineal (Con PCA)	56,5%
Discriminante cuadrático (Sin PCA)	51,8%
Discriminante cuadrático (Con PCA)	51,8%

Para un 50% de las muestras de la matriz utilizadas para validar el modelo, como era de esperar por los casos anteriores, se ha producido un sesgo hacia "BC" en el análisis discriminante lineal con análisis de componentes. Pero en este caso, el sesgo también se ha producido para los dos análisis cuadráticos, tanto con PCA como sin él. Por lo tanto el único análisis real que se obtiene de estas pruebas es el discriminante lineal sin PCA, ya que la precisión proporcionada por los demás no puede considerarse real.

Análisis discriminante lineal sin PCA

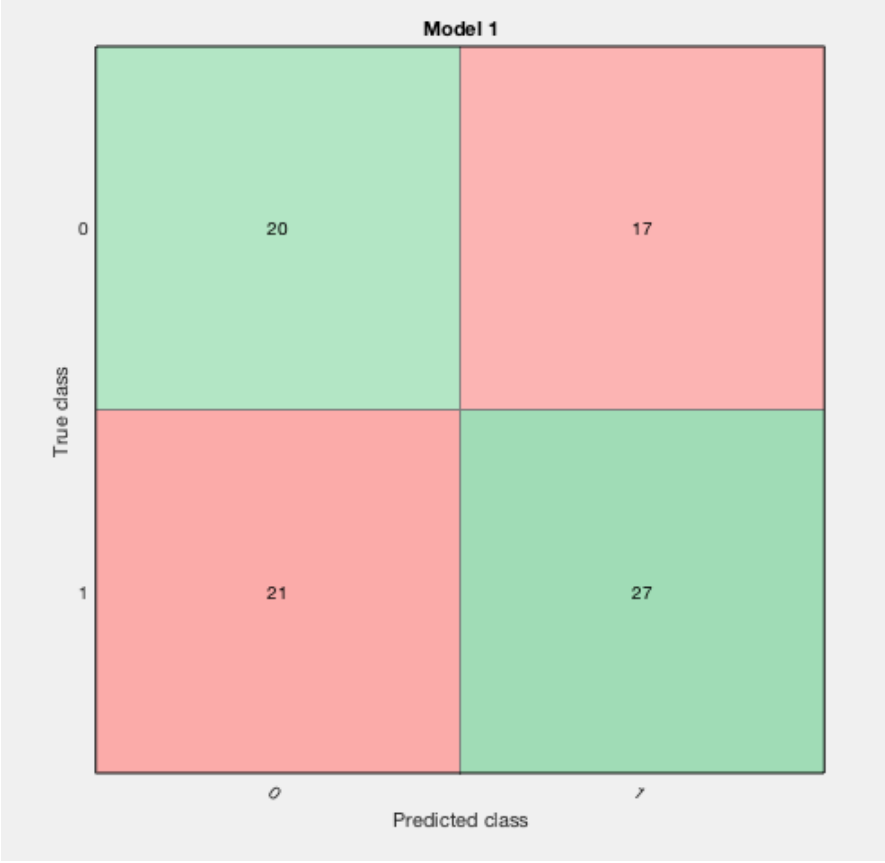

Figura 41: Matriz de confusión para "Xnew_Claseslogicas" con 50% hold out

En base a esta matriz de confusión los resultados parecen haber mejorado en relación a los obtenidos para un 25% de hold out, sobre todo se han equilibrado las clases predichas. En cuanto al porcentaje de acierto de las clases 0 ha disminuido pero para las clases 1 ha aumentado, por lo que en principio parece que el clasificador es mejor, aún estando todavía muy lejos de lo deseado.

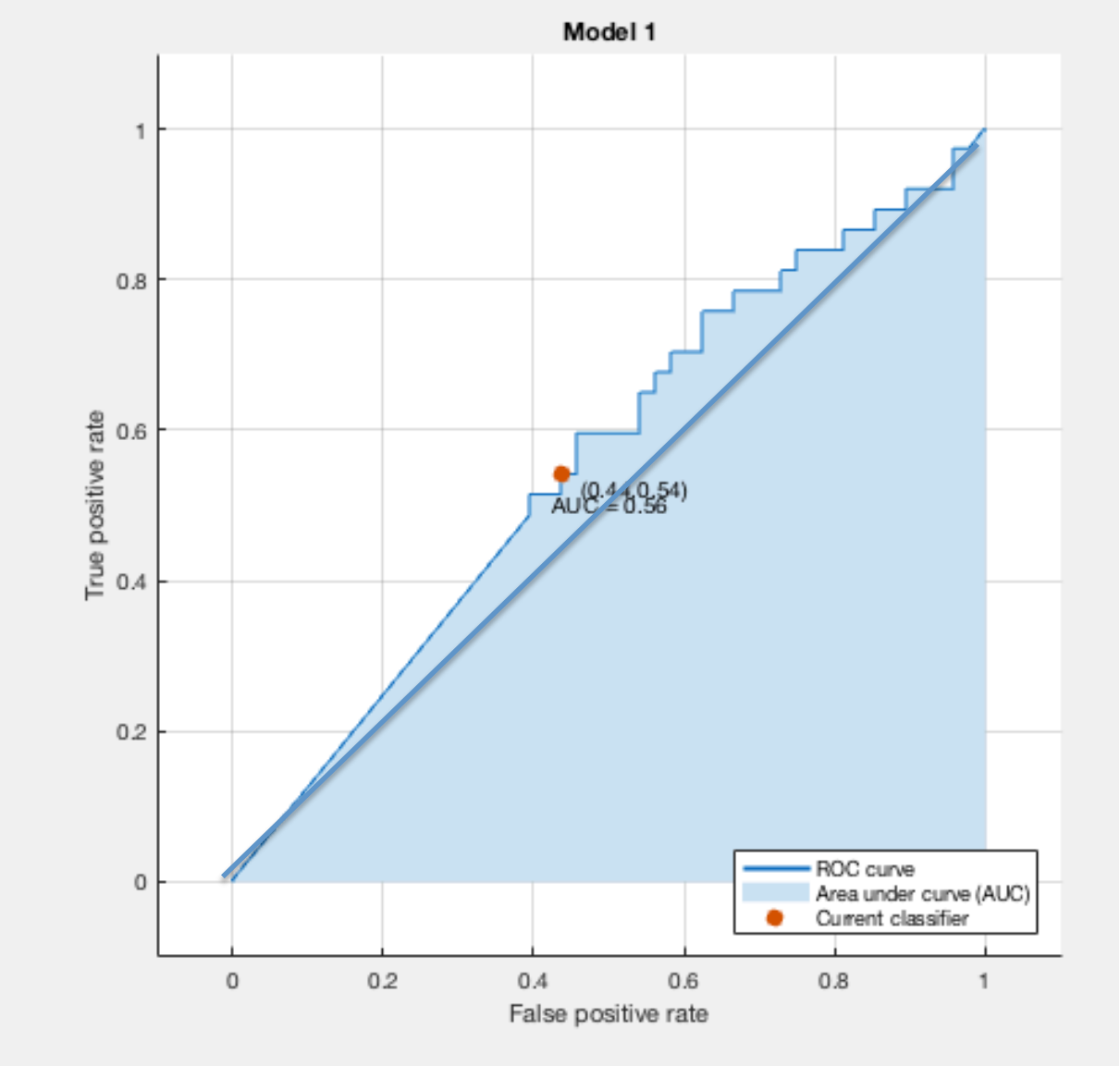

Figura 42: Curva ROC para "Xnew_Claseslogicas" con 50% hold out

En la curva ROC también se ve reflejada esta mejoría, el valor de AUC se ha elevado hasta 0,56. Este valor ya se encuentra más próximo a los obtenidos en estas condiciones para "Xupv", pero todavía el rendimiento sigue siendo menor. En los tres casos con mejores resultados el mejor sistema ha sido el análisis lineal discriminante sin PCA, pero todavía no se han conseguido resultados satisfactorios, los valores de precisión en la predicción siguen siendo muy bajos.

4.5 CONCLUSIONES APARTADO

- Los sistemas entrenados con las muestras con solo dos clases obtienen mejores resultados que los entrenados con muestras con tres clases.

- Las clases que están dando mayores problemas para su clasificación son las clases "CTRL" y "CTRLpost" o la clase 0 cuando se trata con clases lógicas. Esto puede significar que haya algún error en estas.

- No se encuentra una diferencia notable entre los sistemas entrenados con un 25% de hold out y un 50% de hold out, por lo que esto no se considera un factor determinante a la hora de la obtención de resultados.

- No ha habido diferencias significativas en los resultados obtenidos a partir de las matrices "Xupv" y "Xnew", si bien es cierto que los resultados han sido mejores para la matriz de "Xupv", en estas pruebas las diferencias no parecen ser del todo relevantes a la hora de la clasificación. Aún así, si se ha de elegir, la matriz "Xupv" parece más prometedora.

- Cuando se aplica el análisis de componentes principales a los sistemas de análisis se suele producir un sesgo hacia la clase mayoritaria, por lo que de momento parece mejor la no utilización de PCA.

- Los mejores resultados se han obtenido únicamente con dos clases, por lo que a partir de este momento se trabajará con las clases de enfermo "BC" o "1" y sano "0" (agrupa "CTRL" y CTRLpost").

- Los resultados obtenidos en estas primeras pruebas son muy negativos, por lo que deben buscarse soluciones. El primer paso es averiguar que está causando los problemas.

5. PRUEBA DE COMBINACIONES DE ELECTRODOS

Buscando el problema que hace que los resultados de la clasificación no sean los esperados y sean muy bajos, se medita sobre la posibilidad de que esto sea fruto de una sobrecarga de datos al programa, es decir, que al ser la matriz de un tamaño tan voluminoso, el clasificador termine por confundir las muestras, ya que cada una de estas contiene 7729 variables. Para comprobar cuan cierta es esta teoría, se decide analizar las matrices por combinaciones de electrodos. Cada electrodo mide para cada muestra 967 variables, por lo que si en lugar de tener en cuenta las medidas de los ocho electrodos se tienen en cuenta menos, la matriz de datos disminuirá considerablemente, pudiendo comprobarse si el número de variables afecta realmente al entrenamiento. Con esto también se logrará comprobar qué electrodos están dotando al sistema de las mejores medidas y cuales de ellos están causando mayores complicaciones o induciendo a errores. Además, en lugar del método de hold out validation, se utilizará el método de validación cruzada, que consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre las diferentes particiones. Con esto se garantiza que los resultados son independientes de la partición entre los datos de entrenamiento y la prueba, es decir, en el caso anterior podría suponerse que los resultados fuesen equivocados por una mala distribución entre las muestras utilizadas para validar y para calibrar, con la validación cruzada se elimina esta posibilidad.

En definitiva, se realiza, para las dos variables en cuestión, "Xupv" y "Xnew", los cuatro análisis realizados anteriormente, para cada combinación posible de electrodos, con las dos clases lógicas "0" y "1". Desglosando, para la matriz "Xupv", solo con el análisis lineal discriminante sin PCA se realiza:

- 8 análisis para cada electrodo en solitario
- 28 análisis, uno para cada combinación de 2 electrodos
- 56 análisis, uno para cada combinación de 3 electrodos
- 70 análisis, uno para cada combinación de 4 electrodos
- 56 análisis, uno para cada combinación de 5 electrodos
- 8 análisis, uno para cada combinación de 8 electrodos

Esto hace un total de 226 análisis lineales discriminantes sin PCA que abarcan todas las combinaciones posibles de las muestras de pacientes. Este proceso se realiza para los 3 medios de clasificación restantes y para ambas matrices.

Una vez terminados, estos cálculos suponen 1808 sistemas entrenados con diferentes datos de las mismas variables, de los cuales ninguno ha conseguido superar el 60% de eficacia en la

predicción. Los resultados no se adjuntan en esta memoria porque son similares a los obtenidos en el apartado anterior y no aportan nada a las conclusiones.

5.1 CONCLUSIONES DE LA PRUEBA

Tras estos resultados adversos, incluso tras la utilización de la validación cruzada, se concluye que las dimensiones de la matriz no están afectando al entrenamiento del clasificador, y buscando otras posibles alternativas surge el pensamiento sobre la posibilidad que el error lo estuviesen causando las clases "CTRLpost", ya que provienen de muestras de orina de pacientes un tiempo después de haberles realizado la operación, por lo que podría ser que estuviesen causando problemas a la hora de la clasificación.

6. REDUCCIÓN DE LA MUESTRA

Con la finalidad de comprobar lo enunciado anteriormente, se eliminan todas las muestras de clase "CTRLpost" de las matrices de datos "Xupv" y "Xnew", por lo que ahora se trabajará con dos matrices de 106x7729. Estas matrices se llamarán "Xupv_reducida" y "Xnew_reducida" para poder identificarlas con facilidad a lo largo del trabajo restante. En cuanto a las pruebas realizadas en este apartado, son exactamente las mismas que las enumeradas en el apartado anterior, la única diferencia es la eliminación de las clases post y con ellas la disminución del tamaño de ambas matrices de datos.

Una vez realizados los entrenamientos de los sistemas, se comprueba que en la mayoría de los casos los porcentajes de acierto en la predicción han aumentado, sobre todo para el análisis discriminante cuadrático con análisis de componentes principales y validación cruzada, en el cual se obtienen resultados muy superiores a los demás métodos para la mayoría de combinaciones de electrodos posibles. Debido a que el número de análisis (226) es todavía muy extenso, en las siguientes tablas se recogen los resultados más significativos obtenidos. Se expondrán los resultados con una precisión más elevada y se detallará a que combinación de electrodos corresponde para cada matriz de datos.

