UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Informática

Análisis de biomarcadores de cáncer colonorrectal (CCR) con técnicas de machine learning para la priorización en el cribado de colonoscopias.

Trabajo Fin de Grado

Grado en Ciencia de Datos

AUTOR/A: Sebastián Peris, Lucas

Tutor/a: Carot Sierra, José Miguel

CURSO ACADÉMICO: 2023/2024

Resumen

El cáncer colorrectal (CCR) se trata de uno de los cánceres más comunes y mortales en España. Su comportamiento asintomático en estadios tempranos y el rápido aumento en mortalidad en diagnósticos tardíos, proponen la necesidad de establecer programas de cribado para el reconocimiento poblacional y la rápida detección de esta enfermedad. El programa de cribado de la Comunidad Valenciana se ha estandarizado hasta alcanzar la mayor parte de la población objetivo, esto permite reconocer un número mayor de casos de CCR, pero también aumenta el tiempo de espera medio para las pruebas endoscópicas, lo cual es especialmente notable tras la reciente pandemia por COVID19.

En este trabajo se propone la elaboración de un algoritmo para la detección de casos de riesgo y su priorización para la colonoscopia. Para ello, se ha diseñado un proceso de transformación de datos para asegurar la disponibilidad de la información necesaria y relevante. Estos procesos incluyen la recodificación de casos, la selección de variables, la imputación de casos ausentes, el centrado y escalado de los datos y técnicas de muestreo como la generación de instancias artificiales. Tras ello, se ha estudiado el desempeño de una amplia gama de técnicas de clasificación, desde las más simples hasta algoritmos más complejos como las máquinas de vector soporte o los árboles de decisión. No solo se han estudiado estas técnicas de forma individual, sino también la combinación de diversos preprocesos junto con las variantes de los clasificadores, lo que generó un amplio conjunto de metodologías posibles. De entre ellas, se escogió junto con el equipo médico aquella que mejor se pudiera adaptar al contexto y las necesidades del proyecto.

Como resultado de este TFG se ha obtenido un algoritmo de árbol de decisión. Este modelo emplea diversos biomarcadores reconocidos de CCR, con una alta interpretabilidad y un enfoque particular en la reducción de falsos positivos. Cuenta con una sensibilidad de 0.95 y una especificidad de 0.19. Esto permite priorizar alrededor del 5% de los participantes del programa de cribado, donde el 74% de los priorizados padece neoplasia avanzada y un 18.7% CCR.

Finalmente, se han establecido una serie de necesidades y recomendaciones para la implementación del modelo al sistema de priorización de participantes en el programa de cribado.

Palabras clave: CCR, programa de cribado, modelos de clasificación, aprendizaje automático

Abstract

Colorectal cancer (CRC) is one of the most common and mortal cancer in Spain. Due to being asymptomatic and the rapid growth of mortality when diagnosed in late stages, proposes the necessity to establish screening programs for early diagnosis and faster detection. The screening programme in Comunidad Valenciana has been standardized up to the screening of most of the target population, which helps recognize more CRCs

but also increases the average waiting time for endoscopic procedures, especially after the recent COVID19 pandemic.

This paper proposes the development of an algorithm for the detection of at-risk cases and their prioritisation for colonoscopy. To this end, a data transformation process has been designed to ensure the availability of the necessary and relevant information. These processes include case recoding, variable selection, missing case imputation, data centring and scaling, and sampling techniques such as artificial instance generation. Following this, the performance of a wide range of classification techniques has been studied, from the simplest to more complex algorithms such as support vector machines or decision trees. Not only have these techniques been studied individually, but also the combination of different preprocesses together with variants of classifiers, which generated a wide set of possible methodologies. From these, the one that could be best adapted to the context and needs of the project was chosen together with the medical team.

As a result of this thesis, a decision tree algorithm has been derived. This model uses several recognized CRC biomarkers, with high interpretability and a particular focus on reducing false positives. It has a sensitivity of 0.95 and a specificity of 0.19. This allows to prioritize about 5% of the participants of the screening programme, where 74% of those prioritized have advanced neoplasia and 18.7% CRC.

Finally, a series of needs and recommendations have been established for the implementation of the model in the prioritization system for participants in the screening programme.

Keywords: CRC, screening programme, classification models, machine learning

Índice general

1: Introducción

El cáncer colorrectal (CCR) se debe al crecimiento descontrolado de células malignas con capacidad invasiva de las mucosas del colon o del recto. El paso preliminar al desarrollo del CCR es la aparición y el crecimiento de pólipos colorrectales. Un porcentaje de los pólipos colorrectales evolucionarán a CCR. En este complejo proceso están involucrados una serie de factores genéticos, epigenéticos y ambientales. El desarrollo y crecimiento de los pólipos suele ser un proceso. Sin embargo, existen algunos casos en los que este crecimiento es más acelerado y supone un riesgo superior de progresión a CCR.

El CCR es uno de los cánceres más comunes, representando cerca del 11% de todos los tumores diagnosticados en España. Es el segundo tipo de cáncer con mayor mortalidad en nuestro país, por detrás del cáncer de pulmón [1,2]. Esta elevada mortalidad se debe a un diagnóstico tardío en estadios avanzados, ya que el desarrollo de las lesiones pre-neoplásicas es un proceso asintomático.

Para evitar un diagnóstico tardío existe un programa de cribado poblacional cuyo objetivo es, por una parte, resecar los pólipos y evitar su progresión a CCR, y por otra parte, diagnosticar en estadios tempranos aquellos casos con CCR. Existen distintos tipos de cribado de CCR, aunque las pruebas más utilizadas son la prueba de sangre oculta en heces (TSOH), las pruebas inmunoquímicas fecales (FIT), pruebas de mutaciones genéticas y las colonoscopias [3]. Todas estas metodologías tienen como objetivo la detección de casos de riesgo entre una población de mayor riesgo por edad u otros factores de riesgo.

El TSOH es una prueba no invasiva en la que se mide la cantidad de sangre en nanogramos por mililitro (ng/ml) que se encuentra una muestra de heces del sujeto. Un TSOH positivo se ha relacionado con un mayor riesgo de presentar pólipos o CCR. Esta metodología de cribado destaca por su sencillez, su bajo coste y la facilidad con la que se puede generalizar para un gran número de participantes. Aunque como prueba de cribado es adecuada, al tratarse de un resultado dicotómico, no permite la estratificación del riesgo de CCR en aquellos casos con TSOH+. Además, existen otros motivos por los que un paciente pueda resultar positivo tras el TSOH, como la medicación con anticoagulantes y/o antiagregantes o padecer de úlceras, hemorroides o diverticulosis. Es importante destacar que el 50% de los sujetos con TSOH+ no tiene pólipos colorrectales. Dado que esta prueba no permite diferenciar entre quienes tendrán pólipos o incluso CCR, es necesario realizar una colonoscopia posterior.

La colonoscopia es la prueba más sensible y específica para la detección de pólipos y CCR. Esta prueba es tanto diagnóstica, ya que detecta pólipos y cáncer, como terapéutica, ya que permite resecar los pólipos para evitar su progresión a CCR. No obstante, esta prueba es más costosa, invasiva y además tiene pequeños riesgos asociados, como la posibilidad de perforación en el colon y el sangrado gastrointestinal.

De estas pruebas de cribado, se obtienen diferentes biomarcadores sanguíneos y fecales. Los biomarcadores son características definidas que indican procesos biológicos

normales, patologías o respuestas a intervenciones [4]. En el caso del cribado de CCR, los biomarcadores incluyen cualquier propiedad biológica que se pueda medir y se asocie con la presencia de pólipos o de cáncer. En este TFG se han empleado combinaciones de diversos biomarcadores para estratificar el riesgo de CCR antes de la colonoscopia de cribado.

En noviembre de 2005, la Comunidad Valenciana puso en marcha el programa de diagnóstico precoz de CCR, siguiendo una metodología en dos pasos en participantes voluntarios. Este cribado se propone a residentes de la comunidad autónoma de riesgo intermedio, de entre 50 y 69 años que se encuentran asintomáticos**.** A todos ellos, se les invita a participar voluntariamente en el programa y en caso afirmativo se les remite el TSOH de forma bienal. Esta prueba es analizada en laboratorio y en caso de ser positiva, se remite al paciente a la realización de una colonoscopia. La colonoscopia se utiliza como prueba confirmatoria de diagnóstico, para la detección de lesiones precursoras y la eliminación de estas.

En la actualidad se ha estandarizado el proceso de cribado para cubrir la mayor parte del público objetivo, lo cual ayuda a identificar un mayor número de pacientes de riesgo, pero también supone una creciente demanda de colonoscopias; con el consiguiente aumento del tiempo de espera medio para el procedimiento. Además, la pandemia por COVID19 supuso un empeoramiento de las listas de espera que afectó al tiempo hasta realización de pruebas endoscópicas. Dado que la citación para la colonoscopia realiza por orden de llegada a la lista de espera tras un resultado de TSOH+, el tiempo en espera para la técnica es igual entre todos los pacientes. Este aumento temporal puede no ser significativo para la mayoría de los casos, pero es crucial para aquellos pacientes con un CCR no identificado.

1.1. Motivación

El CCR es una enfermedad muy común, que cobra una relevancia superior entre los colectivos más envejecidos de la población. A diferencia de otras enfermedades; y particularmente, otros tipos de cáncer, el CCR es prevenible y tratable. Como se ha mencionado, una colonoscopia puede detectar lesiones precancerosas y pueden eliminarse para evitar su desarrollo en cáncer.

Los avances técnicos y organizativos actuales permiten la recolección de datos masivos y de calidad de una manera protocolizada, que facilita el aprendizaje y el desarrollo de metodologías Esta línea puede ser útil para la estratificación del riesgo de CCR previo a la colonoscopia, la priorización de la prueba en estos casos con un mayor riesgo, y el diagnóstico precoz de los casos con CCR para poder rentabilizar sus opciones terapéuticas.

El presente TFG plantea un algoritmo de decisión clínica en los sujetos incluidos en el programa de cribado de CCR. Además, también permitirá comprender la enfermedad, ya que el propio análisis conlleva nuevas observaciones y planteamientos.

Con lo que a lo personal respecta, siempre he tenido un gran interés por la utilidad social de los datos, y mi implicación en este proyecto me ha llevado a descubrir un área de conocimiento que me apasiona y en la que siento que los científicos de datos tenemos un papel fundamental como colaboradores en la construcción de sistemas de los que cualquiera puede beneficiarse.

1.2. Objetivos

El objetivo de este trabajo es valorar la viabilidad del uso de indicadores sanguíneos y características clínicas para la priorización de pacientes en el proceso de cribado de CCR de la Comunidad Valenciana.

Es fundamental que este proceso no suponga un sobrecoste significativo en el proceso de cribado y que pueda contribuir a reducir los tiempos de espera a aquellos pacientes con mayor riesgo de CCR. De la misma forma, se busca mejorar la comprensión de las tipologías de pacientes que participan en el programa e identificar características comunes y diferencias entre los casos de CCR, los casos de neoplasia avanzada y los pacientes sanos.

El objetivo del trabajo está sujeto a varios condicionantes, que deben tenerse en cuenta a la hora de definir los objetivos específicos:

- Caracterización de las tipologías de los pacientes. Distinguir patrones sobre los indicadores planteados que sean característicos de las tipologías de pacientes mediante técnicas exploratorias y de modelado.
- Clasificación de pacientes. Elaborar metodologías de modelado, técnicas estadísticas y de aprendizaje automático que permitan la clasificación de pacientes en CCR y sanos. Debido al contexto de aplicación del modelo, algunas consideraciones que se deben tener en cuenta son:
    - Dado que a todos los pacientes se les va a realizar la colonoscopia eventualmente, no se pretende encontrar un balance de modelo perfecto, sino uno en el que se pueda maximizar el número de CCRs identificados sin que la priorización de falsos positivos suponga un sobrecoste en los CCRs no identificados.
    - El modelo debe poder ser funcional frente a una situación de desbalanceo de clases, dada la baja prevalencia del CCR
    - Es necesario que el modelo sea interpretable y parta de fundamentos clínicos, debe existir cierta justificación médica para la priorización de los pacientes.
- Aplicabilidad. El enfoque debe usar únicamente la información de la que se dispone de forma previa a la colonoscopia y evitar el uso de atributos de difícil acceso. No se pretende crear un sistema de reordenación de colas complejo, sino un sistema de detección y priorización de casos de riesgo.
- Accesibilidad. La metodología seguida para el objetivo debe ser transparente y compresible para cualquier persona, independientemente de sus conocimientos estadísticos.

1.3. Impacto esperado

Se pretende encontrar una metodología de priorización en base al riesgo de presentar CCR en los pacientes incluidos en el programa de cribado. Para ello, se emplearán datos obtenidos de pruebas de "primera línea" como son la analítica sanguínea, el TSOH y las características de los sujetos incluidos.

A pesar de ello, sería sorprendente que se encuentre un sistema de clasificación donde se pueda identificar con facilidad la mayor parte de pacientes de CCR, esto se debe al componente de aleatoriedad que el cáncer plantea y a la información complementaria de la que no disponemos, como los indicadores genéticos que se han demostrado útiles en el diagnóstico temprano de CCR. También se ha de considerar la dificultad en el análisis de datos con clases desbalanceadas donde el número de casos de la clase minoritaria es muy reducido, lo que supone un obstáculo evidente al abordar el problema desde la perspectiva del aprendizaje automático y los modelos que dependen de grandes cantidades de información. La probabilidad de CCR en los pacientes con TSOH+ es alrededor del 7%.

Por tanto, con este trabajo se espera mejorar la comprensión de las características que componen esta enfermedad y encontrar una metodología que logre reducir altamente el tiempo de espera para colonoscopias de una cantidad reducida de pacientes que se presuponen de mayor riesgo.

2: Antecedentes y estado del arte

2.1. Antecedentes

2.1.1. Cáncer colorrectal

Para comprender mejor el diagnóstico de CCR, es necesario primero entender el desarrollo de pólipos en el colon y el recto. Este proceso se va a explicar en detalle a continuación, y se puede observar también en la figura 2.1.

La división celular en el tejido del intestino y del colón puede verse alterada por mutaciones genéticas que provocan una reproducción celular innecesaria, que causa que se formen pequeñas agrupaciones celulares conocidas como pólipos [5]. Eventualmente, estas células pueden expandirse sobre las paredes del colon y posteriormente a nodos linfáticos, desde donde pueden metastatizar a zonas distantes [6]. Pocos pólipos adquieren estos factores malignos, e incluso los que lo hacen, pueden tardar varios años o incluso una década en esta progresión de pólipo a cáncer. Existen diversas tipologías dentro de los pólipos, los adenomas y los serrados son precancerosos, mientras que los hiperplásicos e inflamatorios son habitualmente benignos [7]. Para facilitar la clasificación de los pólipos, se introdujo el concepto de "neoplasia avanzada", criterio que se aplica cuando se halla por lo menos un adenoma o serrado avanzado en el colon del paciente. La definición de pólipo avanzado (ya sea adenoma o serrado) se aplica cuando el pólipo es mayor o igual a 10 milímetros, es velloso y tiene displasia de alto grado (DAG). Esta terminología será empleada a lo largo del trabajo para discernir los grados de crecimiento de los pólipos.

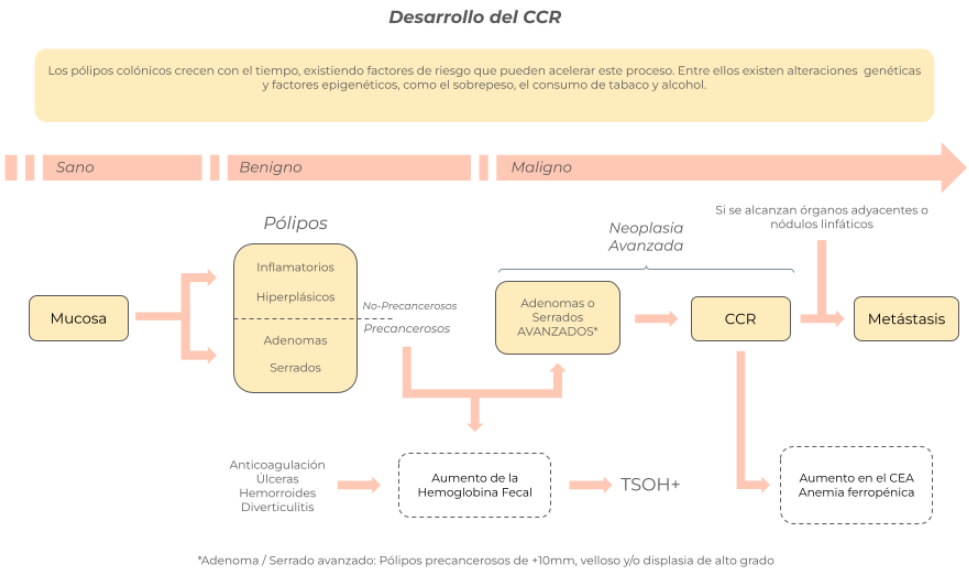


Figura 2.1: Diagrama del desarrollo del CCR.

Como consecuencia del crecimiento de los pólipos; y en particular, de los adenomas, la luz intestinal se ve obstruido por este tejido precanceroso, y con el flujo de elementos en él, los pólipos provocan pequeños sangrados no visibles en las heces, pero detectables tras analizar la muestra de heces. El resultado del TSOH se basa en esta medición, siendo positivo si el paciente tiene un sangrado fecal superior a 100 nanogramos por mililitro, lo que es indicativo de la presencia de pólipos. Se presupone que a mayor es el sangrado, habrá un número o tamaño superior de lesiones que lo producen. Es importante considerar que hay otros factores que pueden provocar sangrados no visibles que no se deben a la presencia de pólipos, como la medicación anticoagulante o la presencia de hemorroides. Es por eso por lo que la hemoglobina fecal será un biomarcador clave a lo largo del proyecto, aunque se deberán tomar algunas consideraciones en su uso e interpretación.

Además de los biomarcadores, se utilizará también información sobre los factores de riesgo del paciente. Los factores de riesgo son características sobre los pacientes que condicionan el desarrollo de pólipos, como la edad, el consumo de tabaco, la dieta o la presencia de otras patologías. Dentro de estos, la edad es uno de los más relevantes, pues se encuentra intrínsecamente relacionado con el tiempo de desarrollo de los pólipos. La prueba de cribado incluye este factor de riesgo dado que el TSOH solo se les realiza a personas de entre 50 y 69 años.

2.1.2. Sistemas de cribado

Es importante mantener presente que el objetivo de este trabajo es mejorar la implementación del sistema de cribado de la Comunidad Valenciana, por lo tanto, es necesario definir los criterios por los que se valoran estos programas. Los criterios por los que se rigen los programas de cribado se definieron por Wilson y Jungner en un artículo oficial publicado por la OMS en 1968. En este se mostraron los objetivos y las características de estos programas [8].

Los criterios en cuestión son los siguientes:

- La enfermedad debe constituir un problema de salud pública, dada una alta morbilidad, mortalidad y complicaciones
- La enfermedad debe tener una fase preclínica en la que se pueda detectar de forma temprana con una mejoría en su pronóstico
- La prueba de screening debe ser válida, sensible, específica y con una reproducibilidad y fiabilidad alta. También debe ser barata, accesible y segura.
- Se debe poder garantizar la confirmación diagnóstica y asegurar un tratamiento adecuado

Para mantener la integridad y utilidad del programa de cribado siguiendo estos principios, se deben tomar algunas consideraciones en el desarrollo de este trabajo para no decrementar las capacidades del programa.

- Esta metodología debe mejorar el rendimiento del programa de cribado, en particular sobre los pacientes de CCR sin empeorar de forma significativa el del resto de participantes.

- La metodología no debe excluir la prueba confirmatoria de diagnóstico (la colonoscopia en este caso); es decir, a todo participante con TSOH+ se le realizará la colonoscopia, independientemente de su riesgo de CCR.

- Para garantizar el bajo coste y la reproducibilidad de la prueba esta metodología no debe depender de información que no se asuma disponible para todos los participantes. Además, se debe evitar el uso de información susceptible a criterios subjetivos o que no esté presente en la mayor parte de casos.

El modelo propuesto en este trabajo debe usar información de fácil acceso, como los datos clínicos, la analítica de sangre y la cuantificación de hemoglobina fecal del TSOH con el objetivo de mejorar la capacidad de cribado del programa, priorizando un número reducido de casos de riesgo, de forma que estos se beneficien de un diagnóstico temprano sin que el resto de los pacientes experimenten un aumento en la espera significativo para el procedimiento confirmatorio.

2.2. Estado del arte

Conforme los métodos de análisis de datos y de inteligencia artificial han mejorado en accesibilidad para la comunidad científica, su uso es cada vez más habitual entre los ensayos clínicos y estudios médicos y el estudio del CCR no es menos. Existen diferentes propuestas para el diagnóstico temprano del cáncer, son muy notables aquellas en las que se hace uso de información genética mediante *Multitarget Stool DNA* [9] o incluso variantes empleando el RNA [10]. En estos trabajos también se incluye información de pruebas FIT y variables clínicas de interés.

Un área donde las técnicas de inteligencia artificial son ampliamente utilizadas es la detección de pólipos durante la colonoscopia [11], donde los modelos reconocen la imagen para facilitarle al especialista la identificación y clasificación de estos.

A pesar de la notabilidad de estas propuestas, se debe recordar que el propósito de este trabajo consiste en emplear la información de la que se predispone para su uso en la priorización de pacientes, no encontrar la forma perfecta de detección del CCR. Es cierto que emplear los resultados de pruebas como el *mt-sDNA, mt-sRNA* o la prueba FIT podría ser útil, o que una categorización más detallada de pólipos en la colonoscopia podría ayudar con la clasificación de pacientes, pero es información de la que actualmente no disponemos.

2.3. Crítica y propuesta

Es evidente que, a mayor cantidad de variables empleadas e información disponible, mejor será la calidad de predicción, pero los estudios previamente mencionados no dependen de ser usables en un contexto particular, valoran la utilidad de ciertas pruebas para el diagnóstico de CCR sin considerar la integración del modelo a un sistema de colas y priorización de pacientes.

La complejidad de los estudios previos es que se basan en análisis de material genético que solamente se hace en centros de referencia. El éxito de nuestro trabajo es la simplicidad del algoritmo de priorización de la colonoscopia mediante variables que pueden ser obtenidas en cualquier hospital.

Se propone reducir la complejidad de los modelos de clasificación actuales, empleando la información relevante para ello y evitando pruebas que supongan un sobrecoste en tiempo y recursos. Se pretende también valorar la suficiencia de información para mantener una calidad de clasificación y utilidad de priorización. De esta forma, la metodología resultante puede no ser la mejor en términos generales, pero sí se pretende que sea la mejor en el contexto que determina este trabajo.

3: Análisis del problema

3.1. Metodología

Para lograr los objetivos que se han establecido previamente, la metodología propuesta se fundamenta en el desarrollo de una estructura automatizable de modelado, o *pipeline*. El diseño de la arquitectura de esta *pipeline* se ha escogido siguiendo una secuencia de mejora iterativa, en la que cada nueva propuesta trata de mejorar la anterior, identificando sus puntos fuertes y débiles con el fin de encontrar un procedimiento lo suficientemente competente como para cumplir las expectativas y los objetivos de este trabajo.

La *pipeline* sigue una estructura en la que cada paso se ejecuta de forma secuencial, comenzando por la preparación de los datos y acabando en la generación de una lista de pacientes propuestos para la priorización de su colonoscopia.

El eje principal sobre el que se fundamente esta metodología es la elaboración de un modelo que afronte el problema como una tarea de clasificación binaria. La elección del modelo de predicción es clave para la *pipeline,* ya que del modelo depende el preprocesado previo de los datos. A pesar de ello, independientemente del modelo escogido, en todos los casos se aplica un filtrado y combinación de variables, un centrado y escalado de los datos y se lidia con los valores ausentes mediante un sistema de imputación. Los criterios seguidos para este tratamiento han sido establecidos de forma cooperativa mediante la combinación de las necesidades clínicas y estadísticas. Todas estas decisiones se verán reflejadas en los apartados 3.2 y en el capítulo 4.

Para comprender el diseño de la *pipeline* y del modelo escogido, es de suma relevancia recalcar uno de los objetivos principales del trabajo: No se pretende encontrar un balance entre la especificidad y la sensibilidad del modelo, sino que lo que se busca es priorizar la máxima cantidad de pacientes que realmente tienen CCR sin que los errores cometidos por el modelo perjudiquen a los casos de cáncer no detectado, aumentando su tiempo de espera. En otras palabras, optimizar el modelo maximizando su sensibilidad y minimizando los falsos positivos. Una vez el TSOH resulta positivo, todos los participantes se realizan una colonoscopia. En sí mismo, el proceso de cribado es un sistema muy específico, ya que pretende detectar todos los casos de CCR posibles, aunque eso suponga realizar pruebas “innecesarias”. Por ello el proceso descrito en este trabajo puede verse como una identificación de pacientes de riesgo, o un descarte de pacientes de bajo riesgo, y reducir el número de falsos positivos favorecerá la obtención de un modelo más específico que sensible.

Otra consideración a tener en cuenta es que a pesar de que el foco esté situado sobre el cáncer de colon, el desarrollo del mismo es un proceso gradual basado sobre el crecimiento de los pólipos, y es preferible priorizar a un paciente neoplásico frente a un paciente sano (entendiendo por “sano” un paciente no-CCR y sin neoplasia avanzada).

A la hora de observar los resultados del *pipeline*, los estadios con prioridad de identificación son; en este orden: CCR, neoplasia avanzada, neoplasia y paciente sano.

La interpretabilidad del modelo es también de alta relevancia en este trabajo, de esta forma se evitarán los conocidos "modelos de caja negra", como los modelos de redes neuronales o de aprendizaje profundo en los que, aún resultando de enorme eficacia en muchas ocasiones, en general es complejo entender el razonamiento de las decisiones del modelo sobre las predicciones. Además, se pretende que el personal médico pueda comprender el sistema de priorización, comprenda que está fundamentado en conocimientos sanitarios y que exista una trazabilidad sobre pacientes particulares, de forma que un especialista pueda observar los motivos por los que un paciente ha sido o no priorizado para su colonoscopia.

Una vez el planteamiento general del *pipeline* haya sido escogido, se debe optimizar el proceso de priorización para dar los mejores resultados posibles. Debido a las características de la tarea en cuestión, no se espera encontrar diferencias significativas en los resultados de los modelos tras ajustar los hiperparámetros del mismo; sin embargo, si que nos permitirá observar diferencias entre versiones de modelos que sigan la misma arquitectura.

Dado que se han empleado dos cohortes de datos distintas y dada la distinción de tipologías de pacientes, diferentes subconjuntos de datos han sido empleadas a lo largo del trabajo. Los conjuntos de datos empleados para cada modelo se mencionarán junto a los mismos. De la misma forma, se han realizado valoraciones acerca del funcionamiento de los modelos teniendo en cuenta estas diferencias tipológicas. Para verificar la veracidad de los modelos frente a la variabilidad de los casos, en todas las arquitecturas se ha aplicado un validación cruzada *10 fold*, en la que se subdividen los datos en entrenamiento y validación, de forma que aprende de un 80% de los datos y se verifica con el 20% restante, este proceso se repite 10 ocasiones, para evaluar el comportamiento del modelo frente a casos no observados previamente y disminuir el riesgo de entrenar o validar con un conjunto de datos que resulte sesgado.

A pesar de que el planteamiento principal se trata de la clasificación binaria en CCR y sanos, se han probado otros planteamientos que exploran posibilidades interesantes para la resolución de dificultades en la tarea de clasificación, como la clasificación multiclase, la generación de instancias artificiales, la modificación de los conjuntos de datos empleados, la predicción de probabilidades o el replanteamiento del *pipeline* general.

Finalmente, para poder evaluar el rendimiento de los modelos empleados, se van a utilizar principalmente las métricas que se obtienen en una matriz de confusión: sensibilidad, especificidad y los dos tipos de error. Como ya se ha explicado previamente, este trabajo tendrá un foco especial sobre la especificidad y el número de predicciones positivas frente al total de predicciones (porcentaje de pacientes que el modelo considera de priorización).

3.2. Requisitos

Una vez se ha comprendido la propuesta del trabajo, se hacen visibles una serie de requisitos o condiciones sobre las que la metodología escogida se debe atener. Como ya se ha explicado, esta no es una tarea de clasificación convencional. El contexto de aplicación es de alta relevancia, el modelo debe ser útil para un programa de cribado de una enfermedad de baja prevalencia y esto limita las posibilidades desde el punto de vista estadístico, pero también le aporta un valor intrínseco al problema y cierta belleza en su solución. Los requisitos principales son los siguientes:

- El modelo debe ser interpretable; es decir, cualquier persona no relacionada con el mundo de la estadística o la ciencia de datos debe comprender los motivos por los que un paciente es o no priorizado. No solo eso, el propósito del trabajo no es desarrollar nuevos biomarcadores para el CCR, sino encontrar la mejor combinación de marcadores ya establecidos para la predicción del CCR. Esto implica que el desarrollo del modelo requiere de un conocimiento elevado de la perspectiva clínica y de la colaboración entre médicos y expertos en datos.

- Dada la baja prevalencia del cáncer frente al número de pacientes sanos, el modelo debe lidiar con la diferencia entre el número de muestras de cada clase, o *desbalance de clases*. Esta cuestión debe ser tratada, ya que de no hacerlo el modelo aprenderá más de la clase mayoritaria, lo que habitualmente empeora la clasificación de la minoritaria.

- Junto con el concepto del desbalance de clases, es importante considerar que la dificultad en la extracción de datos (ya que es dependiente del número de colonoscopias realizadas) causa que se disponga de un número de casos reducido. El problema previo de desbalanceo en muchas ocasiones se resuelve con métodos que requieren un número mayor de instancias, lo cual no será posible para este trabajo y deben considerarse formas de resolver el desbalanceo sin necesitar una mayor N.

- Por último, debido a la componente aleatoria del CCR y por aquellas variables de las que no disponemos en este estudio (como resultados de pruebas genéticas), el modelo debe intentar en la medida de lo posible lidiar con la dificultad de clasificar el desarrollo de pólipos a cáncer. Debido a este factor aleatorio, es algo común encontrar pacientes de alto riesgo sin CCR y pacientes de bajo riesgo pero que han desarrollado cáncer. Es necesario encontrar una metodología que optimice el sistema de priorización asumiendo la dificultad de clasificar ciertos casos.