6.1 RESULTADOS PARA LA MATRIZ INICIAL REDUCIDA "Xupv_reducida"

Resultados para combinaciones de 1 electrodo

Tabla 9: Resultados para combinaciones de 1 electrodo para "Xupv_reducida"

Electrodo	Precisión de la predicción (%)
2º (Rh)	92,45
4º (Au)	91,51
5º (Ag)	90,57
6º (Co)	92,45
7º (Cu)	92,45

Resultados para combinaciones de 2 electrodos

Tabla 10: Resultados para combinaciones de 2 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Ag-Ni	83,96	**Pt-Co**	91,51
Ag-Cu	91,51	**Pt-Ag**	91,51
Ag-Co	92,45	**Rh-Cu**	92,45
Au-Cu	89,62	**Rh-Co**	92,45
Au-Co	90,67	**Rh-Ag**	91,51
Au-Ag	92,45	**Rh-Pt**	86,92
Pt-Cu	90,51	-	-

Resultados para la combinación de 3 electrodos

Tabla 11: Resultados para las combinaciones de 3 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Ag-Co-Cu	91,51	**Pt-Au-Ag**	90,57
Au-Co-Cu	90,57	**Rh-Co-Cu**	92,45
Au-Ag-Cu	92,45	**Rh-Ag-Cu**	91,51
Au-Ag-Co	92,45	**Rh-Ag-Co**	91,51
Pt-Co-Cu	89,62	**Rh-Au-Ag**	91,51
Pt-Ag-Cu	91,51	**Rh-Pt-Cu**	91,51
Pt-Ag-Co	92,45	**Rh-Pt-Co**	91,51

Resultados para la combinación de 4 electrodos

Tabla 12: Resultados para las combinaciones de 4 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Au-Ag-Co-Cu	91,51	**Rh-Au-Ag-Co**	91,51
Pt-Ag-Co-Cu	92,45	**Rh-Pt-Co-Cu**	91,51
Pt-Au-Ag-Cu	92,45	**Rh-Pt-Ag-Cu**	92,45
Pt-Au-Ag-Co	90,57	**Rh-Pt-Ag-Co**	91,51
Rh-Ag-Co-Cu	91,51	**Rh-Pt-Au-Ag**	91,51
Rh-Au-Ag-Cu	91,51	-	-

Resultados para la combinación de 5 electrodos

Tabla 13: Resultados para las combinaciones de 5 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción
Pt-Au-Ag-Co-Cu	91,51	**Rh-Pt-Ag-Co-Cu**	91,51
Rh-Ag-Co-Cu-Ni	82,08	**Rh-Pt-Au-Ag-Cu**	89,62
Rh-Au-Ag-Co-Cu	92,45	**Rh-Pt-Au-Ag-Co**	91,51

Resultados para la combinación de 6 electrodos

Tabla 14: Resultados para las combinaciones de 6 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción
Rh-Pt-Au-Ag-Co-Cu	88,68	**Rh-Pt-Ag-Co-Cu-Ni**	72,64
Pt-Au-Ag-Co-Cu-Ni	72,64	**Rh-Pt-Au-Ag-Cu-Ni**	72,64
Rh-Au-Ag-Co-Cu-Ni	76,42	**Ir-Rh-Au-Ag-Co-Cu**	70,75

Resultados para la combinación de 7 electrodos

Tabla 15: Resultados para las combinaciones de 7 electrodos para "Xupv_reducida"

Combinación de electrodos	Precisión de la predicción (%)
Rh-Pt-Au-Ag-Co-Cu-Ni	74,36

Como puede observarse de estos resultados, las dimensiones de la matriz de muestras influye en una mejor o peor clasificación por parte del sistema entrenado, ya que cuando se aumenta el número de electrodos considerablemente y las variables por muestra empiezan a ser superiores a 5000 la precisión de la predicción comienza a disminuir, también es cierto que los grandes resultados obtenidos para las combinaciones de pocos electrodos pueden estar afectadas por un sesgo, ya que hay una clara dominancia de la clase "1" y es extraño que se obtengan resultados tan elevados con tan poca información. Por otro lado puede intuirse que hay ciertos electrodos que realizan mejores medidas para este análisis, ya que están presentes en los resultados más altos, mientras que otros que apenas aparecen y cuando lo hacen, el resultado tiende a empeorar. Esto último se comprobará en el siguiente apartado, ya que primero se comprueba si los resultados obtenidos se aplican de igual modo a la otra matriz "Xnew_reducida".

6.2 RESULTADOS PARA LA MATRIZ NUEVA REDUCIDA "Xnew_reducida"

En este caso se realizan exactamente los mismos pasos que en el caso anterior, únicamente se modifica la matriz de datos a utilizar para entrenar los sistemas, obteniéndose los siguientes resultados:

Resultados para combinaciones de 1 electrodo

Tabla 16: Resultados para las combinaciones de 1 electrodo para "Xnew_reducida"

Electrodo	Precisión de la predicción (%)
2º (Rh)	91,51
4º (Au)	90,57
5º (Ag)	86,79
6º (Co)	90,57
7º (Cu)	91,51
8º (Ni)	80,19

Resultados para combinaciones de 2 electrodos

Tabla 17: Resultados para las combinaciones de 2 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción
Ag-Cu	90,57	**Rh-Cu**	91,51
Ag-Co	90,57	**Rh-Co**	91,51
Au-Ag	90,57	**Rh-Ag**	90,57
Pt-Cu	90,57	**Rh-Pt**	90,57
Pt-Ag	91,51	**-**	-

Resultados para combinaciones de 3 electrodos

Tabla 18: Resultados para las combinaciones de 3 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Ag-Co-Cu	90,57	**Rh-Ag-Cu**	90,57
Au-Ag-Cu	90,57	**Rh-Ag-Co**	90,57
Pt-Co-Cu	90,57	**Rh-Pt-Cu**	91,51
Pt-Ag-Co	90,57	**Rh-Pt-Co**	90,57
Rh-Co-Cu	91,51	**Rh-Pt-Ag**	90,57

Resultados para combinaciones de 4 electrodos

Tabla 19: Resultados para las combinaciones de 4 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Au-Ag-Co-Cu	90,57	**Rh-Pt-Ag-Cu**	90,57
Pt-Ag-Co-Cu	89,62	**Rh-Pt-Ag-Co**	90,57
Rh-Ag-Co-Cu	90,57	**Ir-Ag-Co-Ni**	90,57
Rh-Au-Ag-Cu	90,57	**Ir-Rh-Ag-Cu**	87,44
Rh-Au-Ag-Co	90,57	**Ir-Rh-Au-Cu**	86,79
Rh-Pt-Co-Cu	90,57	-	-

Resultados para combinaciones de 5 electrodos

Tabla 20: Resultados para las combinaciones de 5 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Pt-Au-Ag-Co-Cu	88,68	**Rh-Pt-Au-Ag-Co**	90,57
Rh-Au-Ag-Co-Cu	88,68	**Ir-Ag-Co-Cu-Ni**	83,02
Rh-Pt-Ag-Co-Cu	90,57	**Ir-Rh-Au-Ag-Cu**	87,74
Rh-Pt-Au-Ag-Cu	90,57	**Ir-Rh-Pt-Ag-Cu**	87,74

Resultados para combinaciones de 6 electrodos

Tabla 21: Resultados para las combinaciones de 6 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)	Combinación de electrodos	Precisión de la predicción (%)
Rh-Pt-Au-Ag-Co-Cu	90,57	**Ir-Rh-Au-Ag-Co-Ni**	86,79
Ir-Au-Ag-Co-Cu-Ni	83,96	**Ir-Rh-Au-Ag-Co-Cu**	83,96
Ir-Pt-Ag-Co-Cu-Ni	81,13	**Ir-Rh-Pt-Ag-Co-Cu**	81,13
Ir-Rh-Ag-Co-Cu-Ni	84,91	**Ir-Rh-Pt-Au-Ag-Cu**	81,13
Ir-Rh-Au-Ag-Cu-Ni	84,91	-	-

Resultados para combinaciones de 7 electrodos

Tabla 22: Resultados para las combinaciones de 7 electrodos para "Xnew_reducida"

Combinación de electrodos	Precisión de la predicción (%)
Ir-Rh-Au-Ag-Co-Cu-Ni	85,85
Ir-Rh-Pt-Au-Ag-Co-Cu	80,19

Estos resultados vienen a confirmar lo supuesto en el caso anterior, el sistema se entrena mejor y clasifica de una manera más eficaz cuando el número de variables por muestras no es excesivamente elevado, o cuando ciertos electrodos no son utilizados en las mediciones. Cuando las combinaciones de electrodos aumentan y el número de variables alcanza valores por encima de 5000 o 6000 variables por muestra el rendimiento del aprendizaje comienza a decaer. Esto también puede ser por lo explicado anteriormente, sobre que esté produciéndose un sesgo hacia la clase mayoritaria y por eso los resultados de la precisión de las predicciones sean tan elevados.

De igual manera ha ocurrido lo esperado para el caso de los electrodos, los mismos electrodos que parecían obtener mejores resultados en el apartado anterior lo corroboran también en este, por lo que el siguiente paso es realizar un análisis para determinar cuales son los mejores electrodos para el trabajo, ya que como ha quedado demostrado, es necesario eliminar los datos de uno o dos de ellos como mínimo para conseguir resultados óptimos.

6.3 CONCLUSIONES DEL APARTADO

- Los resultados comienzan a empeorar una vez se supera un número elevado de variables por muestra, esto puede ser debido al tamaño de la muestra o a ciertos electrodos en particular.

- Ya que en apartados anteriores se comprobó que las dimensiones de las muestras no parecían ser significativas, se entiende que el problema proviene de las medidas de algunos electrodos

- Los resultados han mejorado muy notablemente con la eliminación de las clases "CTRLpost", parece ser que estas variables podían confundir de algún modo al clasificador.

- Puede que muchos de estos resultados no sean reales, ya que al tratarse de un análisis con componentes principales y, según los antecedentes en este trabajo, es probable que muchas combinaciones hayan sido afectadas por el sesgo hacia la clase mayoritaria.

7. Selección de electrodos

Con la finalidad de descubrir que electrodos son los más adecuados para el entrenamiento del clasificador se realizan las siguientes tablas, en las cuales se tiene en cuenta el número de veces que aparece cada electrodo en las combinaciones que han dado mejores resultados en cada caso. Conocido este número y el total de combinaciones elegidas como óptimas simplemente se calcula el porcentaje de ese electrodo en esa serie de combinaciones. Por último, se realiza la media aritmética de los porcentajes de cada electrodo en cada serie de combinaciones. Este estudio se realiza tanto para las combinaciones obtenidas a partir de "Xupv_reducida" como para las combinaciones obtenidas a raíz de "Xnew_reducida", para comprobar cuantos electrodos coinciden como óptimos para ambas matrices de datos.

7.1 RESULTADOS PARA LA MATRIZ INICIAL REDUCIDA "Xupv_reducida"

Tabla 23: Porcentaje de aparición de cada electrodo en las mejores combinaciones para "Xupv_reducida" (1)

Combinación	2 Electrodos		3 Electrodos		4 Electrodos	
Tipo	Veces	Porcentaje (%)	Veces	Porcentaje (%)	Veces	Porcentaje (%)
Ir	0	0,00	0	0,00	0	0,00
Rh	4	30,77	7	46,67	7	63,64
Pt	4	30,77	7	46,67	7	63,64
Au	3	23,08	5	33,33	6	54,55
Ag	6	46,15	10	66,67	10	90,91
Co	4	30,77	8	53,33	7	63,64
Cu	4	30,77	8	53,33	7	63,64
Ni	0	0,00	0	0,00	0	0,00

Tabla 24: Porcentaje de aparición de cada electrodo en las mejores combinaciones para "Xupv_reducida" (2)

Combinación	5 Electrodos		6 Electrodos		Total	
Tipo	Veces	Porcentaje (%)	Veces	Porcentaje (%)	Veces	Porcentaje (%)
Ir	0	0,00	1	16,67	1	3,33
Rh	5	83,33	5	83,33	28	61,55
Pt	4	66,67	4	66,67	26	54,88
Au	4	66,67	5	83,33	23	52,19
Ag	6	100,00	6	100,00	38	80,75
Co	5	83,33	5	83,33	29	62,88
Cu	5	83,33	6	100,00	30	66,21
Ni	1	16,67	4	66,67	5	16,67

Tras estos resultados se escogen 4 electrodos como los electrodos óptimos para los datos de la matriz "Xupv_reducida". Estos electrodos son el electrodo de Plata (Ag), que aparece en un porcentaje del 80% en el total de sistemas entrenados calificados como buenos; el electrodo de cobre (Cu), que aparece en un porcentaje del 66%; el electrodo de cobalto (Co), con un 63% y el electrodo de Rodio (Rh), con un 62%. También puede extraerse que los electrodos que peores resultados están dando son el electrodo de Iridio (Ir) y el electrodo de Níquel (Ni), los cuales tienen un porcentaje de aparición en los sistemas óptimos bastante por debajo de los demás. En cuanto a los electrodos de platino (Pt) y oro (Au) se encuentran en aproximadamente la mitad de las combinaciones, si bien es cierto que la mayoría de las veces aparecen en combinaciones donde hay una gran cantidad de electrodos, mientras que cuando las combinaciones son de 2, 3 o 4 electrodos cuesta más ver su aparición.

7.2 RESULTADOS PARA LA MATRIZ NUEVA REDUCIDA "Xnew_reducida"

Tabla 25: Porcentaje de aparición de cada electrodo en las mejores combinaciones para "Xnew_reducida" (1)

Combinación	2 Electrodos		3 Electrodos		4 Electrodos	
Tipo	Veces	Porcentaje (%)	Veces	Porcentaje (%)	Veces	Porcentaje (%)
Ir	0	0,00	0	0	3	27,27
Rh	4	44,44	6	60	8	72,73
Pt	3	33,33	5	50	4	36,36
Au	1	11,11	1	10	4	36,36
Ag	5	55,56	6	60	9	81,82
Co	2	22,22	6	60	7	63,64
Cu	3	33,33	6	60	8	72,73
Ni	0	0,00	0	0	1	9,09

Tabla 26: Porcentaje de aparición de cada electrodo en las mejores combinaciones para "Xnew_reducida" (2)

Combinación	5 Electrodos		6 Electrodos		Total	
Tipo	Veces	Porcentaje (%)	Veces	Porcentaje (%)	Veces	Porcentaje (%)
Ir	3	37,50	8	88,89	14	30,73
Rh	6	75,00	7	77,78	31	65,99
Pt	5	62,50	4	44,44	21	45,33
Au	5	62,50	6	66,67	17	37,33
Ag	8	100,00	9	100,00	37	79,47
Co	5	62,50	7	77,78	27	57,23
Cu	7	87,50	8	88,89	32	68,49
Ni	1	12,50	5	55,56	7	15,43

Al igual que en el caso anterior, se escogen los 4 electrodos que mejores resultados están aportando para la matriz de datos "Xnew_reducida". Los electrodos escogidos coinciden con los mejores para "Xupv_reducida", como era de esperar, ya que las matrices apenas difieren en sus datos. En este caso el mejor electrodo según la tabla es el electrodo de plata (Ag), con un 79%, seguido por el electrodo de cobre (Cu) con un 68% y el electrodo de rodio (Rh) con un 66%, quedando en cuarto lugar esta vez el electrodo de cobalto (Co) con un 57%, bastante lejos ya del porcentaje del electrodo 5. El electrodo de Iridio (Ir) y el electrodo de Níquel (Ni) han aumentado su

porcentaje en esta ocasión, equiparándose prácticamente a los porcentajes de los electrodos platino (Pt) y oro (Au), todavía un poco por debajo. El problema con estos cuatro electrodos es el mismo, su aparición comienza a hacerse factible una vez el número de electrodos por combinación es superior a 5, más que nada porque no hay otros electrodos que pueden hacer su función, pero no implica que estén mejorando los datos aportados por los 4 electrodos seleccionado como óptimos, tampoco implica que los empeoren, la disminución de la precisión puede ser debida a varios factores.