3.3. Plan de trabajo

Dada la envergadura del estudio y de la implicación del Hospital Clínico de Valencia en el mismo, se definió un plan de trabajo en el que todas las secciones fueran consideradas y que permitiera una mejor comunicación y colaboración entre los equipos.

El proyecto debía comenzar con la comprensión del contexto clínico, del sistema de colas del que se partía y de los factores clínicos considerados para el mismo. El equipo médico se encargaría de la obtención de datos inicial y de la definición de objetivos generales, posteriormente el equipo de datos se encargaría del tratamiento y análisis de datos y la definición de objetivos específicos.

Posteriormente comenzó el modelado del problema siguiendo una estructura de mejora iterativa de modelos, donde el nuevo modelo trataría de solventar las problemáticas o mejorar los resultados del modelo previo, asegurando en todo momento la comprensión del equipo médico, comunicando los métodos, procedimientos y resultados de forma clara y transparente. Este proceso se repetiría hasta encontrar un sistema que ambos equipos consideraran que cumple los objetivos y las expectativas preestablecidas.

Finalmente, se valoraron las implicaciones de la implementación del modelo escogido y se desarrollaron una serie de pautas y consideraciones acerca de la forma de llevar el sistema al contexto real. Se anotarían líneas de trabajo futuras y se valoraría la gestión del proyecto de la forma más objetiva posible.

3.4. Materiales y métodos

3.4.1. Recursos utilizados

Principalmente el trabajo ha sido realizado en Python [12], en su última versión (3.12). Se han empleado diversos ambientes de programación en función de las tareas a realizar, el tratamiento básico de datos se realizó en local en un ordenador de sobremesa, en Visual Studio y SublimeText, ya que no se requería de una gran capacidad computacional. Para el proceso de modelado se programó en Google Colaboratory, para no depender de la capacidad propia del ordenador en las ejecuciones. De forma puntual se utilizó R en Rstudio para realizar consultas y análisis sencillos y rápidos.

Dentro del lenguaje Python, se utilizaron una variedad de librerías gratuitas, entre ellas Pandas [13], sklearn [14], numpy [15], matplotlib [16], seaborn [17], plotly [18] o imblearn [19] entre otras.

Pandas y numpy se han utilizado para el manejo de datos de forma matricial, sklearn para la implementación de los modelos, imblearn para elaborar los modelos de muestreo y matplotlib, seaborn y plotly para elaborar las gráficas y visualizaciones necesarias.

3.4.2. Técnicas utilizadas

Visualización exploratoria

Diagrama de caja-bigotes: Muestra las distribuciones de una variable numérica frente a una categórica, útil para ver las diferencias en una variable particular entre CCR, neoplasia avanzada y sanos.

Matrices de confusión: En forma de tabla muestra el número de casos o proporciones en el cruce de dos variables categóricas. Útil para ver las diferencias entre la variable respuesta y factores categóricos como tabaco o el tratamiento con anticoagulantes.

Matriz de correlación: Muestra la correlación entre todas las combinaciones de variables numéricas del conjunto. Eficaz para observar variables relacionadas entre ellas o con colinealidad.

Modelos de imputación

Se decidió que la forma de lidiar con datos ausentes era la sustitución de estos valores mediante su imputación. Se realizó un estudio acerca de la mejor forma de proceder, y para ello se emplearon métodos como *K-Nearest Neighbors* (KNN), *Multivariate Imputation by Chained Analysis* (MICE) y *MissForest*.

Técnicas de sobremuestreo

Para lidiar con el desequilibrio de clases, se optó por generar instancias de forma artificial para ello se usaron principalmente dos técnicas: *Synthetic Minority Oversampling Technique* (SMOTE) [20] y *Adaptive Synthetic Sampling* (ADASYN) [21].

Modelos de clasificación

Regresión logística: Es uno de los métodos más comunes y sencillos para comenzar una tarea de clasificación binaria y puede ser útil para reconocer algunas variables de relevancia para la predicción

SVC: *Support Vector Classifier* es la variedad de las máquinas de vectores soporte para las tareas de clasificación, que aplica el uso de vectores multidimensionales que actúan de corte para alcanzar la mayor separabilidad en la clasificación posible.

Árboles de decisión: Un árbol de decisión es un modelo donde se escogen de forma iterativa variables del conjunto por las que particionar los datos, proporcionando un camino o "árbol" formado por el conjunto de variables y puntos de corte. Este es un método muy útil para lidiar con clases desequilibradas con alta interpretabilidad.

RandomForest: Este modelo fundamentado en árboles de decisión consiste en combinar versiones ligeramente alteradas de árboles mediante un sistema de ensamblaje que considera los resultados en las predicciones de cada variación árbol de decisión. Este es un modelo que está funcionando realmente bien en los últimos años gracias a sus buenos resultados.

3.5. Marco legal

Este trabajo fue presentado al comité de ética del Hospital Clínico de Valencia y fue aprobado de forma favorable el 29 de julio de 2021. La participación en el estudio es voluntaria y se solicita la autorización de inclusión mediante la firma de un documento de consentimiento informado.

4: Preprocesado y comprensión de los datos

4.1.Descripción del conjunto de datos

El conjunto de datos que se ha utilizado en este estudio se ha obtenido de participantes del programa de cribado del área de salud "Clínico – Malvarrosa". El procedimiento seguido para la obtención de los datos está representado en la figura 4.1.

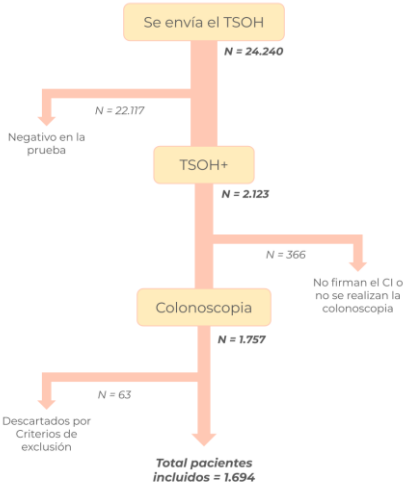


Figura 4.1: Flujo de datos del programa de cribado.

La prueba de TSOH se envió a un total de 24.240 personas, los motivos de descarte para la base de datos son los siguientes:

- El resultado del TSOH del paciente resultó negativo

- El paciente no firmó el consentimiento informado (CI) o no se presentó a la colonoscopia

- El participante cumplía al menos un criterio de exclusión. Estos criterios excluyen a todos los participantes en los que la colonoscopia resulta sucia (valor BOSTON < 5), las colonoscopias en las que no se llegó al punto ciego y los pacientes que tenían enfermedades inflamatorias.

Este flujo de datos resultó en dos cohortes de datos de pacientes válidos para el estudio, una primera cohorte de 1000 pacientes y una segunda de 694.

La información de la que disponemos de cada paciente procede de cuatro fuentes diferentes:

- **Datos clínicos:** Información disponible en los historiales clínicos de cada paciente, aquí encontramos variables básicas como la edad, sexo, peso o la altura, información patológica en variables binarias como diabetes, hipertensión arterial y tratamientos como los anticoagulantes (ACO) o antiagregantes (AG).

- **TSOH:** La prueba de sangre oculta nos aporta principalmente la variable *Hb fecal*, que representa un valor numérico para la cantidad de nanogramos de sangre por mililitro de heces. Esta información es clave pues indica el nivel de obstrucción provocada por los pólipos de colon.

- **Analítica de sangre:** Todos los participantes se realizan una analítica de sangre tras dar positivo en el TSOH, ya sea recetada por el médico de cabecera o el mismo día de la colonoscopia. De esta prueba obtenemos 26 variables numéricas, entre ellas encontramos varios biomarcadores conocidos de CCR.

- **Colonoscopia:** Dado que la colonoscopia es aquello que deseamos priorizar, la información procedente de ella solo puede ser utilizada como variables respuesta. De esta fuente encontramos variables como la presencia de cáncer, de neoplasia avanzada y otros valores referentes a los pólipos.

4.2. Consideraciones clínicas sobre los datos

De forma previa al análisis de datos, en una de las reuniones con el equipo médico se establecieron una serie de consideraciones relevantes de cara al modelado de la *pipeline*. Algunos de estos factores ya se han mencionado, pero agruparlos puede ser útil en la preparación del modelo.

El CCR y la neoplasia avanzada tienen una baja prevalencia, en nuestros datos, estas condiciones representan el 5,74% y el 38,85% respectivamente. Esto implica que solo disponemos de 84 casos de cáncer con los que entrenar el modelo y que se debe de lidiar con el desbalance de clases.

De las variables disponibles, no todas ellas pueden ser utilizadas para la predicción, sino que más bien se usarán para explorar el conjunto de datos. Este es el caso de las variables de colonoscopia, pero también el de otras variables por múltiples razones. Por ejemplo, por cuestiones éticas no se puede emplear el sexo del participante, no se pueden usar los datos de antecedentes por la dificultad de obtener esa información y no se deben emplear los resultados del cuestionario de sintomatología que se le realiza al paciente ya que vencería el propio propósito del cribado.

Hay algunas variables con especial interés médico, esto implica que se deberá observar el efecto de estas variables sobre los modelos y las predicciones. Entre ellas variables como la hemoglobina fecal o la edad son interesantes ya que se utilizan para focalizar el programa de cribado y establecer el positivo en el TSOH. Los factores de riesgo como peso elevado, el tabaco o el alcohol se conoce que pueden acelerar el desarrollo de pólipos en el colon, y finalmente el CEA es considerada un marcador tumoral. A la hora de realizar los análisis expuestos a continuación se situará un foco especial sobre estas variables.

4.3. Preprocesado

El cambio principal que le aplicó a los datos fue descartar a los pacientes que tomaran anticoagulantes y/o antiagregantes, ya que considerando la relevancia que la hemoglobina fecal tendría, incluir pacientes con un sangrado fecal elevado podía generar confusión con esta variable, al excluirse estos pacientes, si el sangrado en heces es elevado, la sospecha de pólipos desarrollados es mayor.

El siguiente paso fue observar las distribuciones de cada variable, lo cual resultó muy útil para detectar valores erróneos y fallos en la codificación. Esos casos fueron revisados individualmente con el equipo médico y se corrigieron los fallos.

Con respecto a la hemoglobina fecal, la medición de la variable sufrió un cambio durante el proceso de recolección de datos. Inicialmente el valor máximo de la variable era de 1000 ng/ml debido a una falta de precisión en la medición. Eventualmente se consiguió aumentar la precisión, pero esta diferencia alteraba de forma negativa el escalado de variables y se prefirió mantener el máximo en 1000 ng/ml y dejar este incremento en precisión para estudios futuros.

En algunos casos se crearon variables combinadas como es el caso del índice de masa corporal (IMC), el cual fue calculado en base al peso y la altura de los participantes. Al tratarse de combinaciones, mantener las variables previas podía provocar una colinealidad en los datos y por eso se decidió prescindir de las variables previas. El siguiente paso en el preprocesado fue lidiar con los datos ausentes, que, aunque no eran demasiado abundantes, se debían considerar.

4.3.1. Datos ausentes e imputación

En primer lugar, se observó la distribución de los datos ausentes, de la cual se extrajo información relevante. La mayor parte de faltantes se encontraba en las variables de la analítica de sangre, aunque en ninguna de las variables la proporción de ausentes superaba el 20%, que es el mínimo que se había establecido para la imputación.

Las variables clínicas y la hemoglobina fecal se encontraban casi completas, las variables de la colonoscopia no tanto, aunque los ausentes se concentraban en las variables referentes al tamaño de los pólipos, mientras que las variables de interés como el CCR o la neoplasia avanzada se mantenían intactas. Otro factor para considerar se trataba de dónde se localizaban los ausentes por filas, donde se podía ver una tendencia a que un mismo participante poseyera múltiples ausentes mientras que muchos otros no contaban con ninguno.

Dado que las variables de la analítica eran las únicas numéricas, se optó por realizar un estudio de métodos de imputación para este conjunto de variables, mientras que las variables restantes serían imputadas por la moda debido a la baja prevalencia de ausentes. El estudio se realizó sobre el total de 31 variables sanguíneas y en él se realizaron pruebas con los modelos de imputación por mediana, KNN, MICE y

MissForest. La metodología seguida para este estudio se puede observar en la figura 4.3.1 (1).

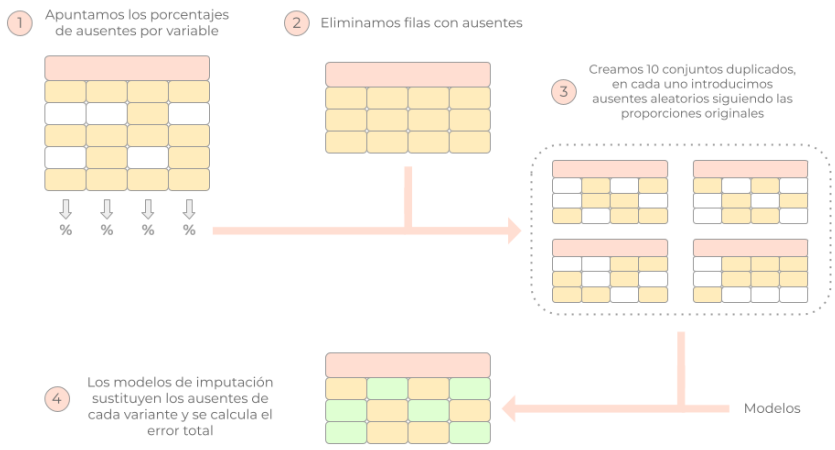


Figura 4.3.1 (1): Metodología para la imputación de datos ausentes.

El estudio se inició aplicando un escalado estándar y anotando las proporciones de ausentes por variable y eliminando todas las filas del conjunto de datos que contuvieran por lo menos un valor ausente. El hecho de que los ausentes se acumularan en las mismas filas fue un factor positivo, ya que debido a ello en este paso tan solo se descartaron alrededor de 300 filas.

Tras eliminar los casos incompletos, se obtuvo un conjunto de datos de menor tamaño, pero sin ningún valor ausente, lo que permitió usarlo de referencia para valorar la calidad de la imputación en el último paso. Sobre estos datos, se crearon 10 variaciones en las que se eliminaron valores de forma aleatoria, pero siguiendo las proporciones de ausentes originales.

Para la imputación se entrenaron los modelos previamente mencionados con diferentes ajustes de parámetros. Cada uno de ellos predijo los ausentes en todos los conjuntos, y como se conocían los valores reales se calcularon las métricas de evaluación y la media de las métricas entre las 10 variantes. Los resultados de los modelos más relevantes se observan en la figura 4.3.1 (2).

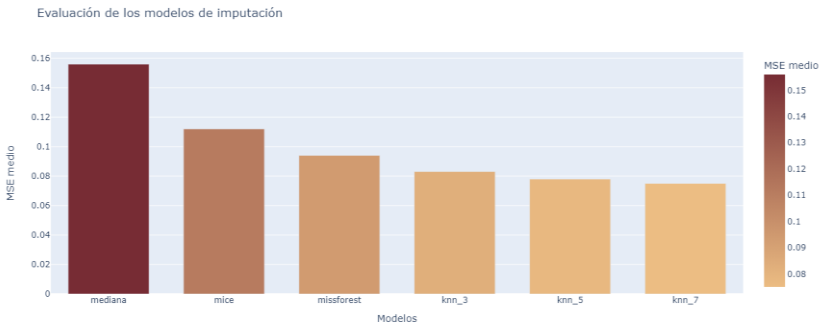


Figura 4.3.1 (2): Resultados de MSE medio de los modelos de imputación de ausentes.