8. Prueba a las mejores combinaciones

Una vez estimados los mejores electrodos para el entrenamiento del sistema por parte de la aplicación se decide realizar un análisis más minucioso a las mejores combinaciones obtenidas, por lo que para cada número de electrodos de 2 a 6 (1,7 y 8 se descartan por obtenerse peores resultados) se escogen las tres mejores combinaciones obtenidas y se realiza la matriz de confusión y la curva ROC a cada una de ellas. Este análisis se realizará para la matriz "Xupv_reducida", descartando la matriz "Xnew_reducida", ya que desde el primer momento se han obtenido mejores resultados para la matriz inicial de la universidad y el hospital, por lo que se asume que el error en la numeración estaba en la hoja de cálculo Excel y no en la matriz de trabajo de Matlab. Las combinaciones elegidas serán las que contengan un mayor número de electrodos óptimos, y se escogen para cada número de electrodos las tres combinaciones más esperanzadoras. Las combinaciones elegidas fueron:

- 2 electrodos: Rh-Ag, Ag-Co y Ag-Cu.
- 3 electrodos: Rh-Ag-Co, Rh-Ag-Cu y Ag-Co-Cu.
- 4 electrodos: Rh-Pt-Ag-Cu, Rh-Ag-Co-Cu y Pt-Ag-Co-Cu.
- 5 electrodos: Rh-Pt-Ag-Co-Cu, Rh-Au-Ag-Co-Cu y Pt-Au-Ag-Co-Cu.
- 6 electrodos: Rh-Pt-Au-Ag-Co-Cu, Rh-Pt-Ag-Co-Cu-Ni y Rh-Au-Ag-Co-Cu-Ni.

Además, para cada combinación específica de electrodos se han detectado las muestras que podrían ser outliers, las cuales han sido eliminadas previamente antes del entrenamiento del sistema de análisis con el Classification Learner. Es decir, se eliminan un número de muestras diferentes para cada combinación. No son eliminadas las mismas muestras para la combinación de Rh-Ag que para la combinación de Ag-Co o Ag-Cu. Con esto se intenta eliminar al máximo posible la posibilidad de datos que causen error en el entrenamiento y clasificación de las muestras.

8.1 RESULTADOS

Combinación Rh-Ag

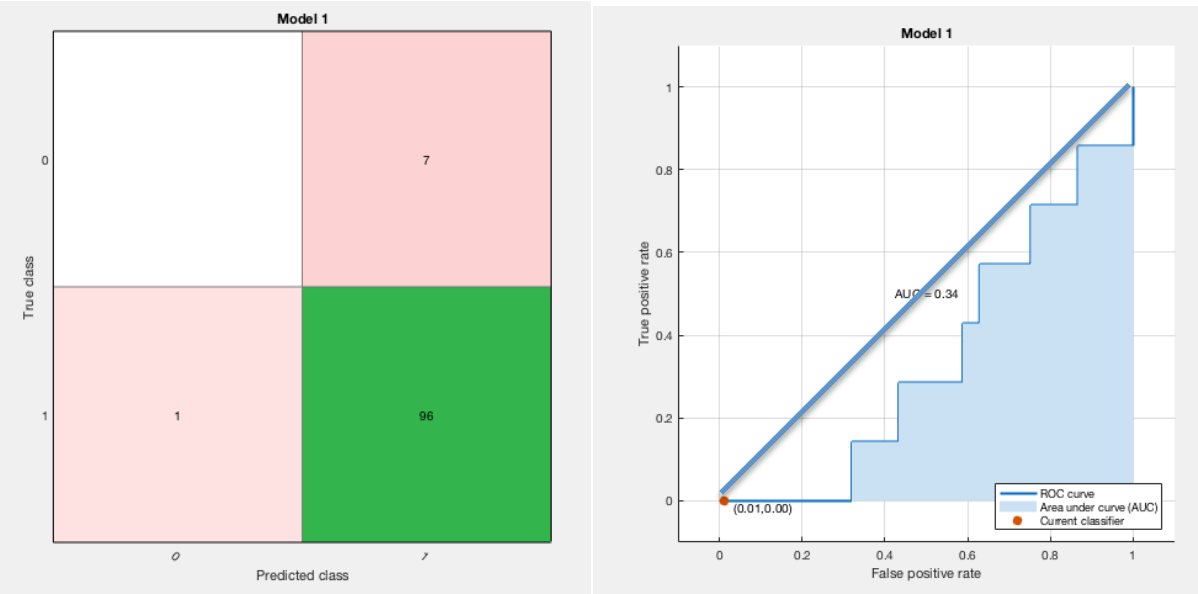

Figura 43: Matriz de confusión y curva ROC para la combinación 25

Como puede observarse, en esta combinación se ha producido un sesgo hacia la clase "1", ya que se han predicho todas las muestras menos una como tal. Aparte, la muestra que se ha predicho como clase "0" ha sido errónea, por lo que el porcentaje de acierto de la clase de pacientes sanos es nula. Si e atiende a la curva ROC se comprueba que el rendimiento es malísimo, ya que el valor del área bajo la curva ni siquiera llega a 0,4. Se encuentra muy por debajo incluso del porcentaje de acierto que se obtendría con el puro azar. Este sistema no puede ser considerado válido para la clasificación de muestras posteriores.

Combinación Ag-Co

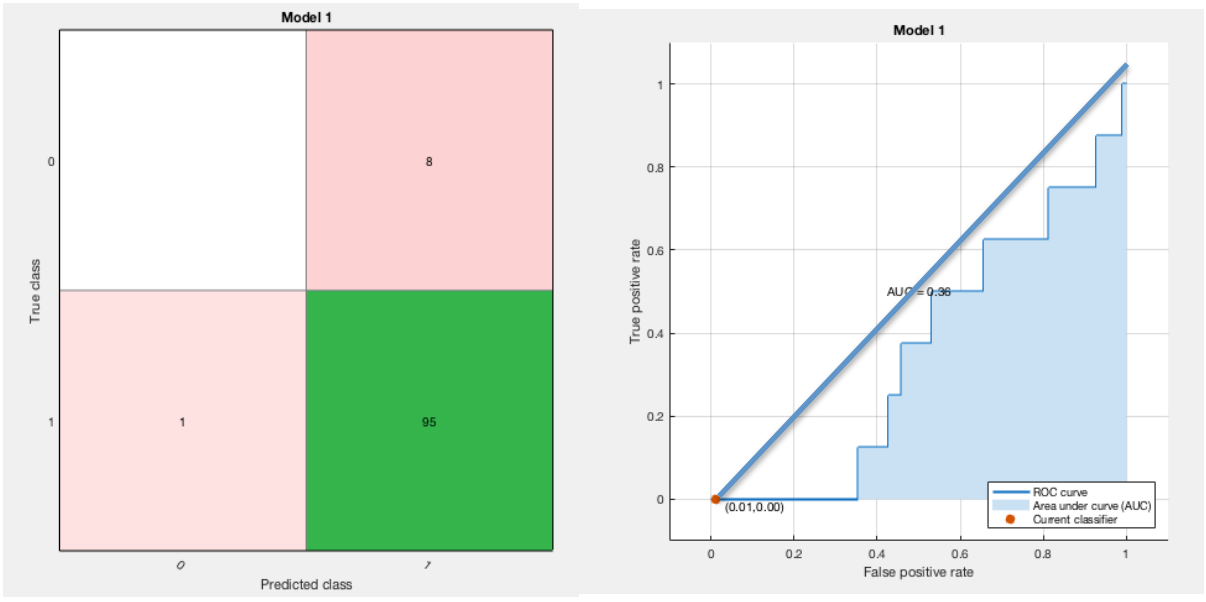

Figura 44: Matriz de confusión y curva ROC para la combinación 56

Para esta combinación ocurre exactamente lo mismo que en el caso anterior, se ha producido un sesgo y no se ha predicho ninguna clase "0" correctamente, por lo que la capacidad del sistema para predecir esta clase es nula. Simplemente al predecir todas las muestras como clase 1, y estas serlo en su gran mayoría, se ha obtenido un porcentaje muy elevado de precisión. En cuanto a la curva ROC se observa a pesar del sesgo que el rendimiento es algo mejor que el anterior. Esto puede indicar que el electrodo de plata funciona algo mejor con el electrodo de cobalto que con el electrodo de rodio.

Combinación Ag-Cu

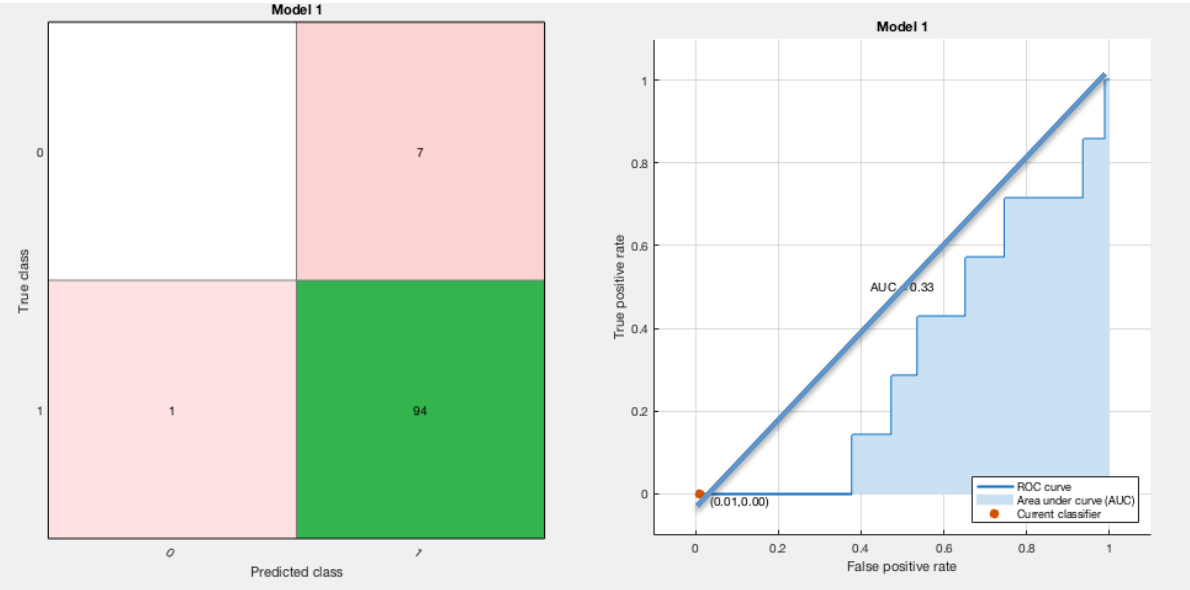

Figura 45: Matriz de confusión y curva ROC para la combinación 57

De igual modo que para las combinaciones anteriores, los resultados prometedores de esta prueba son engañosos, se ha producido un sesgo hacia la clase 1. El sistema no puede ser de ningún modo válido con estos resultados, pero la curva ROC puede servir para extraer datos interesantes, como que el valor del área bajo la curva ha vuelto a decaer, indicando que el electrodo de plata obtiene mejores resultados si se analiza conjuntamente con el electrodo de cobalto que con el de cobre y rodio.

En definitiva, las combinaciones de 2 electrodos no han conseguido los resultados esperados, ya que en todas se ha producido un sesgo hacia la clase 1, esto puede ser debido o a la falta de más variables (número muy pequeño de electrodos) o a la falta de un mayor número de muestras de clase 0. Para esta caso en particular, la conclusión es que ambos factores están perjudicando los resultados de estos sistemas, se necesitan tanto más electrodos de medida como más variedad de muestras.

Combinación Rh-Ag-Co

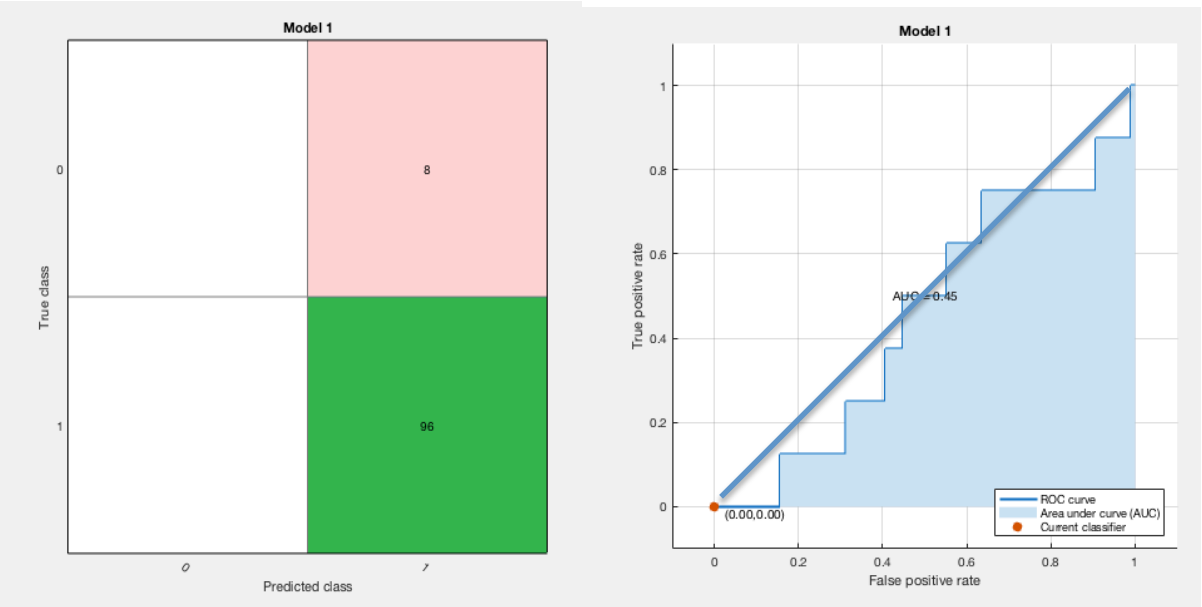

Figura 46: Matriz de confusión y curva ROC para la combinación 256

En este caso ni siquiera se ha llegado a predecir ninguna muestra como clase 0, por lo que en este caso el sesgo producido ha sido total. De ningún modo este sistema puede ser válido en estas condiciones para predecir muestras de futuros pacientes. Lo interesante de esta combinación viene en la curva ROC, ya que parece que al aumentar el número de electrodos aumenta su valor de área bajo la curva. En las siguientes combinaciones se verá si esta mejora se hace factible o es fruto de la casualidad.