La imputación por la mediana se incluyó para comprar resultados y, como era esperado, esta proporcionó peores resultados que cualquier modelo de mayor complejidad. Los mejores resultados se obtuvieron en los modelos de K vecinos más cercanos, con valores muy similares entre su ajuste de número de vecinos. Si bien es cierto que el menor MSE es obtenido con 7 vecinos, se conoce que a mayor número de vecinos mayor es el riesgo de sobreajuste, especialmente considerando el número reducido de casos del conjunto de datos. Por este motivo se optó por sacrificar la mejora del error entre 5 y 7 vecinos y se escogió el de menor K. En cuanto a la distancia empleada se probó tanto con la euclídea como con la distancia de Manhattan, la métrica de evaluación resultó similar en ambos casos, aunque finalmente se usó la euclídea, debido a que se conoce que la distancia de Manhattan funciona especialmente bien en conjuntos de alta dimensionalidad [22] y este no es uno de esos casos.

Tras realizar el estudio y escoger el modelo para la sustitución de ausentes, se retrocedió al conjunto de datos original y se aplicó KNN con distancia euclídea y 5 vecinos sobre las variables numéricas y se imputó usando la moda en las variables categóricas.

4.4. Análisis exploratorio

Antes de comenzar con el modelado de la *pipeline*, se realizó una exploración del conjunto de datos para mejorar la comprensión sobre la información de la que se dispone e intuir relaciones básicas entre las variables.

En primer lugar, se observaron las distribuciones de las variables numéricas. En el anexo 1.1, se observa que las variables seguían mayoritariamente distribuciones normales o distribuciones asimétricas positivas, lo cual es muy común en las variables de la analítica de sangre. Este también es el caso de la hemoglobina fecal, aunque al ajustar el máximo en 1000 ng/ml por el motivo que se comentó anteriormente, la distribución tiene asimetría positiva hasta ese máximo.

El siguiente paso fue comprobar las correlaciones entre estas variables numéricas, para ello se empleó la matriz de correlación que se encuentra en el anexo 1.2. En esta matriz se aprecian correlaciones positivas que corroboran correlaciones teóricas. Algunos ejemplos son cHDL junto con cLDL o el GGT junto con GOT y GPT entre otras. Considerar estas correlaciones puede ser práctico en la interpretación de los modelos, si existe colinealidad entre dos variables a menudo los métodos le otorgan relevancia a una de las dos, mientras que la importancia de la otra se ve reducida ya que ambas variables aportan información similar.

Para entender mejor la relación entre los predictores y las variables respuesta, se construyeron tablas cruzadas entre los atributos predictores categóricos y las variables CCR y neoplasia avanzada. Estas tablas se utilizaron para realizar pruebas de chi cuadrado, los resultados de esta prueba se observan en la figura el anexo 1.3. Aparentemente estos resultados indican que el tabaco es la única variable categórica con efecto significativo sobre la predicción de cáncer, aunque eso no quiere decir que el resto

de las variables no sean de relevancia para los modelos, ya que pueden servir para discernir entre arquetipos de participantes o tener una relación indirecta sobre el CCR.

Se exploraron también las relaciones entre las variables numéricas de la analítica de sangre y las variables categóricas mediante una serie de gráficos de caja-bigotes. En particular, se observaron especialmente aquellos atributos con interés clínico, ejemplos de ello son la edad o el CEA, cuyas visualizaciones se pueden observar en el anexo 1.4. Aunque de forma leve, al desagregar las variables según la presencia de CCR se puede observar cierta diferencia en sus distribuciones. Aunque esto no es determinante, podría ser indicativo de la edad como factor de riesgo en el desarrollo canceroso y el CEA como marcador tumoral.

Se realizó la misma visualización desagregando por CCR, neoplasia avanzada, anticoagulantes, antiagregantes y pacientes sin ninguna condición de las mencionadas, el gráfico en cuestión se trata de la figura 4.4. Este aporta información muy útil, aparentemente los pacientes de CCR, neoplasia avanzada y los participantes que toman anticoagulantes tienden a valor más elevados de hemoglobina fecal. Como era de esperar, las distribuciones de hemoglobina fecal de los pacientes con mayor desarrollo de pólipos se encuentran desplazadas hacia valores más elevados.

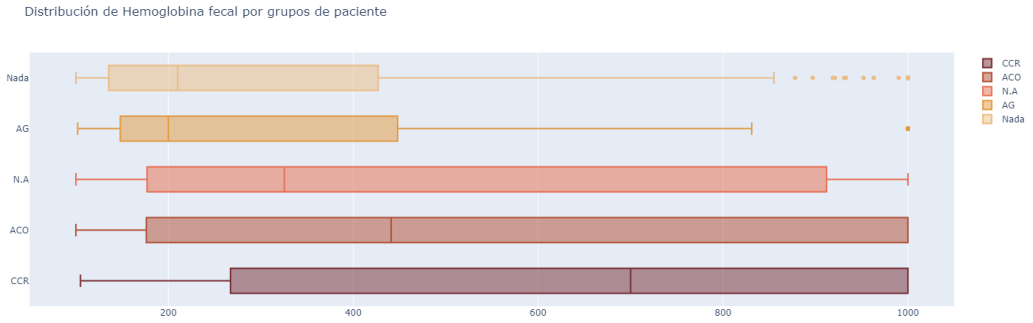


Figura 4.4: Distribución de la hemoglobina fecal, segmentada por tipología de participante.

Este gráfico también sirvió para justificar la exclusión de los participantes con anticoagulantes, ya aumenta la sospecha de un falso positivo en la prueba de cribado. Recordemos que el TSOH trata de detectar obstrucciones causadas por pólipos mediante un aumento en la hemoglobina fecal, si el participante resultó positivo a causa de la medicación y no a causa pólipos en colon, entonces este individuo no debería ser incluido en el conjunto de datos. A pesar de que inicialmente se excluyeron tanto los participantes con anticoagulantes como aquellos con antiagregantes, este gráfico sería empleado más adelante para justificar la inclusión de aquellos con antiagregantes.

5:Modelado

A lo largo de la fase de modelado se han probado una gran variedad de métodos y combinaciones de parámetros y preprocesados que han aportado conocimiento al proceso. Se propuso mejorar los modelos de forma iterativa, de forma que progresivamente se redujeran las limitaciones y errores del modelo previo.

Como ya se ha comentado previamente, los requisitos del trabajo limitan las posibilidades de elección de modelos. Por ello, se descartaron modelos de caja negra como las redes neuronales y se priorizaron modelos interpretables y con facilidad para lidiar con clases desbalanceadas.

5.1.Submuestreo y sobremuestreo

Justificación

Una de las formas principales para trabajar con clases desbalanceadas es tratar de igualar o aproximar el número de instancias de cada clase, ya sea eliminando muestras de la clase mayoritaria o añadiendo muestras a la clase minoritaria. Los métodos que reducen el tamaño de la clase más abundante se conocen como algoritmos de submuestreo, mientras que a los que generan casos nuevos del grupo minoritario se les conoce como algoritmos de sobremuestreo.

La necesidad de considerar formas de lidiar con esta situación proviene de la forma en la que los algoritmos de predicción se entrenan. Esta casuística es similar al motivo por la que centramos y escalamos los datos ya que, si no lo hacemos, los modelos le otorgan mayor importancia a las variables que toman valores más elevados. De la misma manera, si lo entrenamos con un alto desbalance, las predicciones tenderán a ser del grupo donde se concentre el mayor número de muestras, lo que se conoce como sobre-clasificación en la clase mayoritaria.

En este trabajo se han probado diferentes combinaciones de métodos, a continuación, se presentan algunas técnicas y durante cada prueba presentada en este capítulo se mencionará qué preprocesado ha sido utilizado.

Submuestreo

Los algoritmos de submuestreo se dedican a escoger muestras de la clase mayoritaria para su descarte, tratando de minimizar la información pérdida tras su eliminación. El submuestreo no se ha explorado tanto en este trabajo, debido a que eliminar casos de participantes sin cáncer provoca que aumenten rápidamente los falsos positivos de cualquier modelo, lo cual queremos evitar.

En cualquier caso, un método de submuestreo que sí se ha empleado en el desarrollo de la *pipeline* es NearMiss [23], y en particular, su versión NearMiss-3. Esta

variante del método consiste en escoger muestras de la clase mayoritaria asegurándose de que estas muestras no se encuentren próximas a la frontera de decisión.

Sobremuestreo

Estos algoritmos tratan de generar muestras sintéticas de la clase minoritaria usando como referencia los datos de esta. Existen muchos métodos para ello, uno de los más habituales es *Synthetic Minority Oversampling* (SMOTE). Sin embargo, en este caso se usará SMOTE-NC que es una variante que permite generar instancias con tipos de datos mixtos (tanto numérico como categórico).

Otro modelo alternativo es ADASYN, la diferencia principal entre estos métodos es que ADASYN prioriza generar instancias próximas a aquellas que sean difíciles de clasificar. De la misma forma que el SMOTE, este método se debe combinar junto con SMOTE-NC ya que no puede usarse para generar valores categóricos.

5.2. SVC

Justificación

La primera aproximación al problema comenzó empleando el modelo de *Support Vector Machine* (SVM), en particular *Support Vector Classifier* (SVC), que es su variante para tareas de clasificación. Aunque el modelo base no es especialmente bueno en contextos de desbalance, se puede ajustar el parámetro de peso de clases a "balanceado" para otorgarle el mismo peso a ambas clases independientemente del número de casos en cada una.

Otra ventaja del SVC es su interpretabilidad, el modelo se puede comprender usando diferentes visualizaciones. Se pueden observar los coeficientes de un modelo SVC de núcleo (o *kernel*) lineal, se pueden visualizar las fronteras de decisión de forma bidimensional o tridimensional, se pueden calcular los coeficientes SHAP para representar el efecto de los predictores y obtener gráficos de dependencias parciales.

5.2.1. Clasificación binaria

La primera aproximación de SVC tenía el objetivo de predecir la presencia de CCR de forma binaria. Partiendo del conjunto de datos completo tras la estandarización y la imputación con KNN se lidió con el desbalance. Se empleó la combinación de ADASYN y SMOTE-NC para generar instancias hasta llegar a la proporción 1:10, posteriormente se aplicó NearMiss para eliminar individuos de la clase mayoritaria hasta alcanzar la relación 1:2, la cual se usa de forma estándar en casos de alto desbalance [24], ya que generar o eliminar más filas puede provocar peores resultados y susceptibilidad al sobre entrenamiento.

Para la evaluación del modelo se programó de forma manual una versión de validación cruzada de 10 particiones, en la que se aseguró que las instancias artificiales

se usaban únicamente para el entrenamiento, no para la evaluación. Esto permitió entrenar con muestras sintéticas y evaluar con muestras reales previamente no vistas.

En este punto del estudio no fue necesario el ajuste exhaustivo de hiperparámetros, ya que en primer lugar se querían valorar los resultados generales de diferentes metodologías sin necesidad de optimizar la predicción. El resultado del modelo promediado entre las particiones se puede observar en la figura 5.2.1 (1) en forma de matriz de confusión.

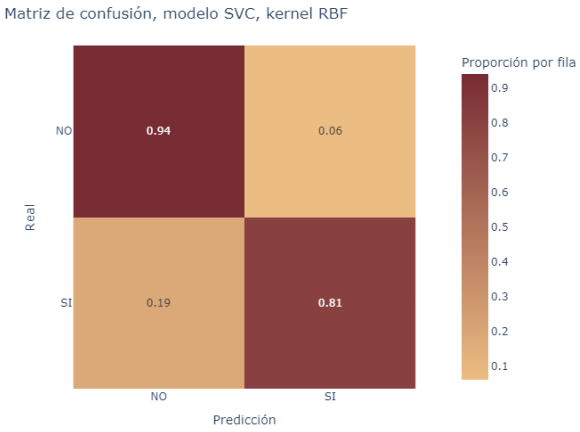


Figura 5.2.1 (1): Matriz de confusión sobre la validación cruzada (10 folds) del modelo SVC de kernel RBF.

Las predicciones obtenidas resultaron ser mucho mejores de lo esperado, esto levantó sospecha de sobre entrenamiento, pero cronológicamente aún no se disponía de la totalidad de la base de datos, y en este punto no se contaba con suficientes casos como para reservar parte de ellos para establecer un conjunto de validación. Por lo tanto, de estos primeros modelos tratamos de comprender la naturaleza del problema de clasificación y elaborar algunos enfoques iniciales.

La interpretación de este modelo fue clave para comprender la relación de los predictores con la variable respuesta y sirvió especialmente para destacar la relevancia de ciertos atributos, lo cual sería muy útil más adelante.

En primer lugar, se entrenó un modelo de SVC de *kernel* lineal, de forma que se pudieran interpretar los coeficientes. La visualización de los coeficientes se puede observar en el anexo 2.1. Algunos atributos destacables son el cLDL, el hierro o la hemoglobina fecal. Si bien puede ser algo útil, interpretar los coeficientes de un modelo de núcleo lineal para extrapolar conclusiones sobre modelos no-lineales no es del todo correcto. Por ello, el siguiente paso fue obtener esta relevancia de atributos mediante la permutación de importancias que ofrece la librería *sklearn*, empleando como métrica el descenso de *accuracy* del modelo como se observa en la figura 5.2.1 (2).

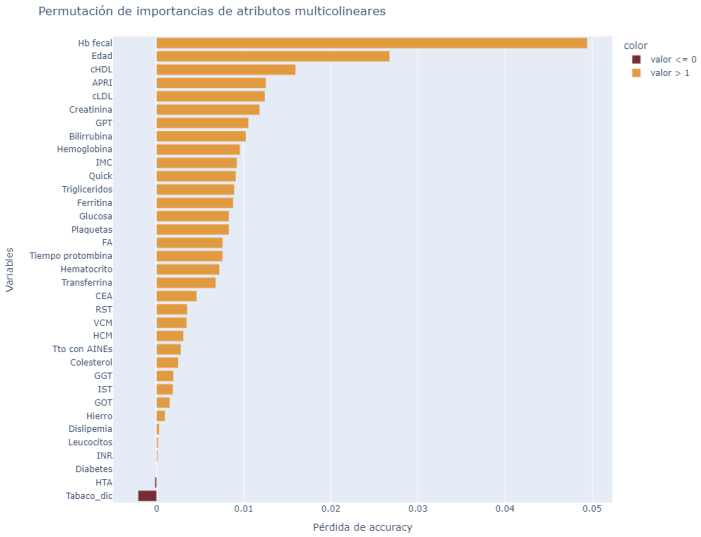


Figura 5.2.1 (2): Permutación de importancias de predictores del modelo SVC.