Combinación Rh-Ag-Cu

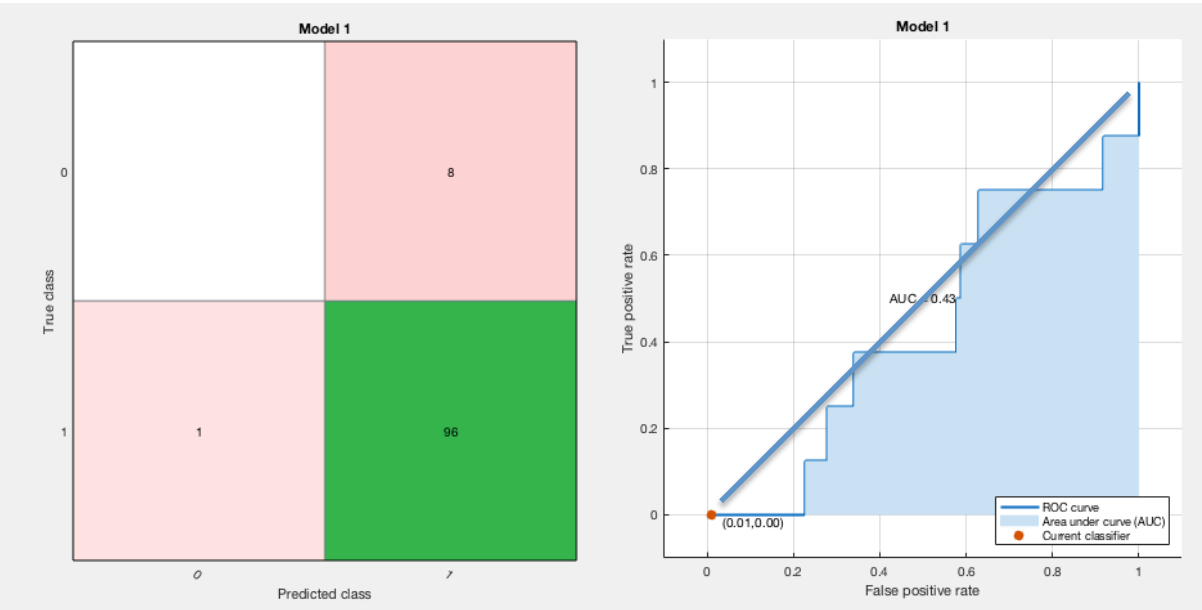

Figura 47: Matriz de confusión y curva ROC para la combinación 257

Se siguen manteniendo los errores de las primeras combinaciones, los resultados tan elevados estaban condicionados porque la mayoría de las clases eran 1. Como dato positivo parece que el valor del área bajo la curva mantiene su tendencia de aumentar en relación al caso de 2 electrodos. También hay que destacar que ha disminuido un poco, pero esto se achaca a la naturaleza de los electrodos, ya que en este caso tampoco se encuentra el electrodo de cobalto, que parece ser el que mejor funciona junto con el electrodo de plata.

Combinación Ag-Co-Cu

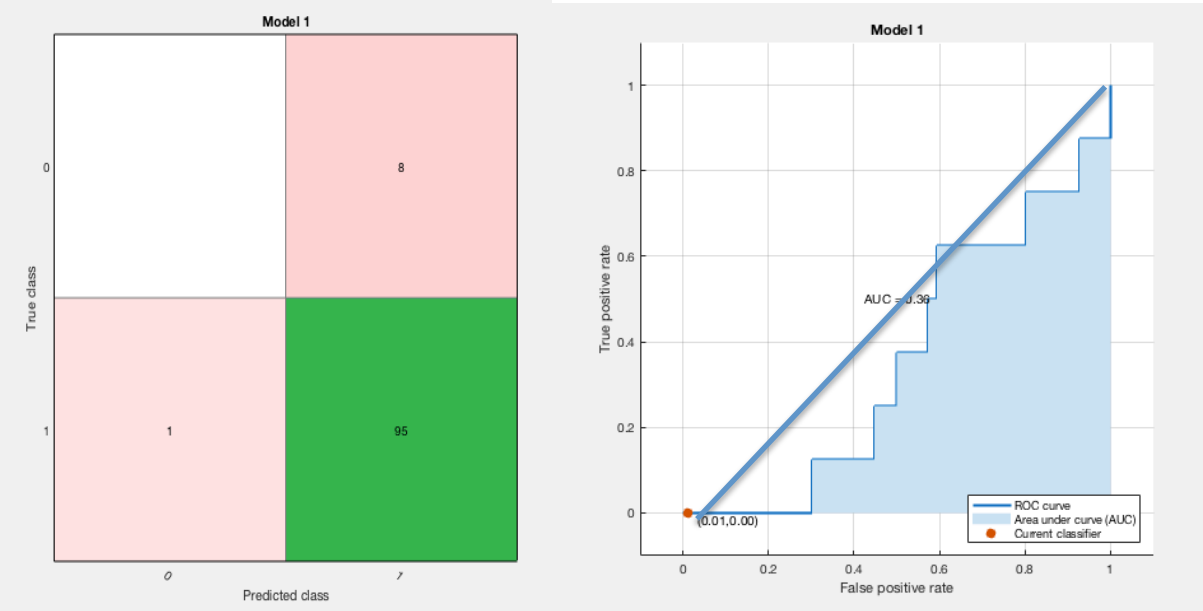

Figura 48: Matriz de confusión y curva ROC para la combinación 567

Para la última combinación de 3 electrodos el resultado no ha sido diferente, se ha producido un sesgo como en los casos anteriores y el sistema no puede ser válido. Además, el valor del área bajo la curva en la curva ROC ha vuelto a disminuir a valores en los que se encontraba para 2 electrodos. Esto indica que estos tres electrodos, a priori los tres mejores por separado, no obtienen el mejor resultado al utilizarse juntos, por lo que cabe la posibilidad de que electrodos con menor porcentaje de importancia en un principio se combinen mejor y obtengan mejores resultados que los electrodos categorizados anteriormente como potencialmente mejores.

Como conclusión, el aumentar a 3 el número de electrodos no ha supuesto una gran mejoría en los resultados, si es cierto que se ha mejorado en los valores de área bajo la curva, pero todavía se sigue produciendo un claro sesgo hacia la clase 1 que impide una correcta clasificación. Como el número de electrodos aún es pequeño y parece ser que se va mejorando, se supone que el bajo número de electrodos era un problema, por lo que falta comprobar si la falta de variedad en las clases lo es también. Para comprobarlo se analizan los datos para las muestras con un mayor número de electrodos.

Combinación Rh-Pt-Ag-Cu

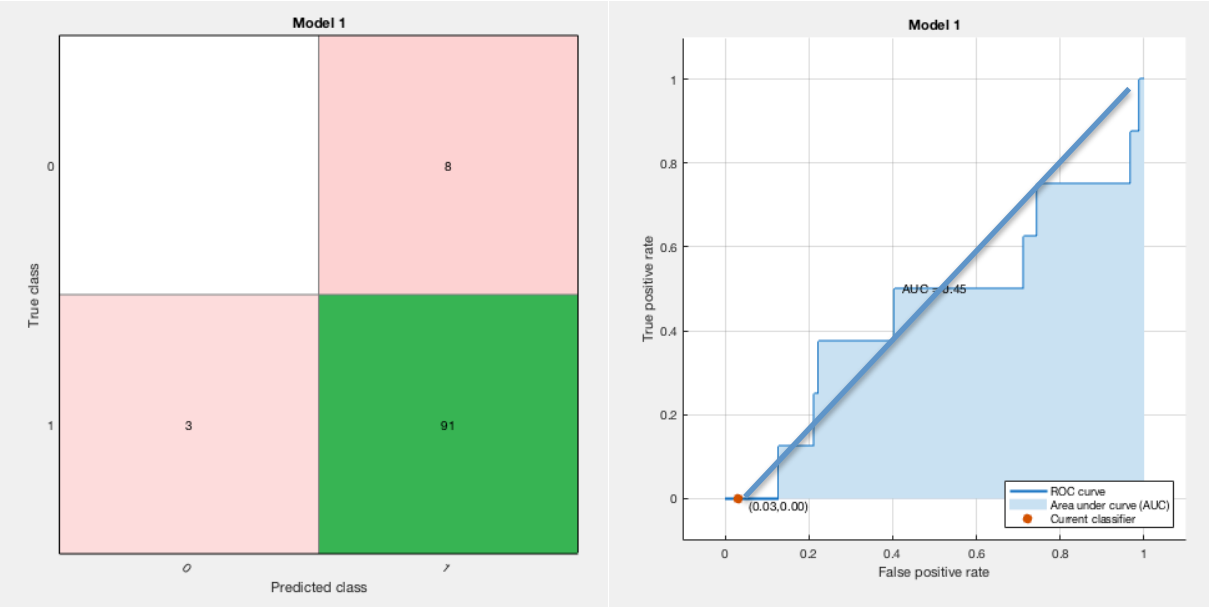

Figura 49: Matriz de confusión y curva ROC para la combinación 2357

Ya con combinaciones de cuatro electrodos se va aumentando el número de variables por cada muestra, y parece ser que aunque se sigue produciendo el sesgo, el sistema comienza a mejorar. Puede verse reflejado en un mayor número de predicciones de la clase 0, aunque estas sean erróneas, por lo que realmente los resultados no han mejorado en gran cantidad. El sistema sigue sin poder predecir ninguna muestra de clase 0 correctamente. El valor de AUC de la curva ROC se mantiene en los valores obtenidos para 3 electrodos, por lo que habrá que ver si con otras combinaciones se consigue seguir mejorándolo.

Combinación Rh-Ag-Co-Cu

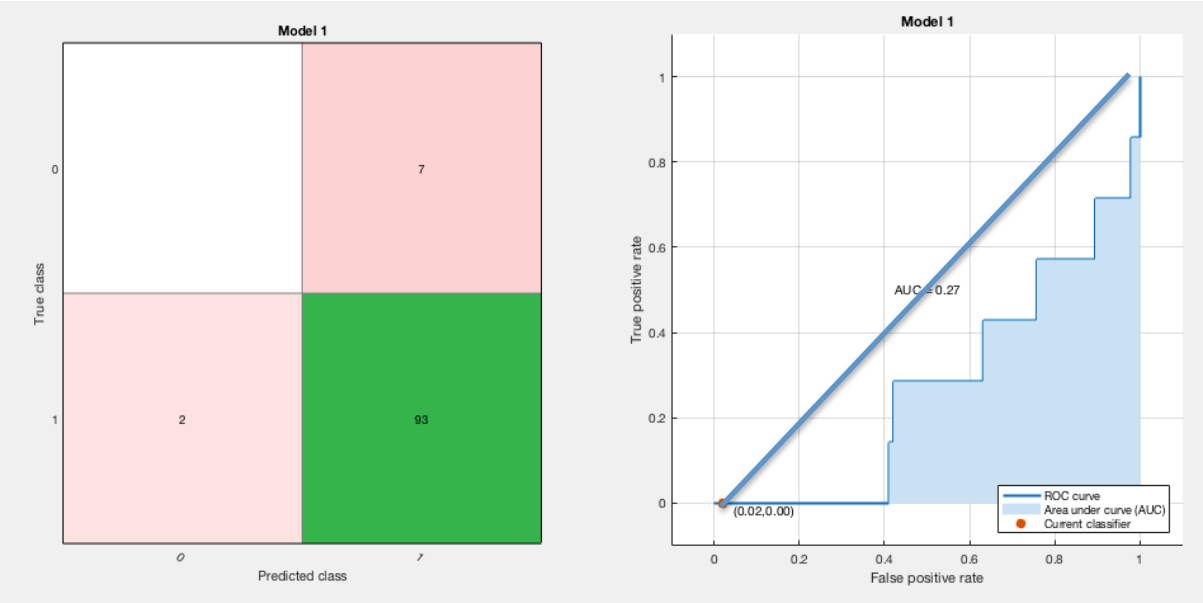

Figura 50: Matriz de confusión y curva ROC para la combinación 2567

Para esta combinación se vuelve a predecir un mayor número de muestras como clase 0, pero como en el caso anterior, el sistema es incapaz de predecir ninguna correctamente. Aunque el sesgo parece que va disminuyendo, los resultados siguen siendo muy malos para lo esperado a priori por los valores obtenidos de predicción facilitados por el clasificador. En cuanto a la curva ROC, en este caso su valor de área bajo la curva es el más bajo hasta el momento, esto es extraño ya que con un mayor número de electrodos este valor tenía una tendencia de mejora, ya que para las demás combinaciones de 4 electrodos no ocurre, se supondrá causa de la combinación, por lo que parece ser que alguno de estos electrodos no trabaja bien con los demás.

Combinación Pt-Ag-Co-Cu

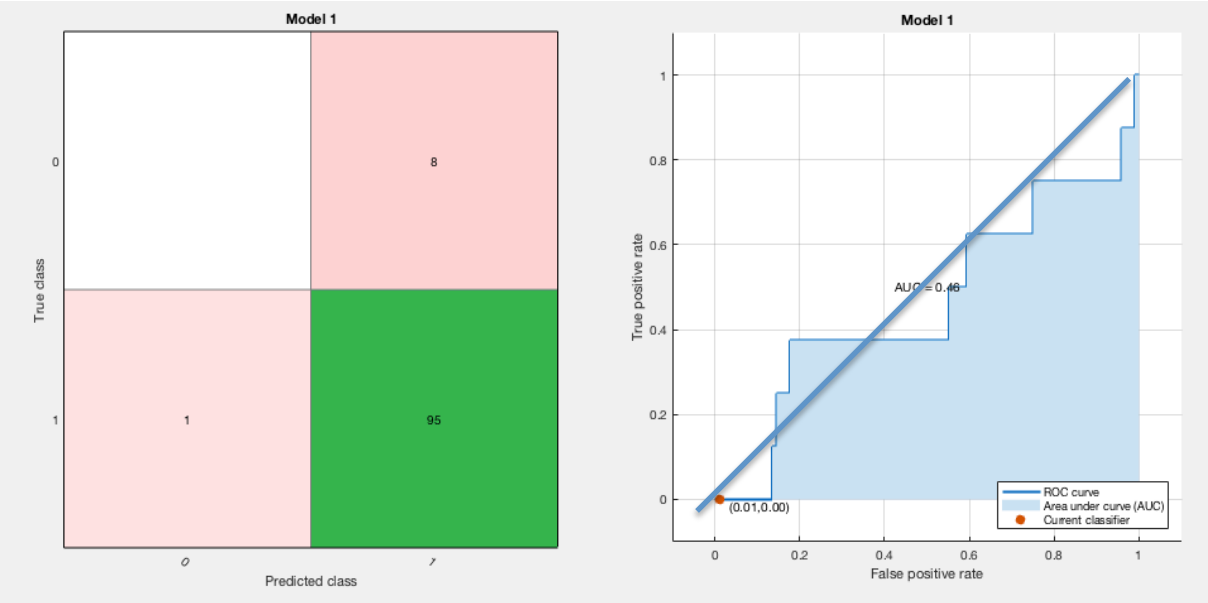

Figura 51: Matriz de confusión y curva ROC para la combinación 3567

Con la última combinación de cuatro electrodos no mejoran las cosas, se sigue produciendo un sesgo. Es más, el sesgo vuelve a pronunciarse al no predecir nada más que una muestra como clase 0. Este sistema sigue sin ser válido para la continuación del trabajo. En cuanto al valor del área bajo la curva se ha remontado un poco en relación a la combinación anterior, por lo que se puede entender que hay dos electrodos que no obtienen muy buenos resultados juntos.

Una vez finalizadas las muestras para las combinaciones de 4 electrodos se ha comprobado que todavía se está produciendo el sesgo que no deja clasificar al sistema de forma correcta. En este caso parece ser que el aumento de un electrodo no está suponiendo una mejora, ya que como mucho se están manteniendo los resultados obtenidos para 3 electrodos. En cuanto a las combinaciones, se sospecha que los electrodos de cobalto y cobre no funcionan del todo bien juntos, en los siguientes apartados se comprobará si esto es real o una simple casualidad.