En este gráfico podemos observar las importancias de atributos para un modelo SVC con núcleo no-lineal, concretamente de *kernel* RBF. En este caso aumenta la relevancia de la hemoglobina fecal y de la edad, entre otras. Es de relevancia mencionar que estos valores pueden verse afectados por la colinealidad entre variables, disminuyendo la importancia de alguna variable relevante. También se puede observar que variables como el tabaco o la hipertensión arterial afectan negativamente a la predicción del modelo, lo que podría marcarlas como variables de confusión. Tras interpretar este gráfico, el último paso fue realizar una visualización de dependencias parciales sobre la variable respuesta (PDP) usando las dos primeras variables del gráfico anterior. En la figura 5.2.1 (3) observamos que la relación es muy clara y alineada con la teoría médica existente. La combinación de edades avanzadas y valores altos en hemoglobina fecal tienden a predicciones positivas en CCR, mientras que participantes jóvenes y con baja hemoglobina fecal tienden a predecirse como no-CCR en este modelo de SVC.

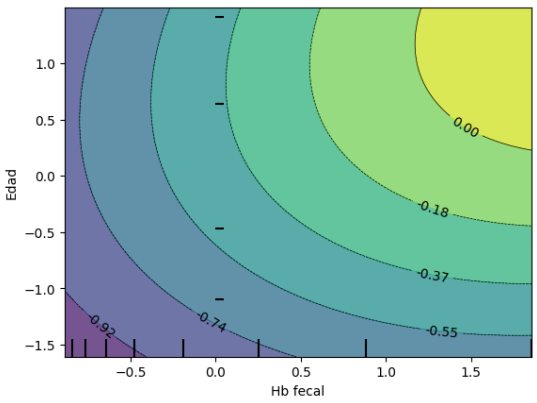


Figura 5.2.1 (3): Gráfico PDP de 2 atributos, hemoglobina fecal frente la edad.

5.2.2. Clasificación multiclase

Justificación

Otra de las aproximaciones que se quiso comprobar utilizando un modelo basado en máquinas de vector soporte fue la clasificación multiclase. Esto se propuso como una forma de clasificar en más de dos indicaros de riesgo; es decir, incluir la neoplasia avanzada como una categoría distinta al participante sano y al participante con cáncer. Este modelo de SVC trataría de predecir tres variables respuesta: participante sano ("Sano"), participante con neoplasia avanzada pero no CCR ("Neoplasia") y participante con neoplasia avanzada y cáncer de colon ("CCR").

Resultados

Los resultados de la figura 5.2.2 de la validación cruzada del modelo no son especialmente positivos, mejoran ligeramente la prevalencia de las condiciones, pero no logran evitar la sobre clasificación en la clase mayoritaria. En particular este es el caso de la clase CCR, en la que el modelo clasifica muy pocas muestras. Estos resultados no fueron sorprendentes debido a las siguientes cuestiones:

- La neoplasia avanzada es un criterio arbitrario escogido por el colectivo médico, útil para categorizar el avance del desarrollo de pólipos pero que no representa nada más allá de su propio crecimiento. Por esto mismo, esta variable respuesta es propensa a generar confusiones en los espacios cercanos al límite de su definición. Por ejemplo, un adenoma de 9mm no se considera neoplasia avanzada por criterio, pero sin embargo se asemeja mucho más a ella que un adenoma de 1mm, pese a tratarse de la misma clase.

- Plantear una clasificación multiclase produce un doble desbalance de clases, puesto que se deben de balancear las tres clases entre sí, dificultando el proceso de sobremuestreo.

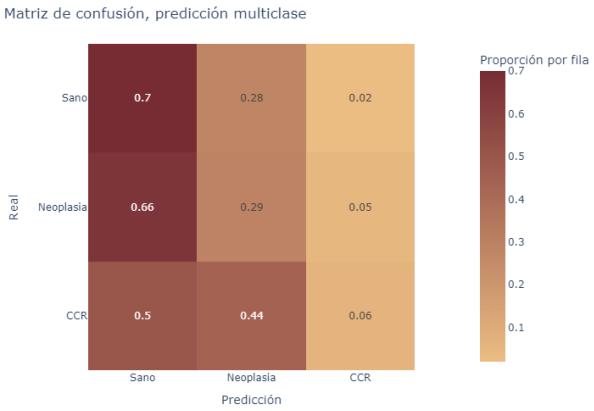


Figura 5.2.2: Matriz de confusión en validación cruzada. Modelo SVC multiclase.

5.3. Árboles de decisión y RandomForest

Justificación

Otra de las propuestas que se planteó inicialmente, se trataba de desarrollar un modelo basado en árboles. Se consideró que esta forma de realizar predicciones podría lidiar mejor con el desbalance de clases manteniendo o incluso mejorando la interpretabilidad del modelo. Al presentar el funcionamiento de los árboles de decisión, el equipo médico se vio muy atraído por esta metodología, puesto que se aproxima en gran medida al planteamiento clínico con el que ellos trabajan habitualmente. De la misma forma, un método ensamblador como *RandomForest* podría ser muy útil para mejorar las predicciones de un único árbol de decisión a costa de una pequeña pérdida de interpretabilidad. Podría parecer que al utilizar un método que emplea muchos árboles como lo hace *RandomForest* se pierde la capacidad de explorar el recorrido de un participante por el circuito del modelo, pero existen formas de escoger el árbol más representativo del conjunto [25].

Resultados

El funcionamiento de *RandomForest* superó los resultados del modelo de SVC; sin embargo, esto supuso que la sospecha de sobre entrenamiento estuviera más consolidada, en el siguiente apartado se explicará como la llegada de 600 nuevos casos afectó al modelado de la *pipeline*. Dada la cantidad de datos, se emplearon 20 árboles en el ensamblador, cada uno de ellos con una profundidad máxima de 10 y no se especificó el número máximo de hojas. La matriz de confusión en validación cruzada de este modelo se aprecia en la figura 5.3 (1).

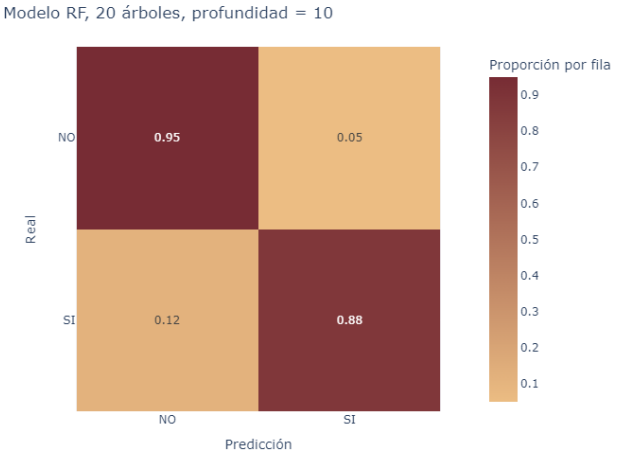


Figura 5.3 (1): Resultados del RandomForest en la validación cruzada.

En cuanto a la interpretación del modelo, en la implementación de *sklearn* es posible acceder al atributo "*feature_importances*" que es computado usando la media y desviación típica del descenso de impureza dentro de cada árbol generado por el modelo. Esto permite crear la siguiente visualización figura 5.3 (2).:

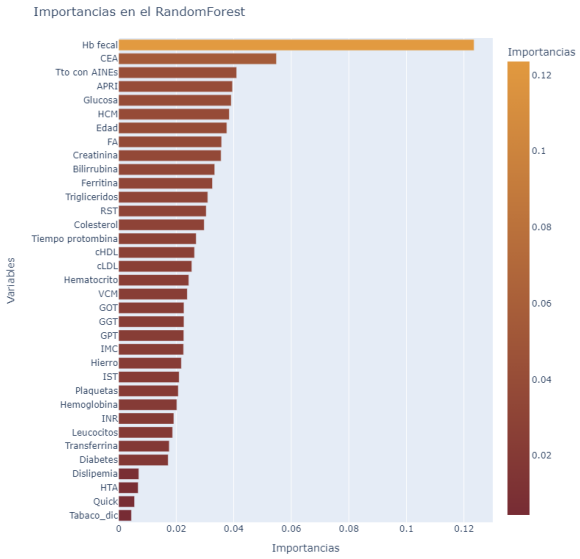


Figura 5.3 (2): Importancia de las variables contribuyentes al modelo de RandomForest.

De la misma forma que en el modelo SVC, la hemoglobina fecal tiene una relevancia muy significativa. En este modelo el CEA cobra una mayor importancia, esta variable sanguínea tiene un interés clínico ya que se considera un marcador tumoral.

5.4. El conjunto de validación

Como se ha mencionado en los apartados previos, se dieron sospechas de sobre entrenamiento, y en cuanto se pudo disponer del conjunto de datos completo, el primer paso que se realizó fue crear un conjunto de validación con el que verificar el funcionamiento de los modelos. Una vez se obtuvieron estos datos, se realizó un diseño de modelado que se presentó al equipo médico, este diseño se puede observar en la figura 5.4 (1).

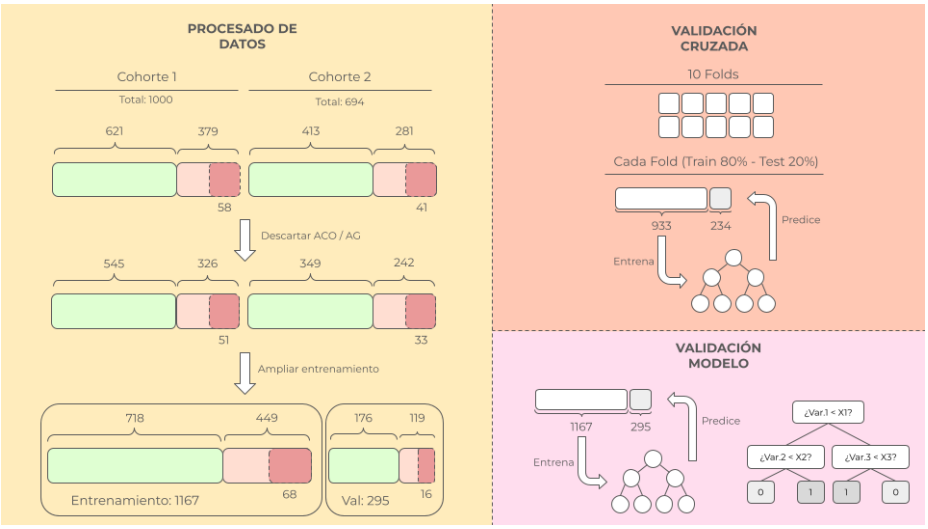
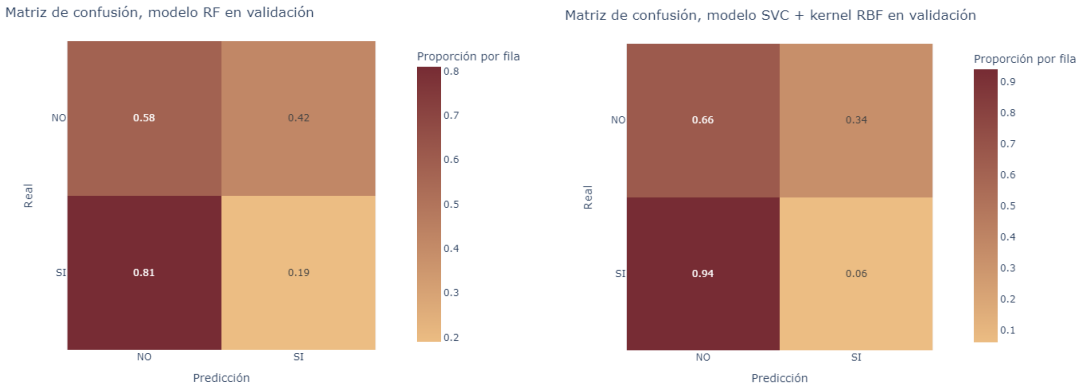


Figura 5.4 (1): Flujo de datos y metodología, tras la llegada de la segunda cohorte de datos.

Para aumentar el tamaño del conjunto de entrenamiento, la nueva cohorte de datos se dividió en dos partes, la primera se sumó al conjunto con el que se elaboraría la validación cruzada del modelo, mientras que la segunda parte se usaría como conjunto de validación. Con estos nuevos datos, se reentrenaron los modelos, las matrices de confusión en la evaluación del modelo resultaban similares pero los resultados en el conjunto de validación confirmaron las sospechas de sobre entrenamiento.

Los resultados en el conjunto de validación se encuentran en las figuras 5.4 (2) y 5.4 (3):



Figuras 5.4 (2) y 5.4 (3): Matrices de confusión de RandomForest y SVC en el conjunto de validación.

En estos resultados se observa que las diferencias entre los resultados de entrenamiento y validación son elevadas y radican particularmente en la alta disminución de sensibilidad de los modelos y del aumento de los falsos positivos. Inicialmente se consideró que se cometió un error o bien en el procedimiento de validación cruzada o bien en la comprobación sobre el conjunto de validación. Sin embargo, en los siguientes apartados se explicará de forma exhaustiva cómo la

complejidad de la circunstancia causó que se pasaran por alto detalles que provocaron este sobre entrenamiento.

5.5. El problema del sobre entrenamiento

Con la democratización del análisis de datos y la efectividad de modelos de alta complejidad, el sobre entrenamiento de los modelos es una problemática que cada vez resulta ser más común. Sobre entrenar un modelo consiste en modelar un problema tratando de que el modelo explique el comportamiento de todos los puntos posibles, lo que provoca realizar inferencias sobre la población que pueden no ser correctas y deberse a cuestiones aleatorias o factores no considerados en el conjunto de datos. A continuación, se detallarán los motivos por los que los modelos previos se encuentran sobre entrenados, y en los aparatados siguientes se plantearán soluciones a este problema.

5.5.1. Generación de instancias

Los modelos de generación de instancias artificiales como SMOTE o ADASYN resultan muy útiles para lidiar con el desbalance de clases. De hecho, estos modelos se proponen habitualmente para lidiar con tareas con altísimo desbalance como lo son la detección del fraude bancario o la detección de correos electrónicos *spam* [26]. Estos modelos resuelven el desbalance, pero son muy dependientes de la calidad de los datos de los que parten; es decir, que la muestra de la clase que se quiera aumentar sea suficientemente representativa de la población real, pero a la vez que exista suficiente diversidad como para que las muestras generadas no resulten idénticas entre ellas. Este último es el problema que se encuentra en el conjunto de datos, se generan demasiadas muestras para la cantidad de datos de los que se parten, eso causa que, aunque en la validación cruzada se reserven casos reales para la validación, los casos generados por SMOTE-NC se parecen tanto a los reales que a efectos prácticos es como si se estuviera validando con muestras con las que se ha entrenado, lo que causa resultados extrañamente positivos en la validación cruzada.

5.5.2. Particiones y sobreajuste del modelo

Otra problemática asociada al sobre entrenamiento es no forzar un límite de ajuste de parámetros, lo que de forma natural opta por optimizar al máximo los modelos, incluso provocando sobre entrenamiento en ocasiones.

En el caso de *RandomForest* esto es más sencillo de visualizar, si no se especifica la profundidad de los árboles, el número máximo de hojas o el número de particiones el modelo busca reducir la entropía a toda costa, generando árboles muy complejos con particiones que, en ocasiones, disgregan apenas unas pocas muestras. Un árbol de alta

complejidad identificará patrones en los datos de entrenamiento que no necesariamente sean comportamientos reales en la población general.

5.6. Árbol de decisión simplificado

Para solucionar el problema del sobre entrenamiento se propuso reducir la complejidad de la metodología por dos vías. Por un lado, reduciendo o eliminando el proceso de generación de instancias artificiales y por otro descartando parámetros de los árboles de decisión para evitar un sobreajuste en base a una inferencia incorrecta.

De esta forma se optó por entrenar un único árbol de decisión, sin utilizar modelos de muestreo y por tanto ajustando el parámetro de peso de clase en "balanceado" para que ambas clases contaran con el mismo peso en el modelo sin importar el número de individuos en cada grupo. La profundidad del árbol y el número de hojas máximo sería ajustada manualmente, detectando particiones que dividieran un número muy pequeño de muestras. Este fue un proceso iterativo en el que progresivamente se fue reduciendo el tamaño del árbol hasta encontrar un árbol sin particiones con el riesgo de provocar sobre entrenamiento. Este proceso culminó con árboles de 3 alturas y un número máximo de 6 hojas. Una vez obtenido el tamaño ideal de los potenciales árboles, existían diversas variantes de modelos, que empleando diferentes atributos obtenían resultados muy similares. En el siguiente apartado se explicará el flujo de decisión llevado para la selección del mejor árbol de entre las variantes similares.

5.6.1. Selección del mejor árbol

El árbol escogido debía acotarse a los objetivos generales del trabajo, recordemos que estos son: interpretabilidad, máxima sensibilidad reduciendo los falsos positivos y al tratarse de un árbol de baja complejidad, evitar particiones que no sean útiles para la clasificación del CCR. Con esto hago referencia a aquellas particiones que generan una buena división en los datos, pero ambos lados de la partición producen la misma etiqueta.

De entre los árboles restantes, estos se presentaron al equipo médico y se valoró la teoría clínica detrás de estos. Este proceso se realizó con unos pocos árboles y el más atractivo desde la perspectiva analítica y médica fue el que se muestra en la siguiente figura 5.6.1 (1):

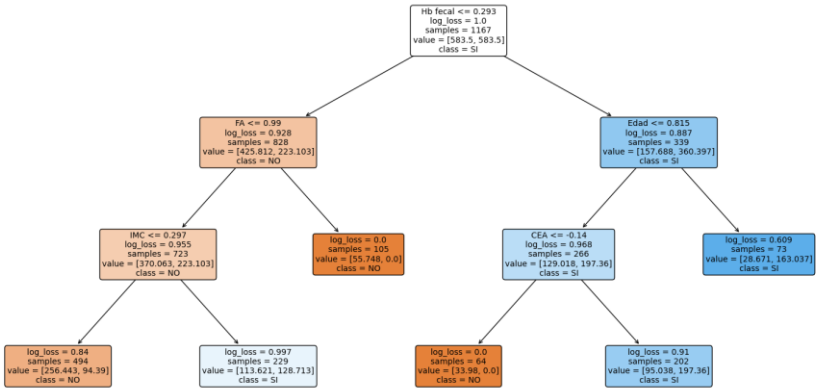


Figura 5.6.1 (1): Árbol de decisión simplificado.

La matriz de confusión sobre el conjunto de validación asociada al árbol de decisión previo es la siguiente figura 5.6.1 (2):

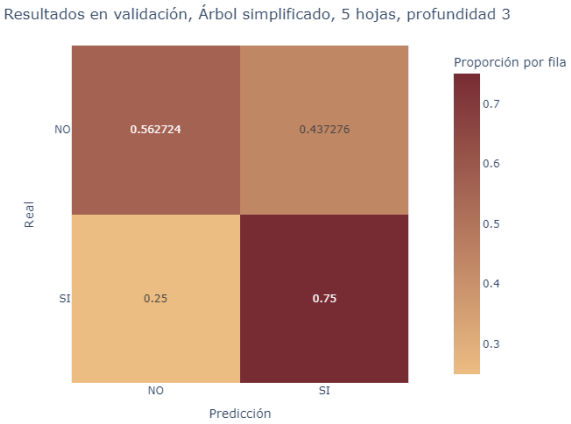


Figura 5.6.1 (2): Matriz de confusión del árbol de decisión simplificado en validación.

Estos resultados fueron muy positivos teniendo en consideración la dificultad de clasificar una enfermedad con un componente probabilístico tan alto. Sin embargo, a pesar de tratarse de un modelo con notable sensibilidad, la única problemática restante se trataba de reducir el número de falsos positivos, ya que no es asumible priorizar un 43,7% de los participantes sanos porque representaría un aumento del tiempo de espera demasiado elevado para el resto de los casos.

En cuanto a la interpretación de resultados, este modelo destaca en este aspecto. Dado que se trata de un único árbol de un tamaño reducido, los especialistas de salud pueden seguir el recorrido de cada caso por el árbol, para acabar comprendiendo la clasificación de un participante en una clase u otra. De la misma forma, se pueden observar las primeras particiones del árbol en una, dos o tres dimensiones. Un ejemplo con el punto de corte de la hemoglobina fecal se puede observar en la figura 5.6.1 (3), el resto visualizaciones relevantes se encuentran en el anexo 3.1. Es importante recordar

que para entrenar los modelos los datos se han sometido a un escalado estándar, por lo que aún es posible a los valores originales si se cuenta con la media y desviación típica originales.

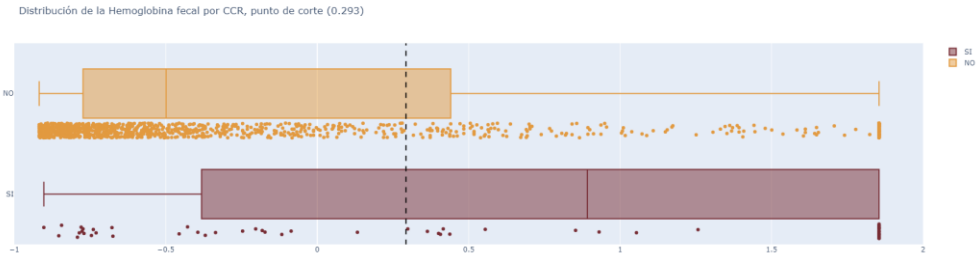


Figura 5.6.1 (3): Distribución de la hemoglobina fecal por CCR, mostrando el valor de la partición del árbol simplificado.

Los valores de puntos de corte en las variables empleadas en el modelo tras deshacer el escalado son los siguientes:

Variable	Valor	Unidad
Hb fecal	493	Ng/mL
Edad	66	Años
FA	105.5	mU/mL
IMC	28.7	---
CEA	1.36	Ng/mL

5.6.2. Ajuste de la frontera de decisión

Con el objetivo de disminuir el número de falsos positivos, el siguiente ajuste del modelo se trató de cambiar el criterio de entrenamiento del modelo y codificar que el modelo devolviera la probabilidad de CCR en vez de un valor categórico. El nuevo criterio de entrenamiento sería la pérdida logarítmica, ya que esta produce resultados más precisos en la predicción de probabilidades, y de esta forma, al contar con valores de 0 a 1, se podría ajustar la frontera de decisión entre la clase negativa y positiva.

Para ello se empleó el modelo previo de árbol de decisión simplificado en una prueba de validación cruzada múltiple, ajustando puntos de corte de 0.5 a 1. Esto provocó que en cada iteración el modelo fuera menos proclive a priorizar al participante, reduciendo los falsos positivos, pero también disminuyendo la sensibilidad del modelo. El gráfico a de la figura 5.6.2 muestra la sensibilidad del modelo frente a la proporción de falsos positivos a través de diversos valores de frontera.

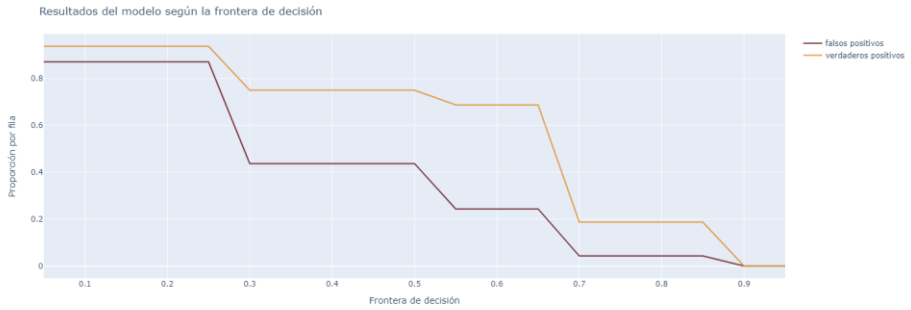


Figura 5.6.2: Falsos positivos frente verdaderos positivos según la frontera de decisión establecida.

5.7. Simplificación del muestreo

Se realizó también otra prueba empleando árboles de decisión simplificados. Esta trataba de retomar el concepto de sobre y submuestreo, pero desde una aplicación mucho más reducida de esta, combinando una reducción del 10% de muestras de la clase mayoritaria aplicando NearMiss y una generación con SMOTE-NC del 50% de casos reales. Con ello se obtuvo un conjunto de datos con un desbalance de proporción 1:10 frente al 5,74% de prevalencia del CCR en los datos originales. Esta aproximación no pretendía igualar las clases, sino aliviar ligeramente el desbalance.

Desafortunadamente, tras añadir nuevos casos y descartar otros no supuso diferencias en los árboles generados, y estos empleaban las mismas particiones con muy sutiles diferencias en los valores de algunas particiones. Los resultados en validación cruzada resultaron muy similares y dado que el árbol de este modelo resultó casi idéntico al previo las predicciones sobre el conjunto de validación fueron las mismas.

5.8. Modelo de regresión

La última prueba que se realizó en el modelado de la *pipeline* fue plantear una metodología muy distinta a lo realizado hasta el momento, pero con un razonamiento médico detrás.

A lo largo del trabajo, surgió la sospecha de que los factores que afectan al desarrollo de pólipos no necesariamente tenían por qué tener un efecto sobre la aparición de CCR, aunque sí se sabe que el desarrollo de pólipos, efectivamente, tiene un efecto sobre el cáncer. Esta sospecha se da tras comprender que, mientras que la neoplasia avanzada es una categorización humana y no un comportamiento natural, el cáncer si que puede poseer otras propiedades y características que lo distinguen de pólipos del mismo tamaño y tipología.

Hasta el momento se ha planteado la tarea reconocimiento del desarrollo de pólipos y la de CCR como una misma, en esta metodología se planteó la posibilidad de dividir el proceso en dos pasos: Predicción de desarrollo de pólipos y predicción de CCR. El modelo de predicción de cáncer podría tratarse del mismo, incluyendo las predicciones de un modelo previo con enfoque sobre el crecimiento de las lesiones en el colon. Estas predicciones podrían ser sustitutivas de la hemoglobina fecal, atributo que hasta el momento se ha empleado para representar la obstrucción intestinal que los pólipos provocan, representando con ello su desarrollo.

A pesar de que no se contaba con el número y tamaño exacto de todos los pólipos del participante, se realizó una estimación de la suma total de masa que los pólipos representaban empleando diferentes atributos obtenidos de la colonoscopia. Antes de elaborar este modelo de predicción previo, se añadió la variable compuesta al modelo de árbol de decisión simplificado que anteriormente se utilizó. De esta forma se ratificaría la utilidad de establecer un modelo para la predicción de esta variable, y que posteriormente se utilizara para la predicción de CCR. Lamentablemente, los resultados no se vieron alterados en absoluto, el árbol de decisión encontró una colinealidad muy elevada entre la hemoglobina fecal y la variable de masa total de pólipos, al ser variables correlacionadas y ser la hemoglobina fecal la que más capacidad predictiva tenía, la otra no aportaba información y el modelo se mantenía intacto.

A pesar de que los resultados no fueron óptimos, se sigue considerando una buena aproximación, a la espera de mejorar la calidad de los datos y de establecer un indicador compuesto que pueda ser más representativo del desarrollo de pólipos que la hemoglobina fecal.

6: Resultados y discusión

De entre la diversidad de modelos y metodologías que se han comprobado en la fase de modelado, el árbol de decisión simplificado de la figura 5.6.1 (1) es el modelo que mejor satisface las necesidades planteadas inicialmente, y es por ello por lo que se escogerá como el modelo definitivo para la priorización de pacientes.

6.1.Resultados metodológicos

A lo largo del proceso de modelado, se han probado una gran variedad de técnicas multimodales, este proceso de mejora iterativa que se ha seguido, ha tenido en todo momento la finalidad de encontrar debilidades y fortalezas de cada una de las diferentes aproximaciones al problema. Este ha sido el caso hasta que se encontró el modelo de árbol de decisión simplificado, el cual gracias se escogió gracias al aprendizaje obtenido al elaborar los modelos previos.

Existen una variedad de conclusiones extraíbles de este proceso, para facilitar la comprensión de resultados y conocimiento extraído, a continuación, se describirán aquellos aprendizajes que resultaron fruto de la prueba las diferentes aproximaciones.

La aproximación inicial consistía en resolver la priorización siguiendo una clasificación binaria, ya que, tras valorar el funcionamiento de la clasificación en CCR, neoplasia avanzada y sano, se detectó que esto solo agravaba el desequilibrio entre las clases y creaba una mayor complejidad, sin que necesariamente mejorara la calidad de la priorización.

Posteriormente, se plantearon modelos como SVC y RandomForest, los cuales se entrenarían tras una generación de instancias artificiales, con el objetivo de minimizar el problema del desequilibrio de clases. Esto inicialmente parecía una buena opción, pero tras la llegada del conjunto de validación y la verificación del sobreajuste de los modelos, se descartó rápidamente una generación de instancias a tal escala, dado que las muestras sintéticas resultaban ser muy similares entre sí.

Tras observar este suceso, se optó por simplificar la metodología, descartando la aplicación de SMOTE-NC, escogiendo el modelo de árbol de decisión por su interpretabilidad y reduciendo el tamaño de estos hasta un punto en el que todas las particiones fueran de utilidad para la predicción y no hubiera indicios de sobreajuste. De esta forma, se escogió el tamaño óptimo de los árboles (3 alturas y un máximo de 7 hojas) y se revisaron las alternativas de resultados similares con el equipo médico, tratando de asegurar un modelo con el máximo respaldo clínico teórico.

Finalmente, para reducir el número de falsos positivos se estudió el ajuste de una frontera de decisión superior a 0.5, de forma que el modelo fuera reticente a predecir casos positivos, reduciendo así el número de falsos positivos. En acuerdo con el equipo

médico se estableció que el máximo asumible era un 20%, por lo que se fijó la frontera de decisión en 0.7, ya que era el primer valor de frontera menor al máximo de falsos positivos establecidos. De esta forma se obtuvo el modelo definitivo: un árbol de decisión binario, de 3 alturas y 6 hojas, que emplea variables reconocidas como biomarcadores de CCR y cuya frontera de decisión se encuentra desplazada para reducir los falsos positivos.