Combinación Rh-Pt-Ag-Co-Cu

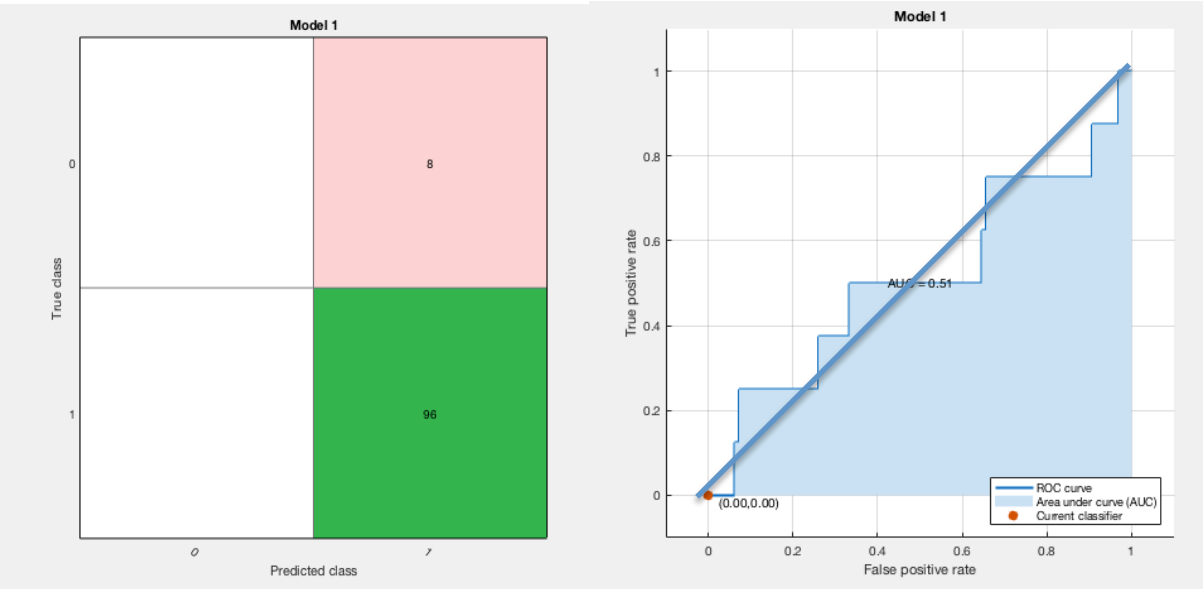

Figura 52: Matriz de confusión y curva ROC para la combinación 23567

A pesar de que no se ha predicho en ninguna ocasión una clase 0, los datos de la curva ROC vuelven a mejor al haber aumentado un electrodo más en los análisis. Si esta tendencia sigue produciéndose es posible que finalmente se consiga eliminar el sesgo. La otra posibilidad es que el sesgo venga producido por una falta de muestras de clases 0, pero por los resultados de precisión de las combinaciones con un mayor número (que decaen), parece ser que el sesgo deja de producirse, por lo que el problema también viene de el número de electrodos.

Combinación Rh-Au-Ag-Co-Cu

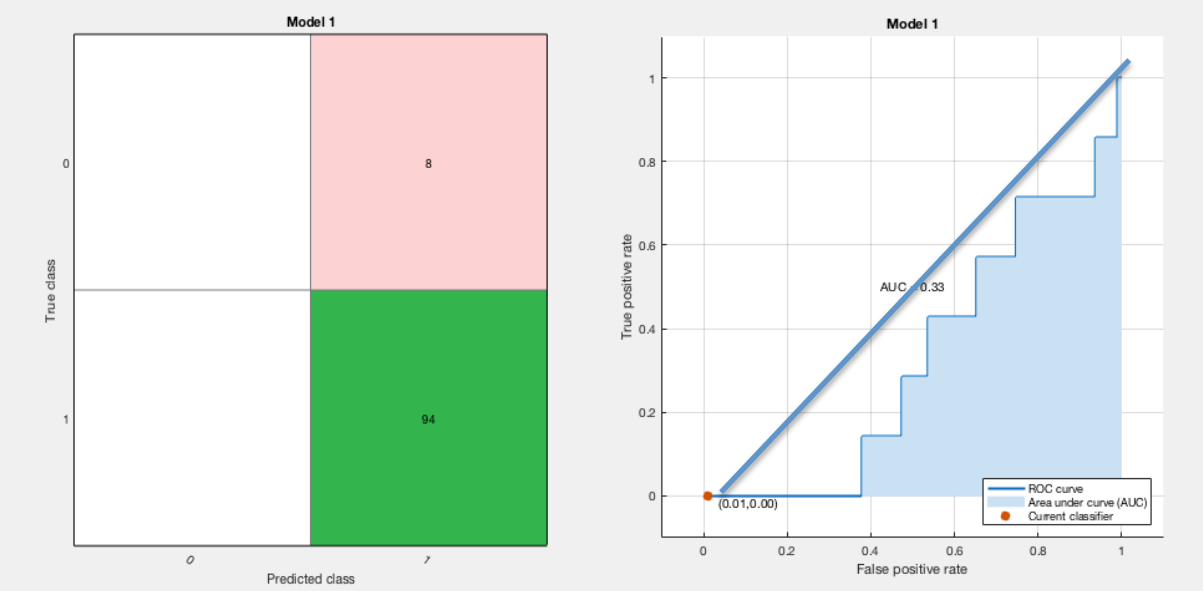

Figura 53: Matriz de confusión y curva ROC para la combinación 24567

En esta combinación vuelven los resultados negativos, no ha sido capaz de predecir en ninguna ocasión la clase 0, ni bien ni mal, simplemente el sistema no realiza ninguna predicción, todo lo predice como clase 1. Además el área de la curva ROC ha disminuido, por lo que se entiende que la inclusión del electrodo de oro no mejora los datos aportados por los electrodos que ya estaban anteriormente.

Combinación Pt-Au-Ag-Co-Cu

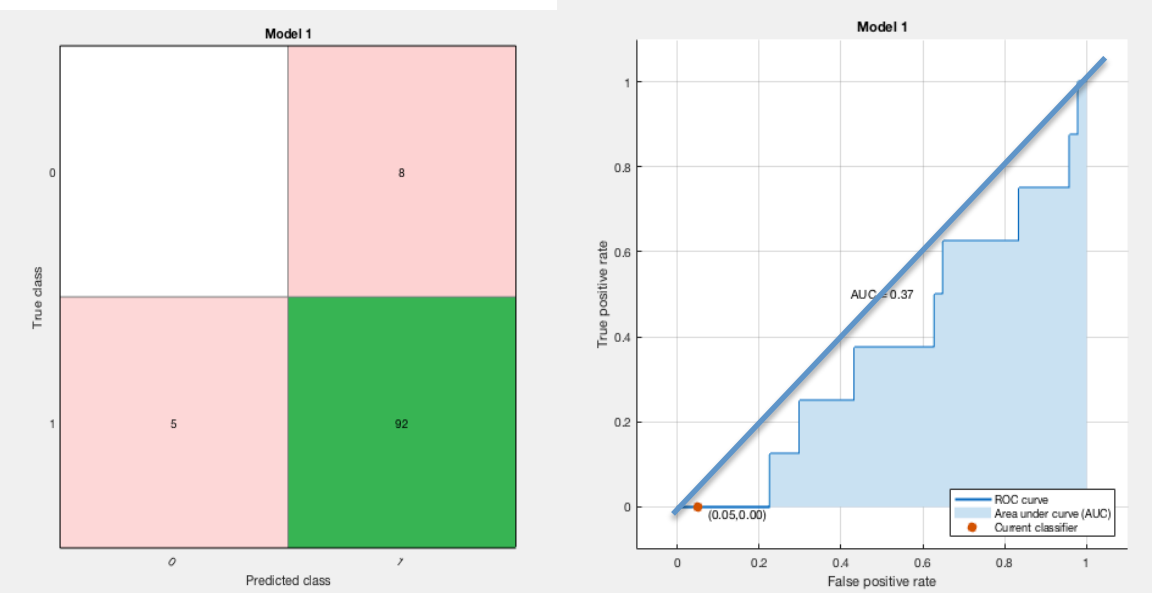

Figura 54: Matriz de confusión y curva ROC para la combinación 34567

Para la última combinación de 5 electrodos se ha logrado predecir clases 0 de nuevo, pero en ningún caso de un modo correcto. Sigue existiendo algún problema que impide que el sistema clasifique de forma correcta las muestras de clase 0. Además, los valores de la curva ROC oscilan para cada combinación, mejorando en algunas ocasiones para luego obtenerse resultados pésimos en otras. En este caso puede comprobarse, como si se atiende a la matriz de confusión puede pensarse sobre una mejoría ya que intenta predecir la clase no sesgada, pero al analizar el valor del área bajo la curva este indica que los resultados de este sistema son pésimos

Una vez terminadas las combinaciones de 5 electrodos no se ha observado ninguna mejoría real desde las combinaciones de 3, por lo que parece ser que estos nuevos electrodos introducidos no están aportando ninguna mejoría. Aún así todavía no se ha trabajado con ninguna combinación que contenga los electrodos de Iridio o Níquel, por lo que puede ser que en el siguiente apartado estos electrodos aporten buenos resultados, o por otro lado, que estos se mantengan igual o incluso empeoren. En base a los resultados previos, debería dejar de producirse el sesgo una vez se llega a los 6 electrodos, ya que los valores de precisión decaen al 80 o 70%. Esto indica que se están prediciendo un mayor número de clases 0 y esto aumenta las posibilidades de que el sistema comience a clasificarlas correctamente.

Combinación Rh-Pt-Au-Ag-Co-Cu

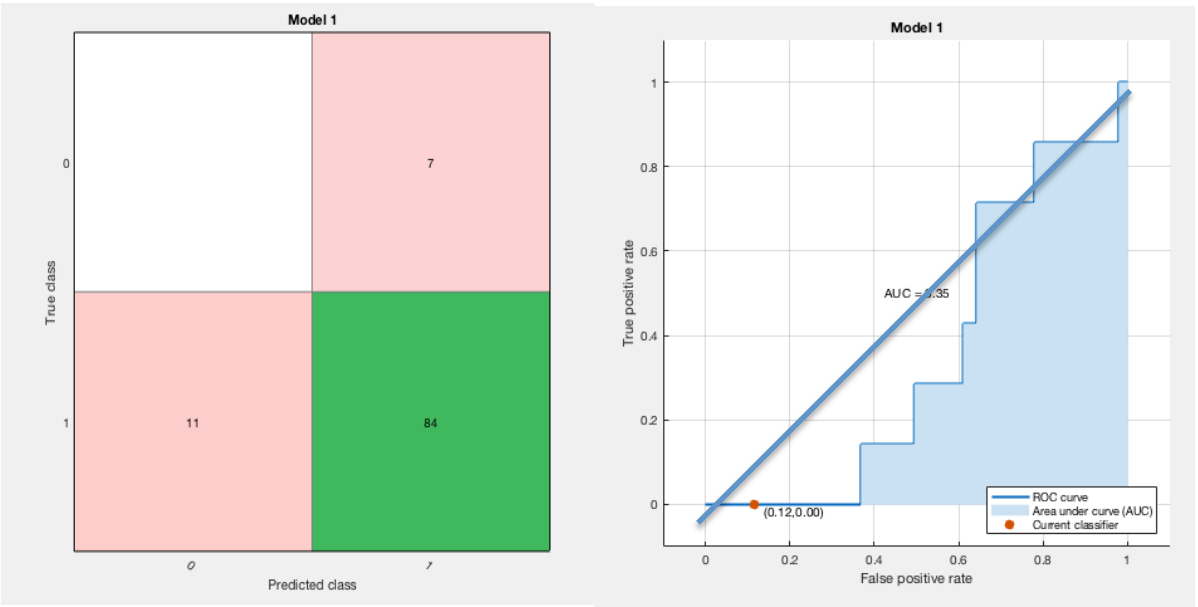

Figura 55: Matriz de confusión y curva ROC para la combinación 234567

Como se preveía se ha producido un mayor número de predicciones de la clase 0 que en los cases anteriores, por lo que el sesgo sigue produciéndose, pero parece ser que está remitiendo. El resultado de área bajo la curva de la curva ROC es muy bajo, pero esto es debido a que se han producido un gran número de errores al clasificar clases 0 de manera errónea. Esto se ve reflejado en la figura, ya que en la zona de falsos positivos el área de la curva aumenta, mientras que para valores bajos de la zona de falsos positivos es prácticamente nulo.

Combinación Rh-Pt-Ag-Co-Cu-Ni

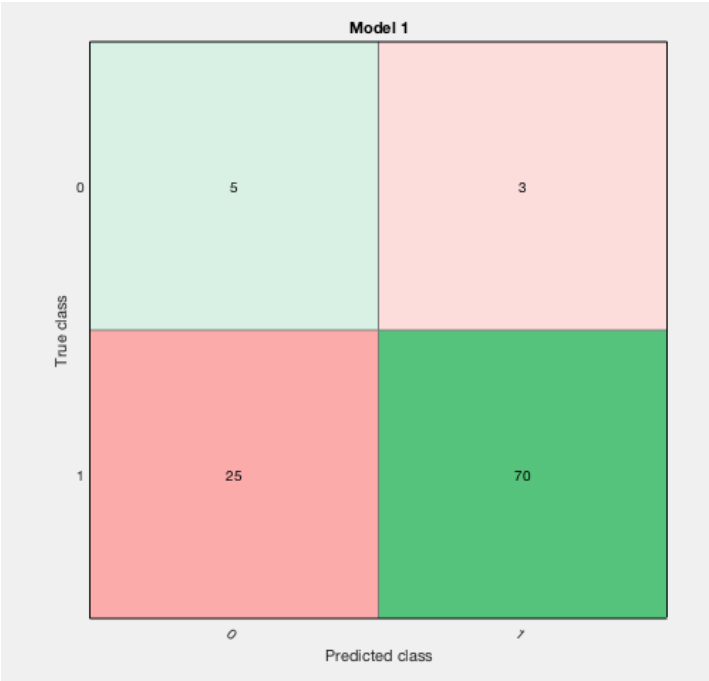

Figura 56: Matriz de confusión para la combinación 235678 (1)

Para esta combinación por fin se consigue eliminar el sesgo hacia la clase 1. El sistema ya es capaz de predecir las dos clases. Para la clase 0, que sigue siendo la clase conflictiva, se han predicho con acierto 5 muestras de 8 existentes, pero para predecir de forma correcta estas 5 muestras se ha errado en 25. Para la clase 1 se predice con acierto la mayoría de las veces, pero esto puede ser debido a que la mayoría de muestras son de esta clase. Para valorarlo, es necesario examinar las matrices de confusión desglosadas:

Figura 57: Matriz de confusión para la combinación 235678 (2)

De esta matriz se puede extraer que un 62% de las muestras de clase "0" son predichas con acierto mientras que un 74% de las muestras de clase "1" también lo son. Estos valores comienzan a acercarse a los resultados que se esperan con estos tipos de análisis, pero todavía se deben optimizar en gran medida para aumentar estos porcentajes. De todos modos, con 6 electrodos ya se ha conseguido no solo que el sistema pueda predecir alguna clase 0 correctamente, sino que el resultado no ha sido muy bajo. Esto indica que esta combinación de electrodos comienza a trabajar de forma correcta, cosa que todavía no se había conseguido sin que se produjese un sesgo. Esto puede ser debido o al mayor tamaño de la matriz de datos o a la inclusión del electrodo de Níquel, que no había sido utilizado hasta el momento y parece aportar muy buenos resultados al sistema.

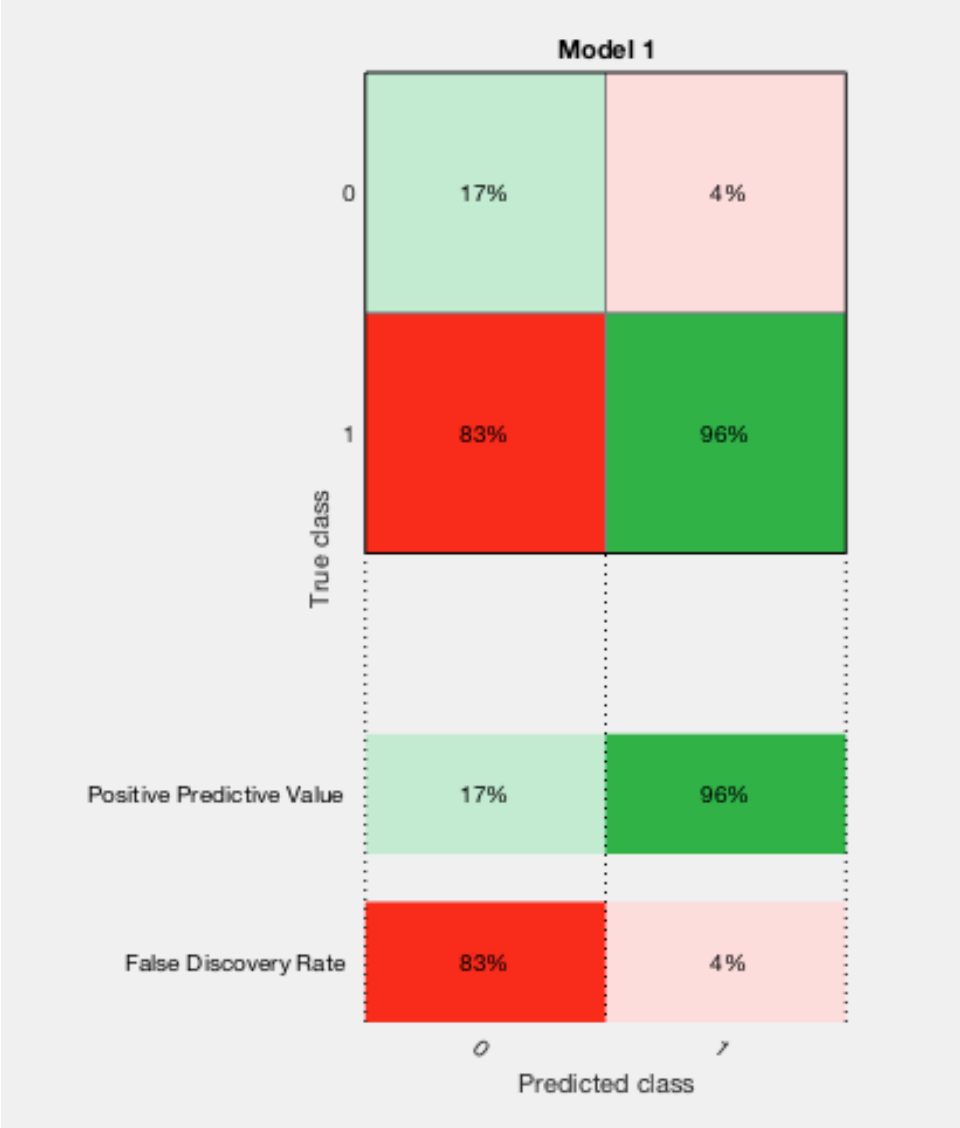

Figura 58: Matriz de confusión para la combinación 235678 (3)

De esta matriz desglosada se extraen resultados peores, de todas las muestras predichas como clase "0", únicamente se ha acertado en un 17% de las ocasiones. En contraposición, de todas las muestras que se han predicho como clase "1" se ha acertado el 96% de las veces. Por un lado se tiene un ratio muy malo de predicción para una clase y por otro uno muy bueno para la otra. Esto es lógico, ya que la clase 0, a pesar de haber comenzado a ser predicha, sigue causando problemas en su clasificación. Estos problemas se suponen debidos a una falta de muestras de clase 0 en la matriz, por lo que el sistema no posee información suficiente para poder discriminar esta clase de una muestra de clase 1.

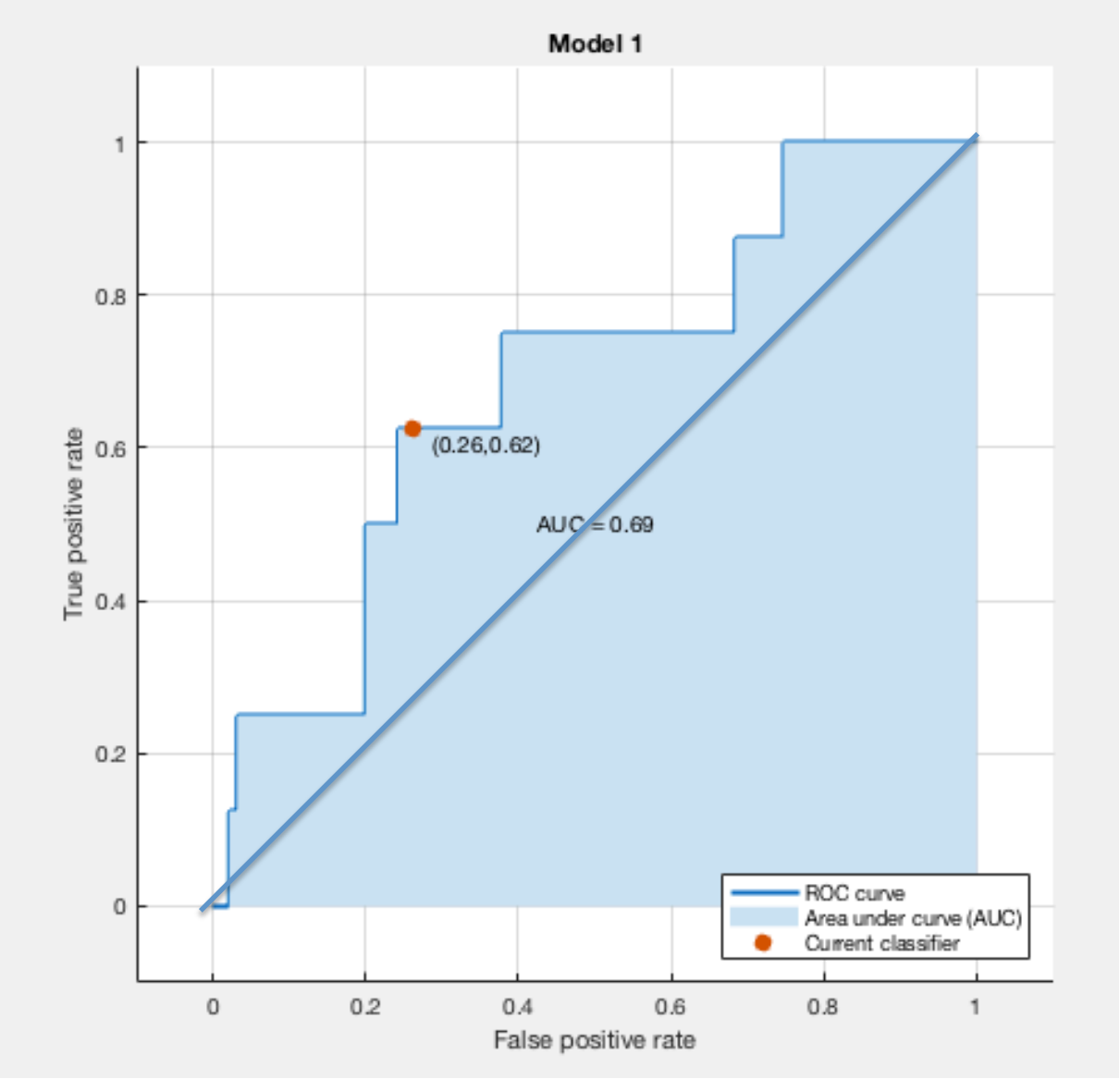

Figura 59: Curva ROC de la combinación 235678

En este caso se obtiene una curva ROC mucho mejor que las obtenidas anteriormente, ya se supera completamente la diagonal y su valor de área bajo la curva es de 0,69. Según la clasificación comentada al comienzo esto corresponde a un sistema de clasificación regular, pero acercándose ya a valores próximos a un clasificador bueno, por lo que esta combinación se considera de momento la mejor opción para el análisis de estas matrices. El aumento del número de electrodos a 6 ha dado sus frutos, en la siguiente combinación se comprobará si estos siguen para otras combinaciones y si el electrodo de Níquel puede tener algo que ver en esta mejoría.

Combinación Rh-Au-Ag-Co-Cu-Ni

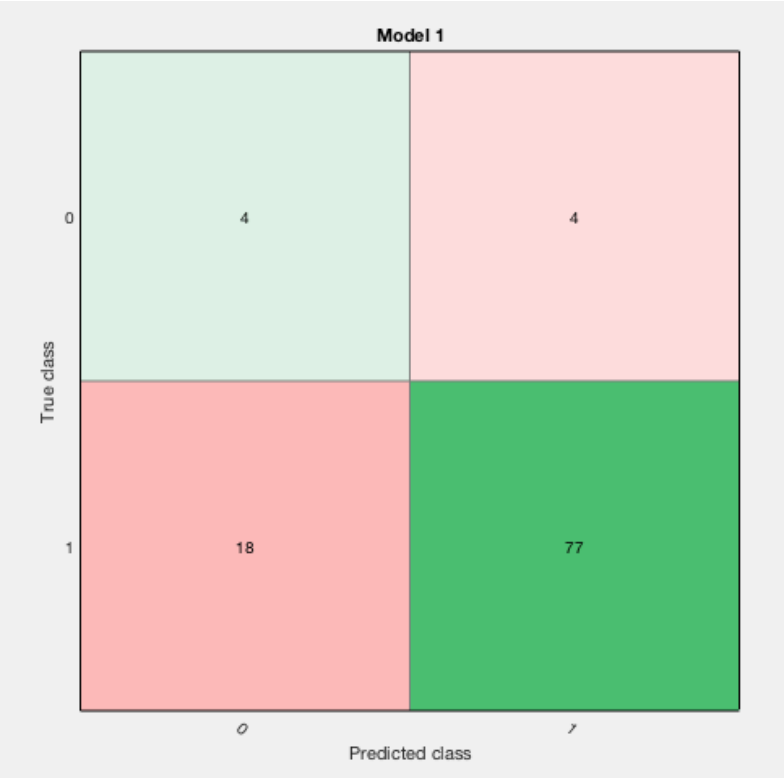

Figura 60: Matriz de confusión de la combinación 245678 (1)

En esta combinación se comprueba la tendencia de mejoría mostrada en la matriz anterior. Se comprueba que a partir de 6 electrodos comienza a dejar de producirse el sesgo hacia la clase 1 y el sistema comienza a clasificar. Con esto y los resultados anteriores, se puede considerar seis como el número óptimo de electrodos a utilizar para las medidas, ya que si se aumenta el número se dejará de producir el sesgo con total seguridad, pero los resultados del porcentaje de precisión disminuyen. A continuación se analizan las demás matrices de confusión para comprobar como de bueno es este sistema.

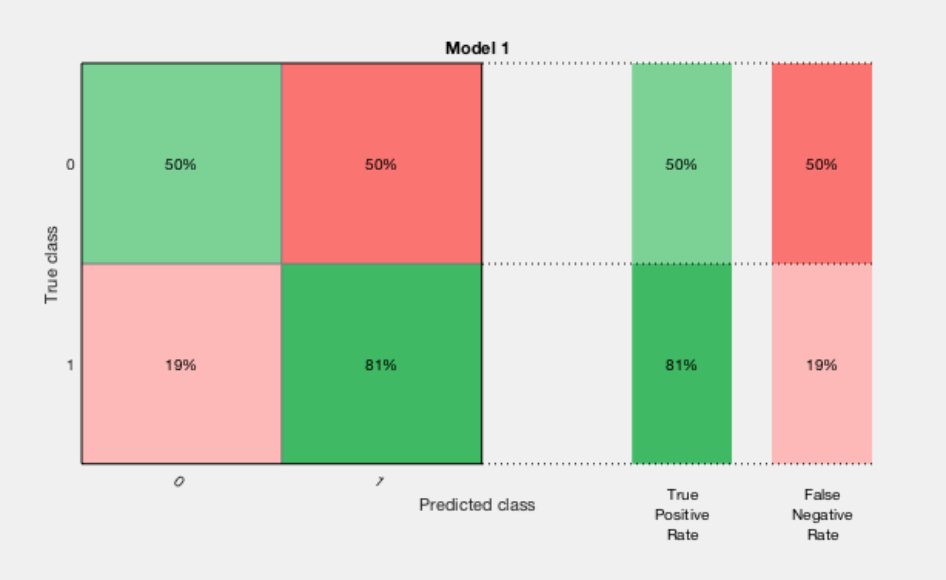

Figura 61: Matriz de confusión para la combinación 245678 (2)

De acuerdo con esta matriz, el sistema entrenado es capaz de predecir correctamente el 81% de las muestras de clase 1 y el 50% de las muestras de clase 0. Sigue existiendo un problema con las muestras de clase 0, como era de esperar por el bajo número de estas en la matriz. Lo importante es que los resultados obtenidos en la anterior combinación se mantienen, por lo que se afianza el número de electrodos como el óptimo para la clasificación y el electrodo de Níquel como necesario.