6.2. Interpretación analítica

Los resultados del modelo definitivo tras ajustar la frontera de decisión sobre el conjunto de validación son los siguientes: figura 6.1

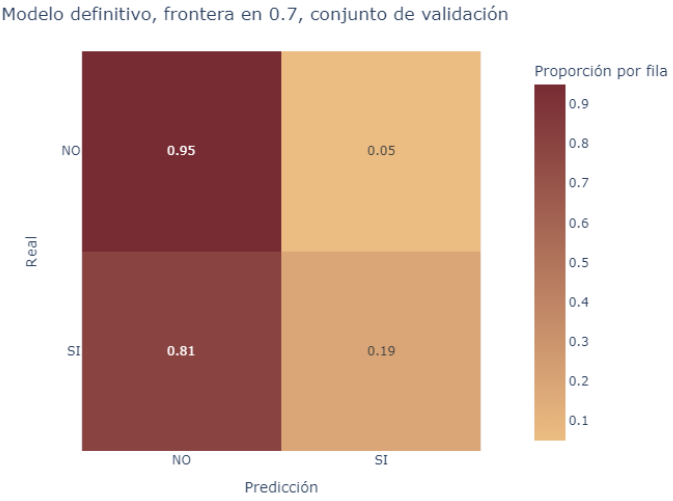


Figura 6.1: Resultados en validación del modelo de árbol de decisión simplificado, con frontera de decisión en 0.7.

Este modelo parte de la priorización del 5,4% de los participantes y logra detectar el 18,7% de los casos de CCR. Teniendo en consideración que se ha minimizado en gran medida el número de falsos positivos, el modelo tiene una sensibilidad relativamente alta. Otro factor que tener en cuenta es que no todos los falsos positivos son iguales; es decir, es preferible priorizar a un participante con neoplasia avanzada que un participante sano, pese a que ninguno de los dos sea paciente de CCR. Este es un punto fuerte del modelo, y en particular, del ajuste de la frontera de decisión, ya que el 74% de los pacientes priorizados tienen neoplasia avanzada. Esto quiere decir que habitualmente se priorizarán pacientes con un alto desarrollo de pólipos, y que una parte de ellos tendrán CCR.

6.3. Interpretación clínica

Como ya se ha mencionado, un punto fuerte del modelo es la sencillez para su interpretación por parte de un especialista del área sanitaria, para valorar los motivos de priorización de un participante tan solo hay que observar las variables que se incluyen en el árbol y realizar el recorrido por las particiones de forma manual. Por otro lado, el uso de estas variables para la priorización se encuentra respaldado por motivos clínicos. La hemoglobina fecal representa la obstrucción intestinal y por tanto el desarrollo de pólipos, un IMC elevado y una mayor edad se tratan de factores de riesgo para la enfermedad, la fosfatasa alcalina se está estudiando como biomarcador de CCR y el CEA es un marcador tumoral reconocido.

Retomando el conjunto de datos, se comprobaron los resultados del modelo sobre el conjunto de participantes medicados con antiagregantes, ya que el efecto sobre la hemoglobina fecal era menor. Los resultados del modelo en validación fueron positivos y se pueden observar en el anexo 3.2. Por lo que, pese a que el modelo no fuera entrenado con casos de antiagregantes, sí que podía resultar útil para la priorización de este grupo de individuos. Esto aumentó un poco más el colectivo de participantes candidatos a la priorización, mejorando el número total de casos de CCR prevenidos de forma temprana, dejando fuera del sistema únicamente a los pacientes bajo la medicación con anticoagulantes.

6.4. Implementación del sistema de priorización

A pesar de que la implementación final del modelo se escapa del alcance de este trabajo, es necesario dejar anotadas algunas pautas para su implementación en el sistema de colas. Este modelo necesita una capacidad de computación mínima, por lo que es fácilmente reproducible a lo largo del tiempo. Esto permite que el sistema se pueda ejecutar de forma frecuente para reconsiderar los casos priorizados y actualizar la lista de espera conforme se realicen las colonoscopias.

El equipo médico propuso una ejecución semanal, donde a los priorizados se les realizaría la colonoscopia los lunes, esto es muy útil organizativamente porque permite asegurar que cada lunes el número de colonoscopias con neoplasia avanzada o CCR es mayor, y se pueden reorganizar horarios y quirófanos acorde a ello. La ejecución del modelo puede aplicarse de forma automática a los participantes cuyos resultados se obtienen en la semana previa, o con mayor antelación, en función del tamaño de la lista de espera.

6.5. Retroalimentación y evaluación del modelo

Con la llegada de nuevos participantes y colonoscopias realizadas, si se automatiza el proceso de recolección de datos es muy sencillo reentrenar el modelo para ajustar de forma más precisa los puntos de corte de cada partición del árbol o incluso, cuando el número de casos sea suficiente, incluir nuevos atributos de partición que contribuyan positivamente a la clasificación de participantes. Se propone que, tras la aplicación del modelo, se establezca un sistema de evaluación continua del árbol en términos de analítica de datos y un sistema sencillo de revisión de casos particulares por parte de los especialistas en casos de necesidad de comprobaciones manuales.

7: Conclusiones

7.1. Cumplimiento de objetivos

Los resultados que la *pipeline* desarrollada ha producido cumplen los requisitos establecidos al inicio del trabajo, se ha podido lidiar con el desbalance de clases empleando un método como lo es un árbol de decisión ajustado a otorgar el mismo peso entre clases, se ha conseguido priorizar a un número óptimo de pacientes de CCR sin que el número de falsos positivos priorizados implique un riesgo en el tiempo de espera y el modelo definitivo está basado en biomarcadores clínicos preestablecidos, con una capacidad de interpretación más que suficiente para el grupo de especialistas que puede utilizarlo.

Más allá de los objetivos del proyecto, a nivel personal y de equipo de trabajo se ha profundizado mucho más en el conocimiento acerca de la enfermedad y del desarrollo de pólipos. Se han comprendido los factores que tienen efectos sobre el CCR y cómo las diferentes tipologías de pacientes cuentan con particularidades que las distinguen en una tarea como de priorización como esta.

Desde la perspectiva del análisis de datos, entender que en muchas ocasiones más complejidad no asegura mejores resultados ha sido una labor compleja, pero que desde lo personal considero que aporta una visión más realista del trabajo que los científicos de datos desempeñamos. Finalmente, elaborar un modelo que no solo fuera preciso sino también justo y moral ha sido un objetivo clave que se ha cumplido ampliamente.

7.2. Propuestas de mejora y trabajos futuros

Una vez concluido el trabajo, es cierto que a lo largo de él se han apreciado ciertas áreas mejorables, las cuales se van a destacar a continuación para que puedan ser consideradas en futuros proyectos.

Mejorar el conjunto de datos

El mayor foco de mejora del estudio se encuentra entorno al conjunto de datos. Hay que considerar que es habitual que estudios como este cuenten con un número reducido de casos, en particular debido al coste temporal y económico de las pruebas, de la dificultad de la recolección de datos y, especialmente, de la baja prevalencia de la enfermedad. A pesar de ello, hubiera sido deseable contar en la base de datos con más casos de CCR positivo, lo cual hubiera sido especialmente útil a la hora del modelado y potencialmente hubiera mejorado la eficacia del sobremuestreo.

Estudiar las variables genéticas

Por otro lado, en el capítulo 2, se ha mencionado que las variables de carácter genético tienen una alta capacidad predictiva sobre el cáncer, en el futuro se podría realizar un estudio específico para valorar si la mejora predictiva que supone la inclusión de estas variables podría compensar el coste de introducir una prueba genética como esta, considerando en todo momento que el objetivo es encontrar un proceso rápido, barato y eficaz de cribar según el riesgo de CCR.

Mejora de precisión en la hemoglobina fecal

Una mejora sencilla que podría haber contribuido muy positivamente en el estudio es la estandarización del aumento de precisión de la hemoglobina fecal. Como se comentó al presentar el conjunto de datos, en parte del conjunto de datos los valores de hemoglobina fecal superiores a 1000 ng/ml se codificaron como 1000 debido a una falta de precisión en la toma de datos, mientras que en el resto del conjunto sí que se pudo anotar el valor real pesa a superar los 1000ng/ml. Esta circunstancia es muy significativa dada la altísima relevancia que le otorgan todos los modelos a esta variable, usándose como nodo raíz en cualquier versión de los árboles de decisión, potencialmente una mayor precisión en esta variable también podría ayudar con la distinción de los participantes medicados con anticongelantes y posteriormente valorar su inclusión en la *pipeline* y haciendo a estos posibles beneficiarios del modelo de priorización.

Random Forest simplificado

Otra propuesta que se planteó fue el entrenamiento de un modelo de *RandomForest* sin la generación de instancias artificiales y con los parámetros empleados para el modelo definitivo. Esto se planteó ya que como bien se ha mencionado, el problema de sobreajuste se produjo debido a la generación de instancias con el sobremuestreo y del excesivo número de particiones de los árboles, no intrínsecamente del modelo de ensamblaje de árboles. Este modelo podría ser interesante, pues se podría realizar una votación combinada entre los árboles que se valoraron en el apartado 5.6.1.

Regresión para la predicción de pólipos

Una última alternativa o complemento a esta mejora podría ser una elaboración más completa de la propuesta realizada en el apartado 5.7. Esto se podría conseguir mejorando la calidad de los atributos obtenidos en la colonoscopia para obtener un indicador compuesto que sea representativo del desarrollo de pólipos. Como se ha explorado durante todo el trabajo, la hemoglobina fecal es muy relevante ya que los pólipos obstruyen el tracto intestinal aumentando el sangrado en heces. Sin embargo, esta variable genera confusión con aquellos participantes con sangrados que se deben a otras causas, encontrar una variable sustitutiva y que prevenga está confusión podría ser beneficioso.

8: Bibliografía

[1] Sociedad Española de Oncología Médica. (2022) *Las cifras del cáncer en España 2022*. https://seom.org/images/LAS_CIFRAS_DEL_CANCER_EN_ESPANA_2022.pdf

[2] Sociedad Española de Oncología Médica. (2023) *Las cifras del cáncer en España 2023*. https://seom.org/images/Las_cifras_del_Cancer_en_Espana_2023.pdf

[3] Gupta, S. (2022) *Screening for colorectal cancer*. Hematology/Oncology Clinics of North America, 36(3), pp. 393–414. doi:10.1016/j.hoc.2022.02.001.

[4] Biomarkers Definitions Working Group. (2001) *Biomarkers and surrogate endpoints: preferred definitions and conceptual framework*. Clin Pharmacol Ther. 2001 Mar;69(3):89-95. doi: 10.1067/mcp.2001.113989. PMID: 11240971.

[5] Mayo Clinic Staff (2023). *Colon polyps*. Mayo Clinic. Disponible en: https://www.mayoclinic.org/diseases-conditions/colon-polyps/symptoms-causes/syc-20352875 (Accedido: 02 Febrero 2024).

[6] Simon K. *Colorectal cancer development and advances in screening*. Clin Interv Aging. 2016 Jul 19;11:967-76. doi: 10.2147/CIA.S109285. PMID: 27486317; PMCID: PMC4958365.

[7] Your Colon or Rectal Pathology Report: Polyps (including Serrated Adenomas) | American Cancer Society. Disponible en: https://www.cancer.org/cancer/diagnosis-staging/tests/biopsy-and-cytology-tests/understanding-your-pathology-report/colon-pathology/colon-polyps-sessile-or-traditional-serrated-adenomas.html (Accedido: 07 Febrero 2024).

[8] Wilson, James Maxwell Glover, Jungner, Gunnar & World Health Organization. (1968). *Principles and practice of screening for disease*. World Health Organization. https://iris.who.int/handle/10665/37650

[9] Barnell EK, Wurtzler EM, La Rocca J, Fitzgerald T, Petrone J, Hao Y, Kang Y, Holmes FL, Lieberman DA. *Multitarget Stool RNA Test for Colorectal Cancer Screening*. JAMA. 2023 Nov 14;330(18):1760-1768. doi: 10.1001/jama.2023.22231.

[10] Pickhardt PJ, Correale L, Hassan C. *PPV and Detection Rate of mt-sDNA Testing, FIT, and CT Colonography for Advanced Neoplasia: A Hierarchic Bayesian Meta-Analysis of the Noninvasive Colorectal Screening Tests*. AJR Am J Roentgenol. 2021 Oct;217(4):817-830. doi: 10.2214/AJR.20.25416. Epub 2021 Mar 11. PMID: 33703913.

[11] Wang KW, Dong M. *Potential applications of artificial intelligence in colorectal polyps and cancer: Recent advances and prospects*. World J Gastroenterol. 2020 Sep 14;26(34):5090-5100. doi: 10.3748/wjg.v26.i34.5090. PMID: 32982111; PMCID: PMC7495038.

[12] Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

[13] McKinney, W., & others. (2010). *Data structures for statistical computing in python*. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

[14] Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12(Oct), 2825–2830.

[15] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020). *Array programming with NumPy*. Nature, 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2

[16] Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95.

[17] Waskom, M., Botvinnik, Olga, O.; Kane, Drew, Hobson, Paul, Lukauskas, Saulius, Gemperline, David C, ... Qalieh, Adel. (2017*). mwaskom/seaborn: v0.8.1* (Septiembre 2017). Zenodo. https://doi.org/10.5281/zenodo.883859

[18] Inc., P. T. (2015). *Collaborative data science*. Montreal, QC: Plotly Technologies Inc. Recuperado de: https://plot.ly

[19] Imbalaced-Learn documentation# (2024) *imbalanced-learn documentation v0.12.3*. Disponible en: https://imbalanced-learn.org/stable/# (Accedido: 20 Marzo 2024).

[20] N. V. Chawla, K. W. Bowyer, L. O.Hall, W. P. Kegelmeyer, *SMOTE: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 321-357, 2002.

[21] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. *Adasyn: adaptive synthetic sampling approach for imbalanced learning*. In 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328. IEEE, 2008.

[22] Aggarwal, C.C., Hinneburg, A., Keim, D.A. (2001). *On the Surprising Behavior of Distance Metrics in High Dimensional Space*. En: Van den Bussche, J., Vianu, V. (eds) Database Theory — ICDT 2001. ICDT 2001. Lecture Notes in Computer Science, vol 1973. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44503-X_27

[23] Inderjeet Mani and I Zhang. *Knn approach to unbalanced data distributions: a case study involving information extraction*. In Proceedings of workshop on learning from imbalanced datasets, volume 126. 2003.