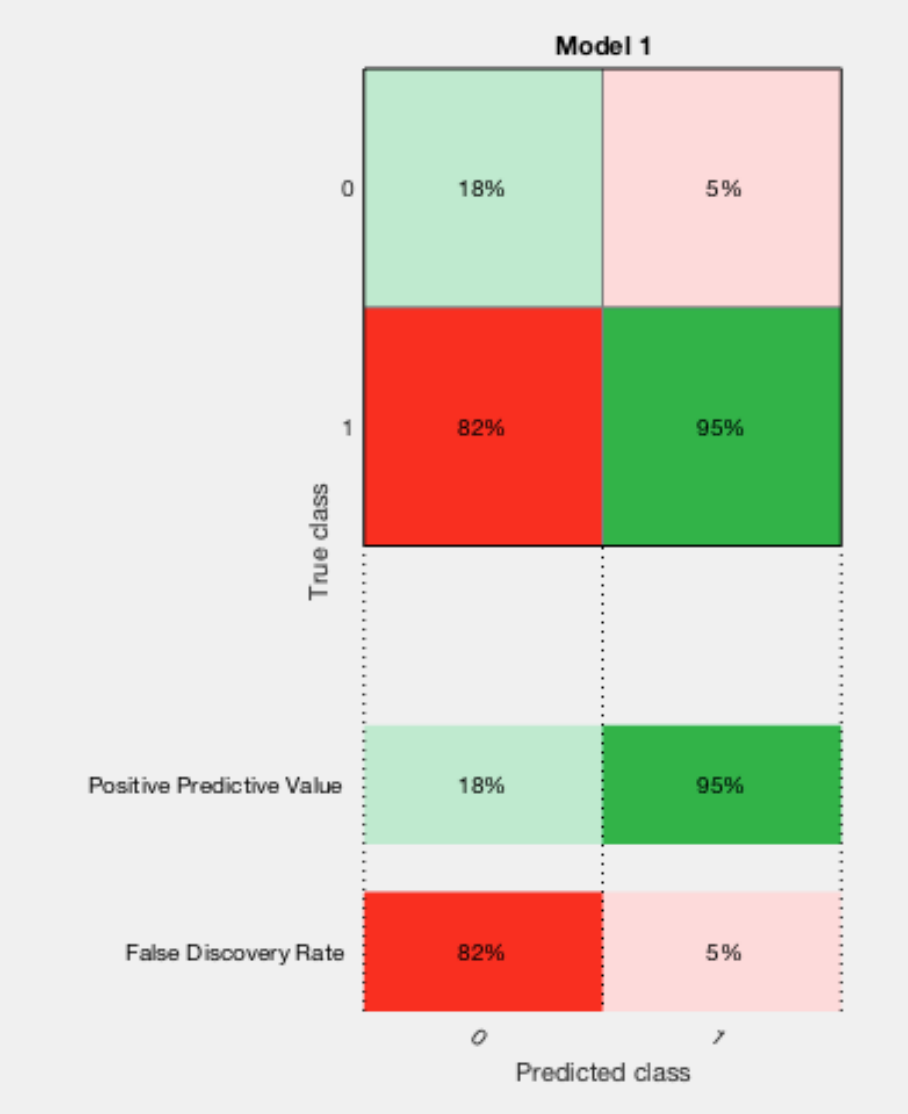

Figura 62: Matriz de confusión de la combinación 245678 (3)

En este caso la matriz de confusión indica que un 95% de las muestras predichas como clase 1 eran realmente de esa clase, este porcentaje es muy bueno, por lo que parece que el sistema para la clasificación de las muestras de clase 1 no tiene ningún problema. Por otro lado, solo un 18% de las clases predichas como clase 0 resultaron luego ser correctas, por lo que la situación es análoga a la anterior, existe un problema para clasificar las muestras de clase 0. Este problema podría solucionarse con la adición de más muestras de este tipo de clase. A continuación se analiza la curva ROC para comprobar el rendimiento de este sistema de clasificación.

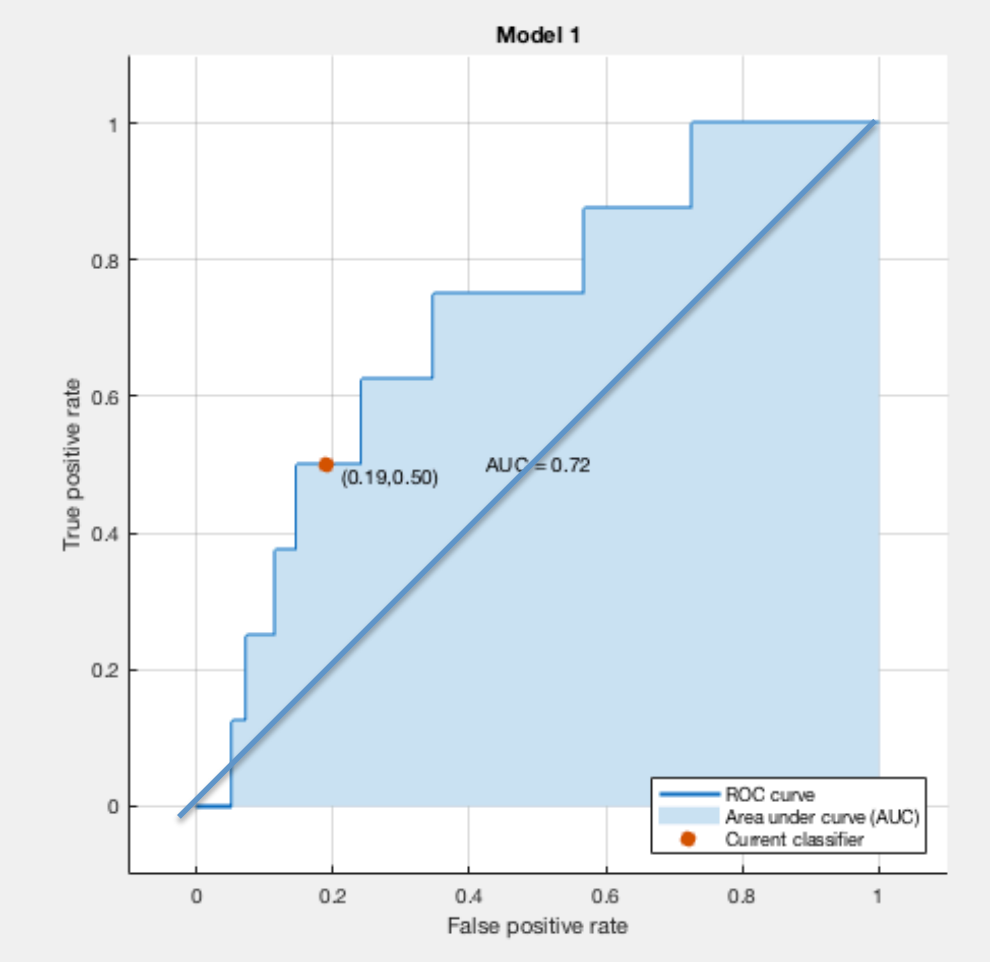

Figura 63: Curva ROC de la combinación 245678

La curva ROC indica que los resultados de esta prueba son bastante buenos. El valor del área bajo la curva es de 0,72, el más elevado conseguido hasta este momento. Se considera un sistema de clasificación regular, pero está al límite de un sistema de clasificación bueno. La mejoría ha quedado confirmada con estas dos combinaciones de 6 electrodos, sobre todo en esta última. La diferencia entre las dos últimas combinaciones es que en la primera se ha utilizado el electrodo de platino y en esta última se ha sustituido el de platino por el de oro. Aunque los resultados son parecidos, parece ser que esta combinación de 5 electrodos funciona mejor con el electrodo de oro que con el de platino.

8.2 CONCLUSIONES

- A pesar de que previamente los resultados obtenidos para los análisis discriminantes cuadráticos con PCA pareciesen muy buenos, la gran mayoría estaban afectados por un sesgo y sus valores de precisión eran erróneos.

- Este sesgo deja de producirse para las combinaciones de 6 electrodos, por lo que combinaciones menores con este tipo de análisis no son viables sin que se produzca un error.

- Mediante las combinaciones de 6 electrodos pueden obtenerse resultados prometedores, comparándolos con los resultados para las demás combinaciones, se concluye que el número ideal de electrodos para estas muestras es 6.

- El mayor problema actual es la predicción de las clases "0", las cuales el programa no consigue predecir correctamente. Esto se cree debido a una falta de cantidad de este tipo de muestras, por lo que podría solucionarse añadiendo más muestras de este tipo a la matriz de datos.

- El electrodo de Níquel parece ser el electrodo fundamental que hace que no se produzca la situación de sesgo. A pesar que este electrodo en pequeñas combinaciones no obtiene buenos resultados al encontrarse en una combinación mayor consigue que el sistema clasifique de forma correcta.

9. Prueba de aleatoriedad

Dado que en el apartado anterior se ha comprobado que con una análisis discriminante cuadrático con PCA pueden llegar a obtenerse buenos resultados para ciertas combinaciones de electrodos, en este apartado se intentará comprobar si la posición o el número de las muestras influye en los resultados del aprendizaje y entrenamiento del sistema. Además, con esta prueba se podrá discernir si existen muestras que están generando errores o malas clasificaciones.

Esta prueba se realizará tanto a las matrices de datos iniciales como a las matrices reducidas, para tener una base en la cual poder guiarnos con mayor facilidad una vez se realicen a las matrices reducidas y poder comparar los resultados obtenidos, los cuales, a priori, deben ser mejores para las matrices reducidas.

9.1 RESULTADOS PARA LA MATRIZ NUEVA "Xnew"

En primer lugar se trabajará con "Xnew", para ello, se crea una nueva matriz compuesta por las muestras y los valores de la matriz "Xnew" pero desordenando las muestras aleatoriamente. Una vez el orden de las muestras está aleatorizado se dividen en secciones de 20 muestras cada una. A cada una de estas secciones se le realiza un análisis discriminante cuadrático con análisis de componentes principales, obteniéndose de este modo una precisión en la predicción dada por el programa de:

Tabla 27: Resultados prueba aleatoriedad "Xnew" (1)

Paquete	1	2	3	4	5	6	7	8
Precisión	0,60	0,30	0,55	0,50	0,55	0,50	0,40	0,55

Los valores obtenidos son parecidos a los conseguidos en el primer apartado, donde ya se realizó este análisis a las matriz completa, pero ahora se ha comprobado que dependiendo de las muestras que se utilicen el rendimiento de la clasificación puede doblarse o reducirse a la mitad.

A parte de calcular la precisión, con el sistema de análisis creado para cada paquete de datos, se han predicho las clases que corresponderían a las muestras de la propia matriz "Xnew" sin desordenar. Es decir, se han utilizado los sistemas creados a partir de los paquetes para predecir las clases propias de la matriz de la que provienen los paquetes. Con esto, se obtiene una matriz de 8 columnas, donde cada columna contiene las clases predichas para la matriz según el sistema de cada paquete. Esta matriz se ha denominado "resultados_new".

La finalidad de realizar esta matriz es poder comparar ahora la precisión de los sistemas creados, no con las muestras que el propio programa conoce si no con las muestras ordenadas que nosotros conocemos. Para realizarlo, se compara cada columna de la matriz "resultados_new" con las clases de la matriz "Xnew", obteniéndose como resultado:

Tabla 28: Resultados prueba aleatoriedad "Xnew" (2)

Paquete	1	2	3	4	5	6	7	8
Precisión	0,44	0,54	0,57	0,53	0,55	0,49	0,60	0,57

De acuerdo con estos resultados, se difiere de la precisión dada en un principio por la aplicación de entrenamiento, sobre todo en los valores pico, en los valores intermedios si se comprueba que la precisión apenas varía y, como en el primer caso, se encuentra alrededor del 50% en la mayoría de los casos. Esto puede indicar que los resultados pico obtenidos son fruto del azar, ya que al seleccionar paquetes de muestras pequeños cabe la posibilidad de que esas muestras se clasifiquen de mejor o peor modo.

9.2 RESULTADOS PARA LA MATRIZ INICIAL "Xupv"

En segundo lugar, se realizará el mismo análisis a la variable "Xupv", es decir, se partirán las muestras de la matriz en paquetes de 20 aleatoriamente, y cada paquete servirá para entrenar un sistema de análisis discriminante cuadrático con PCA, obteniéndose las siguientes precisiones en la validación:

Tabla 29: Resultados prueba aleatoriedad "Xupv" (1)

Paquete	1	2	3	4	5	6	7	8
Precisión	0,70	0,55	0,50	0,55	0,50	0,30	0,70	0,55

Como en el caso anterior, estos resultados parecen indicar que el orden o la existencia o no de ciertas muestras en el entrenamiento es influyente a la hora de la obtención del clasificador. Para comprobar cuan veraces son estos resultados, se aplicará el mismo procedimiento que en el caso anterior, pero en este caso se intentarán predecir las clases de la matriz "Xupv" con cada sistema creado para cada paquete de muestras. De igual modo se crea la matriz de resultados y se comparan las clases predichas con las clases reales, obteniéndose la siguiente tabla:

Tabla 30: Resultados prueba aleatoriedad "Xupv" (2)

Paquete	1	2	3	4	5	6	7	8
Precisión	0,59	0,54	0,46	0,58	0,56	0,54	0,50	0,57

Tal cual ocurrió en esta prueba para "Xnew", los valores de la precisión tienden a acercarse al 50% cuando son validados con las clases externas, algunos valores suben y otros bajan, pero los valores más alejados de la media siempre tienden a estabilizarse, por lo que se entiende que estos valores atípicos que aumentan o empeoran en porcentaje son fruto de la casualidad al coger muestras aleatorias.

Tras estas pruebas se concluye que en las matrices iniciales ni la posición de las muestras ni el tamaño de las matrices es influyente a la hora de entrenar el sistema de clasificación. Este análisis sirve de punto de partida para la realización de esta prueba a las matrices sin las clases "CTRLpost", que podría llevar a engaño como se ha comprobado en las primeras pruebas donde se obtuvieron valores más extremos, que luego realmente no eran ciertos.

Una vez se ha comprobado el funcionamiento del método y las confusiones que este puede dar, como para algún paquete de datos un valor más elevado de lo que en realidad es, se realiza la prueba a las variables reducidas, es decir, habiendo eliminado las muestras con clase "CTRLpost".

9.3 RESULTADOS PARA LA MATRIZ NUEVA REDUCIDA "Xnew_reducida"

Se comienza con la matriz "Xnew_reducida", en este caso, y debido a la poca cantidad de clases "CTRL" solo se puede dividir la matriz en dos paquetes de muestras, ya que el sistema necesita al menos 3 muestras de clase "CTRL" para poder realizar el análisis, por lo que si se realizan más divisiones de paquetes, no todos alcanzarán esta condición, y el programa no puede entrenar el sistema y se produce un error. Por lo demás el procedimiento es el mismo, se reordenan las muestras aleatoriamente y se crean dos paquetes, para cada uno de estos se genera un sistema de análisis discriminante cuadrático con PCA, obteniéndose las siguientes precisiones:

- Paquete 1: 67,92 %
- Paquete 2: 84,91 %

Igualmente, con cada paquete se predicen las clases de la matriz reducida "Xnew_reducida", y como en los apartados anteriores se comparan las predicciones con las clases reales, dando como resultado:

- Paquete 1: 50,00%
- Paquete 2: 47,65%

Según estos resultados los valores obtenidos de precisión mediante la aplicación de entrenamiento son erróneos, esto puede ser debido a que al haber pocas clases de "CTRL" el sistema al utilizar PCA y validación cruzada clasifique como "BC" casi todas las muestras, pero al analizar muestras de otra matriz cometa errores. Lo positivo de estos resultados es que se comprueba una duda creada en apartados anteriores, sobre si se estaba realizando un sesgo hacia la clase "BC". Tras estos datos se comprueba que en este caso no es así, ya que el porcentaje es demasiado bajo, si se hubiese producido el sesgo la mayoría de muestras habría sido predichas con éxito ya que la mayoría son de clase "BC".