[24] Brownlee, J (2021). *SMOTE for Imbalanced Classification with Python*. Disponible en: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (Accedido 20 Mayo 2024)

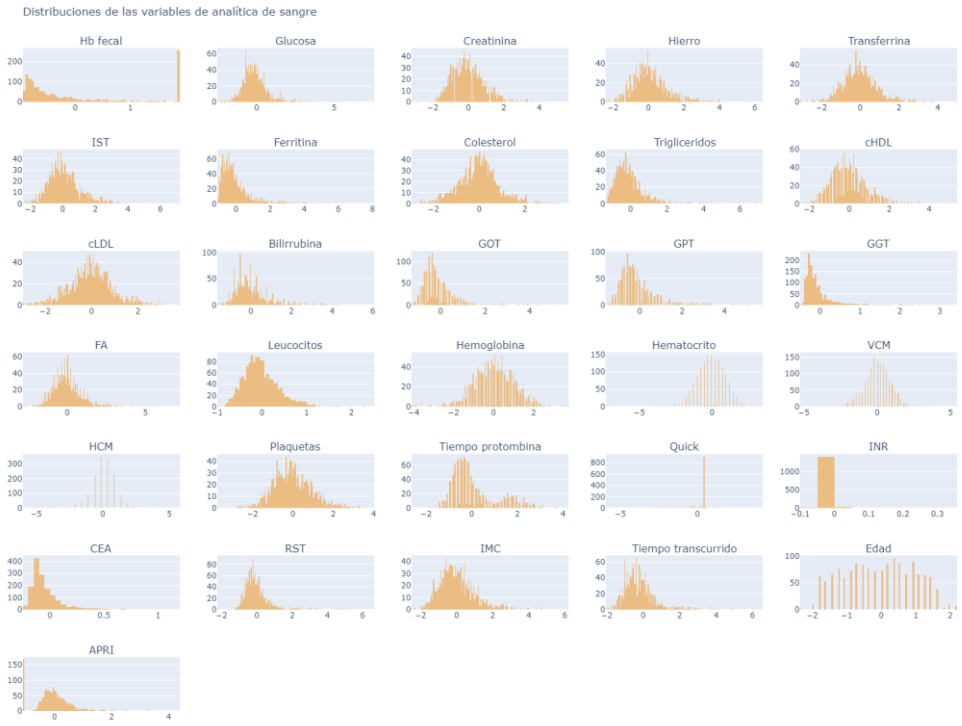
[25] Weinberg, A.I., Last, M. *Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification*. J Big Data 6, 23 (2019). https://doi.org/10.1186/s40537-019-0186-3

[26] Swastik Satpathy (2024). *SMOTE for Imbalaced Classification with Python*. Disponible en: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/ (Accedido: 22 Mayo 2024)

9: Anexos

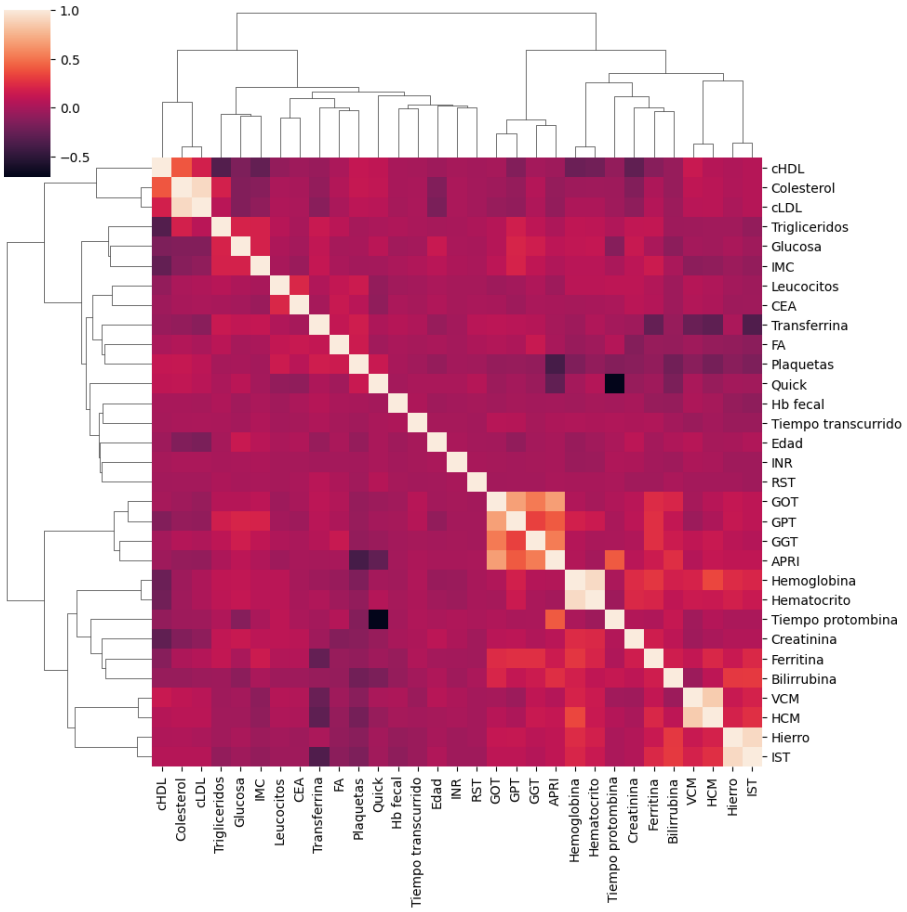
9.1.Anexo 1: Análisis exploratorio

El primer paso llevado a cabo en el análisis exploratorio fue visualizar las distribuciones de las variables numéricas de forma univariante. En el anexo 1.1 observamos que todas las variables siguen una distribución normal o una distribución asimétrica, habitualmente con asimetría negativa.



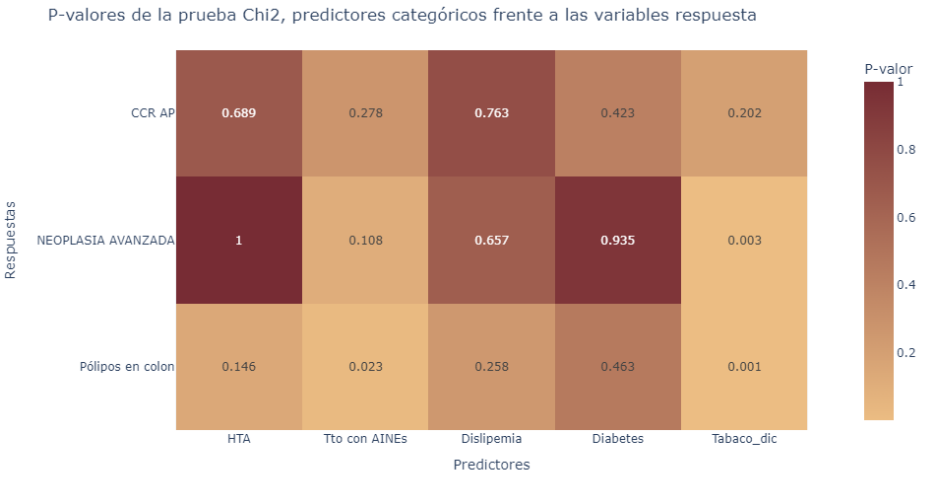
Anexo 1.1: Gráfico univariante de distribuciones de las variables numéricas.

Para observar las correlaciones entre estas variables numéricas se obtuvo una matriz de correlación combinatoria que se observa en el anexo 1.2, además se aplicó un sistema de agrupación por *clusters* siguiendo el método de Ward junto con una distancia euclídea. Las agrupaciones formadas no son especialmente relevantes, ya que se conoce que las variables analíticas tienen cierta independencia entre sí, aunque hay algunas agrupaciones lógicas como las variables referentes al hierro, como Hierro, Ferritina o Transferrina o marcadores como GGT, GPT, GOT y APRI que también forman una agrupación entre ellos.



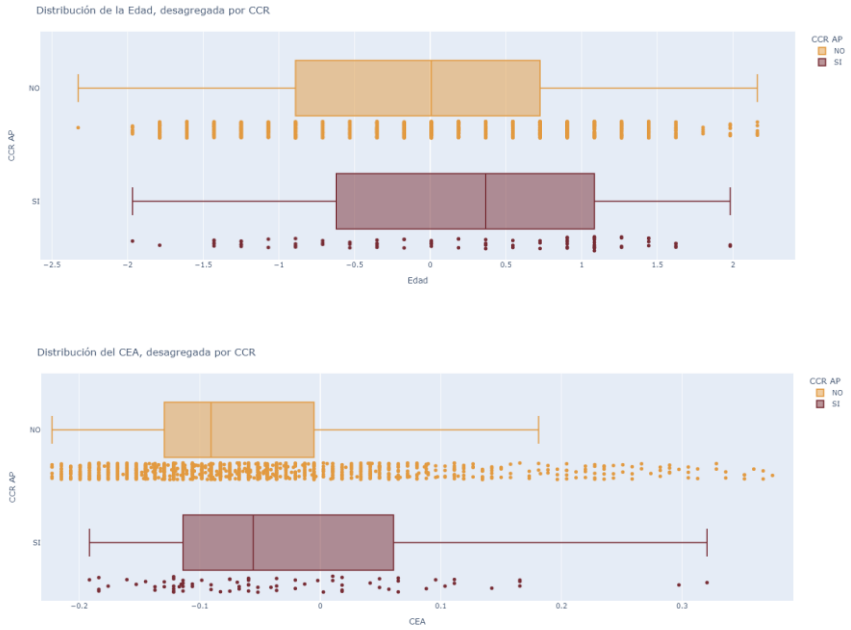
Anexo 1.2: Matriz de correlación de las variables numéricas, agrupados en según un modelo jerárquico

En cuanto a las variables categóricas, se realizó una prueba “chi cuadrado” para entender la relación entre las diferentes variables respuestas y las condiciones categóricas utilizadas como predictores. Esta matriz se observa en el anexo 1.3.



Anexo 1.3: Gráfico de p-valores de la prueba Chi-2 entre predictores y respuestas.

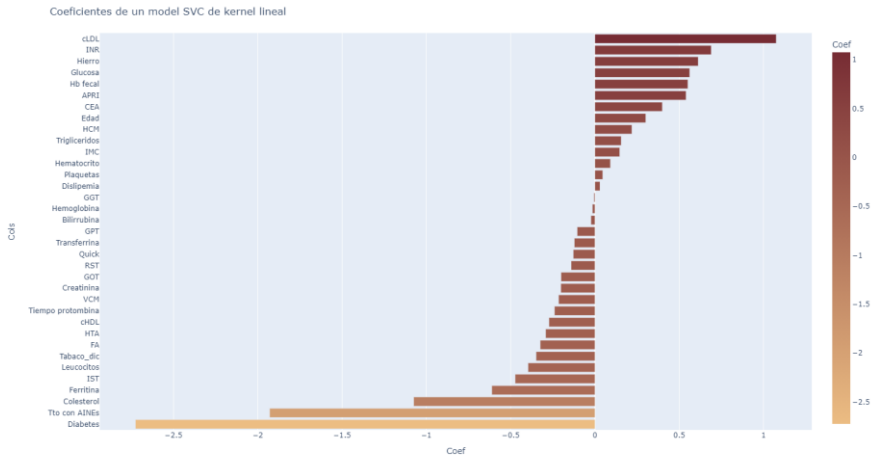
Por último, se observó la distribución de variables numéricas frente a la variable respuesta CCR tal como se hizo con la hemoglobina fecal. En el anexo 1.4 se encuentran dos visualizaciones de caja-bigotes entre el CEA y la Edad frente al CCR.



Anexo 1.4: Gráficos caja-bigotes. CEA frente a CCR y Edad frente a CCR..

9.2. Anexo 2: Modelado

Al emplear modelos de SVC que utilizan núcleos no-lineales, se pierde la trazabilidad de los coeficientes del modelo, y con ello su interpretación. Una alternativa a la interpretación es el entrenamiento de un modelo lineal para extrapolar sus conclusiones sobre el modelo no-lineal. Esto en muchas ocasiones no es una buena metodología, pero en este caso se muestra en el anexo 2.1 con tal de demostrar esta posibilidad, independientemente de su viabilidad.



Anexo 2.1: Coeficientes del modelo SVC con núcleo lineal.

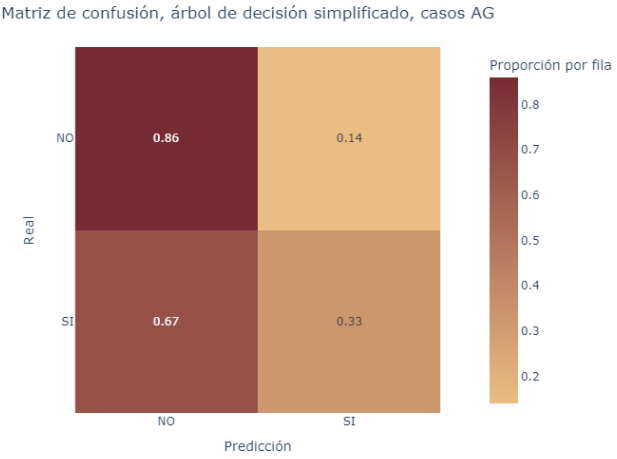
9.3. Anexo 3: Resultados

Una vez se ha establecido el modelo de árbol de decisión, se pueden visualizar los primeros puntos de corte establecidos por el algoritmo en el entrenamiento, tanto de forma univariante como se ha mostrado con la hemoglobina fecal, como de forma bivariante como se muestra en el anexo 3.1.

Anexo 3.1: Gráfico de dispersión entre las variables de cada rama del árbol.

En la primera de las visualizaciones se observa la rama izquierda del árbol. Considerando las etiquetas que se producen tras las particiones, en esta rama el modelo priorizará únicamente a aquellos participantes que se encuentren en el cuadrante superior izquierdo. De la misma forma, en la rama derecha del árbol, el modelo solo descartará para la priorización a aquellos que se encuentran en el cuadrante. Es importante considerar que estas visualizaciones no serán del todo precisas tras el ajuste de la frontera de decisión del modelo.

En cuanto a los resultados del modelo sobre el grupo medicado con antiagregantes que previamente se había excluido, en el anexo 3.2, se pueden observar los resultados del modelo al predecir estos casos.



Anexo 3.2: Resultados del modelo definitivo sobre los casos de antiagregantes.

9.4. Anexo 4: ODS

Este trabajo fomenta de forma activa el cumplimiento de diversos *Objetivos de Desarrollo Sostenible* (ODS). Estos objetivos fueron planteados por las naciones unidas con el ánimo de mejorar la calidad de vida de los individuos, la vida en sociedad y la relación con nuestro entorno. En este proyecto cobran una mayor relevancia los siguientes objetivos:

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No procede
ODS 1. Fin de la pobreza.				x
ODS 2. Hambre cero.				x
ODS 3. Salud y bienestar.	x			
ODS 4. Educación de calidad.				x
ODS 5. Igualdad de género.			x	
ODS 6. Agua limpia y saneamiento.				x
ODS 7. Energía asequible y no contaminante.				x
ODS 8. Trabajo decente y crecimiento económico.				x
ODS 9. Industria, innovación e infraestructuras.	x			
ODS 10. Reducción de las desigualdades.	x			
ODS 11. Ciudades y comunidades sostenibles.				x
ODS 12. Producción y consumo responsables.				x
ODS 13. Acción por el clima.				x
ODS 14. Vida submarina.				x
ODS 15. Vida de ecosistemas terrestres				x
ODS 16. Paz, justicia e instituciones sólidas.				x
ODS 17. Alianzas para lograr objetivos.				x

- **3: Salud y bienestar**. Este trabajo se basa fundamentalmente en el diagnóstico temprano de una enfermedad con alta mortalidad, lo que mejora significativamente su tratamiento y disminuye el impacto negativo de la misma, mejorando así el bienestar y la salud de los ciudadanos.

- **9: Industria, innovación e infraestructura.** El planteamiento del proyecto parte de una innovación en el sistema de listas de espera, aplicando metodologías de análisis de datos que no se aplican por el momento en este campo. Más allá del resultado utilitario, este trabajo supone una innovación y aprendizaje en el área de aplicación.

- **10: Reducción de las desigualdades.** Los sistemas de cribado generalizados se plantean precisamente para evitar las diferencias sociales entre los participantes con el ánimo de detectar enfermedades sobre una población general, independientemente de su poder adquisitivo, procedencia o estatus social. Este trabajo promueve la igualdad de beneficios sanitarios para todos los participantes.