9.4 RESULTADOS PARA LA MATRIZ INICIAL REDUCIDA "Xupv_reducida"

En último lugar se realizará la prueba a la matriz "Xupv_reducida", los pasos son exactamente iguales que para "Xnew_reducida", por la falta de clases "CTRL" solo es posible crear dos paquetes de muestras aleatorias. Estos dos paquetes son entrenados como en los casos anteriores para poder obtener las siguientes precisiones:

- Paquete 1: 49,06%
- Paquete 2: 90,57%

Según estos resultados los porcentajes difieren en gran cantidad, por lo que para cerciorarse de cual es el correcto la mejor solución es avanzar con la prueba y comparar los valores predichos para "Xupv_reducida" por los dos paquetes con los valores reales. Tras la comparación, se ha obtenido:

- Paquete 1: 43,53%
- Paquete 2: 57,06%

Se comprueba de este modo que el resultado real es el de una precisión cercana al 50%, como en los casos anteriores. Con esta prueba se demuestra que la aleatoriedad de las muestras no influye a la hora de entrenar el sistema, así como tampoco el tamaño de los paquetes de muestra. Además, se ha descubierto que el problema no son las clases "CTRLpost", ya que tras su eliminación los resultados obtenidos siguen siendo prácticamente los mismos. Se deduce que el problema podría ser de los electrodos, por esto, se realiza la siguiente prueba a las mejores combinaciones de electrodos, para la cual se han obtenido los mejores resultados, y que todavía esta por ver si estos resultados son reales o fruto de errores y la casualidad.

10. PRUEBA CON LAS CLASES "CTRLpost" A LOS MEJORES RESULTADOS

Tras lo acontecido en el apartado anterior, donde se ha comprobado que los buenos resultados obtenidos al eliminar las clases CTRLpost era debido a que se producía un sesgo hacia la clase mayoritaria, debe comprobarse si esto ocurre de igual manera para los sistemas propuestos con las mejores combinaciones de electrodos, para los cuales se han obtenido unos resultados satisfactorios, pero que según las curvas ROC y las matrices de confusión parece ser que se está cometiendo un sesgo. Para terminar de comprobarlo, se utilizarán estos sistemas entrenados para predecir las clases "CTRLpost" eliminadas para el entrenamiento y ver si el sistema es capaz de clasificarlas en clases diferentes o simplemente reconoce todas como clase "1".

10.1 RESULTADOS

Combinaciones de 2 electrodos

Tabla 31: Predicción de las muestras "CTRLpost" con las combinaciones de dos electrodos

Combinaciones de 2 electrodos						
Combinación	Rh-Ag		Ag-Co		Ag-Cu	
Clase	BC	CTRL	BC	CTRL	BT	CTRL
Resultado	63	0	63	0	63	0

De acuerdo con estos resultados, se comprueba que la precisión dada por la aplicación es engañosa, ya que se ha producido un sesgo hacia la clase "BC". El porcentaje de éxito en la predicción es elevado porque la mayoría de clases son "BC".

Combinaciones de 3 electrodos

Tabla 32: Predicción de las muestras "CTRLpost" con las combinaciones de 3 electrodos

Combinaciones de 3 electrodos						
Combinación	Rh-Ag-Cu		Rh-Ag-Co		Ag-Co-Cu	
Clase	BC	CTRL	BC	CTRL	BT	CTRL
Resultado	63	0	63	0	63	0

También para combinaciones de 3 electrodos se comprueba que la precesión dada por la aplicación es engañosa, ya que se ha producido un sesgo hacia la clase "BC". El porcentaje de éxito en la predicción es elevado porque la mayoría de clases son "BC".

Combinaciones de 4 electrodos

Tabla 33: Predicciones de las muestras "CTRLpost" con las combinaciones de 4 electrodos

Combinaciones de 4 electrodos						
Combinación	Rh-Pt-Ag-Cu		Rh-Ag-Co-Cu		Pt-Ag-Co-Cu	
Clase	BC	CTRL	BC	CTRL	BT	CTRL
Resultado	63	0	63	0	63	0

Como en los casos anteriores, se comprueba que la precisión dada por la aplicación es engañosa, ya que se ha producido un sesgo hacia la clase "BC". El porcentaje de éxito en la predicción es elevado porque la mayoría de clases son "BC".

Combinaciones de 5 electrodos

Tabla 34: Predicción de las muestras "CTRLpost" con las combinaciones de 5 electrodos

Combinaciones de 5 electrodos						
Combinación	Rh-Pt-Ag-Co-Cu		Rh-Au-Ag-Co-Cu		Pt-Au-Ag-Co-Cu	
Clase	BC	CTRL	BC	CTRL	BT	CTRL
Resultado	63	0	63	0	63	0

Incluso para la combinación de 5 electrodos sigue comprobándose que la precesión dada por la aplicación es engañosa, ya que se ha producido un sesgo hacia la clase "BC". El porcentaje de éxito en la predicción es elevado porque la mayoría de clases son "BC".

Combinaciones de 6 electrodos

Tabla 35: Predicción de las muestras "CTRLpost" con las combinaciones de 6 electrodos

Combinaciones de 6 electrodos						
Combinación	Rh-Pt-Au-Ag-Co-Cu		Rh-Pt-Ag-Co-Cu-Ni		Rh-Au-Ag-Co-Cu-Ni	
Clase	BC	CTRL	BC	CTRL	BT	CTRL
Resultado	63	0	41	22	38	25

Por último, para la combinación de 6 electrodos por fin se obtienen resultados valorables, ya que aunque para la primera combinación se repita el fracaso de los casos anteriores, para las otras dos combinaciones se han obtenido buenos resultados, ya que a pesar de que su porcentaje de precisión sea menor que en otras combinaciones, este es real. Además, las dos combinaciones prácticamente coinciden en la clasificación de la muestra, siendo dos modelos diferentes, por lo que puede considerarse que los resultados obtenidos parecen ser buenos. Aún con esto, el porcentaje debe mejorarse todavía para poder considerarlo un diagnóstico fiable.

10.2 CONCLUSIONES

- Se ha comprobado que el orden de las muestras dentro de la matriz no es influyente a la hora de entrenar un sistema de predicción.

- Se ha comprobado que el número de muestras tampoco lo es, a no ser que se llegue a casos extremos, donde hay demasiadas muestras de una clase y muy pocas de otra.

- Para las combinaciones de 6 electrodos se han obtenido buenos resultados, estos podrían mejorarse con un entrenamiento con más clases "CTRL", ya que el número de clases "BC" es mucho más elevado.

- Para combinaciones de electrodos menores a 6 se produce un sesgo, este puede ser debido también a la poca cantidad de clases "CTRL" y el gran número de clases "BC", por lo que quizás con un número más equitativo combinaciones de 4 o 5 electrodos puedan ser de igual modo tan buenas o mejores que combinaciones de 6 electrodos.

- Los electrodos que mejores resultados dan para la clasificación son el de plata, cobalto, cobre y rodio, mientras que el electrodo que menos resultados positivos aporta es el electrodo de iridio.

- Se ha conseguido un sistema de predicción basado en un análisis discriminante cuadrático con PCA con una precisión de aproximadamente 75%. Este valor puede mejorarse optimizando el sistema, sobre todo con la inclusión de más muestras de clases "CTRL".

- La adición del electrodo de níquel a las combinaciones de 5 electrodos consigue eliminar el sesgo del sistema y se comienzan a obtener buenos resultados, esto convierte al electrodo de níquel, junto con el electrodo de plata, en los dos electrodos fundamentales para la realización de las medidas, a pesar de que a priori el electrodo de níquel obtuviese malos resultados, se ha comprobado que su presencia en las combinaciones es necesaria.

11. Bibliografía

[1] Instituto Nacional del Cáncer NIH. https://www.cancer.gov/images/cdr/live/CDR765552-750.jpg (5 de Agosto 2017).

[2] Instituto Nacional del Cáncer NIH. https://www.cancer.gov/espanol/tipos/vejiga/paciente/tratamiento-vejiga-pdq (5 de Agosto 2017).

[3] American Cancer Society: Cancer Facts and Figures 2017. Atlanta, Ga: American Cancer Society, 2017.

[4] J. Olsson, P. Ivarsson, F. Winquist, Talanta, Vol. 76, pp. 91-95, 2008.

[5] A. Gutés, F. Céspedes, M. del Valle. Electronic tongues in flow analysis. Analytica Chimica Acta 600 (2007) 90-96.v

[6] E. Baldeón. Tesis Doctoral: Desarrollo de la técnica de lengua electrónica voltamétrica para la determinación de la capacidad antioxidante total de extractos de plantas y frutas peruanas.

[7] M. Alcañiz Fillol. Tesis Doctoral: MIguel Alcañiz Fillol. Diseño de un sistema de lengua electrónica basado en técnicas electroquímicas voltamétricas y su aplicación en el ámbito agroalimentario. Universidad Politécnica de Valencia (2011).

[8] Patrik Ivarsson, Christina Krantz-Rülcker, Fredrik Winquist and Ingemar Lundström. A voltametric electronic Tongue (2005) S-SENCE and Laboratory of Applied Physics, Linköping University, SE-581 83 Linköping, Sweden.

[9] Manual de utilización de Matlab https://es.mathworks.com/help/ (Agosto 2017)

[10] Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report. Palo Alto (USA): HP Laboratories; (2004).

[11] Implementation of an electronic nose to detect patients with COPD from exhaled breath

Cristhian Manuel Durán Acevedo Adriana Eugenia Velásquez Carvajal Oscar Eduardo Gualdron Guerrero .Revista Ingeniería y desarrollo (2012).

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCUELA TÉCNICA SUPERIOR INGENIEROS INDUSTRIALES VALENCIA

TRABAJO FIN DE MASTER EN INGENIERÍA QUIMICA

PRESUPUESTO DEL TRABAJO

1. Presupuesto

Este documento tiene como finalidad la estimación del coste del proceso de elaboración de este Trabajo final de máster. Para realizar la estimación, se tendrá en cuenta tanto los costes de realización del trabajo por parte de todas las personas implicadas como los costes de la instrumentación utilizada. Debe señalarse que para la realización de este trabajo las muestras ya se proporcionaron medidas, por lo que ni los aparatos de medición ni el coste del personal de medida como mano de obra serán considerados en este presupuesto.

1.1 MATERIALES

Los materiales que han sido utilizados para la realización de este TFM han sido un ordenador portátil y el programa informático Matlab. A estos materiales se le asignan los siguientes costes:

- Ordenador portátil
  - Coste del ordenador: 1350 €.
  - Años de amortización del ordenador: 5 años.
  - Horas aproximadas de uso del ordenador al año: 1700 horas.
  - Horas de uso del ordenador en 5 años: 8500 horas

$$Coste\ ordenador\ = \frac{1350\ €}{8500\ h}\ =\ 0{,}16\ €/hora$$

- Licencia de estudiante de Matlab
  - Coste de la licencia: 0 €

1.2 PERSONAL

Con la finalidad de estimar los costes de mano de obra en la realización de este trabajo se utiliza como apoyo la información proporcionada por el Centro de apoyo a la innovación, la investigación y la transferencia de tecnología (CTT) de la Universidad Politécnica de Valencia (UPV). Esta información se encuentra divulgada en su página web (http://i2t.webs.upv.es/2-serv-upv/serv-upv-2/Tarifas2007.htm).

En base a estos datos, los precios que se van a estimar son:

- Proyectista: 19,60 €/hora
- Tutor de la escuela: 26,79 €/hora
- Cotutor del proyecto: 27 €/hora

1.3 CUADROS DE PRECIOS

Cuadros de precios de mano de obra

MANO DE OBRA					
Nº	**Ud.**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio Total (€)**
1.1	Horas	Proyectista	350	19,69	6891,5
1.2	Horas	Tutor académico	50	26,79	1339,5
1.3	Horas	Cotutor académico	5	27	135,0
Total mano de obra				**Suma total**	8366,0

Cuadro de precios de materiales

MATERIALES					
Nº	**Ud.**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
2.1	horas	Ordenador portátil	296	0,16	47,36
2.2	unidad	Licencia Matlab	1	0	0,00
Total materiales				**Suma total**	47,36

1.3.1 Cuadros de precios desglosados en tareas

Tarea ID01

DESCRIPCIÓN: ID01					
Planteamiento del TFM y definición de los objetivos a conseguir					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	6	19,69	118,14
M.O.2	Horas	Tutor académico	5	26,79	133,95
M.O.3	Horas	Cotutor académico	5	27	135,00
Total				**Suma total**	387,09

Tarea ID02

DESCRIPCIÓN: ID02					
Análisis previo y búsqueda de información					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	70	19,69	1378,3
M.O.2	Horas	Tutor académico	10	26,79	267,9
MAT1	Horas	Ordenador	40	0,16	6,4
Total				**Suma total**	1653,2

Tarea ID03

DESCRIPCIÓN: ID03					
Elaboración del trabajo					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	195	19,69	3839,55
M.O.2	Horas	Tutor académico	15	26,79	401,85
MAT1	Horas	Ordenador	195	0,16	31,20
MAT2	Horas	Licencia Matlab	180	0	0,0
Total				**Suma total**	4272,6

Tarea ID04

DESCRIPCIÓN: ID04					
Seguimiento y reuniones					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	15	19,69	295,35
M.O.2	Horas	Tutor académico	15	26,79	401,85
Total				**Suma total**	697,20

Tarea ID05

DESCRIPCIÓN: ID05					
Diseño del documento final y corrección de errores					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	50	19,69	984,5
M.O.2	Horas	Tutor académico	3	26,79	80,4
MAT1	Horas	Ordenador	50	0,16	8,0
Total				**Suma total**	1072,9

Tarea ID06

DESCRIPCIÓN: ID06					
Realización y preparación de la presentación					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	10	19,69	196,90
M.O.2	Horas	Tutor académico	1	26,79	26,79
MAT1	Horas	Ordenador	8	0,16	1,28
Total				**Suma total**	224,97

Tarea ID07

DESCRIPCIÓN: ID07					
Realización y preparación de la documentación técnica final					
Identificador	**Unidades**	**Concepto**	**Cantidad**	**Precio unitario (€)**	**Precio total (€)**
M.O.1	Horas	Proyectista	4	19,69	78,76
M.O.2	Horas	Tutor académico	1	26,79	26,79
MAT1	Horas	Ordenador	3	0,16	0,48
Total				**Suma total**	106,03

1.3.2 Resumen de tareas y presupuesto final

Tarea	Descripción	Coste (€)
ID01	Planteamiento del TFM y definición de objetivos a conseguir	387,09
ID02	Análisis previo y búsqueda de información	1653,20
ID03	Elaboración del trabajo	4272,6
ID04	Seguimiento y reuniones	697,20
ID05	Diseño del documento final y corrección de errores	1072,90
ID06	Realización y preparación de la presentación	224,97
ID07	Realización y preparación de la documentación técnica final	106,03
Total	Coste total de realización del TFM	**8413,99**

10	Análisis voltamétrico de orinas de cáncer de vejiga mediante una lengua electrónica. Elaboración de un modelo de detección