Universitat Oberta de Catalunya

uoc.edu

Técnicas de Machine Learning aplicadas a la búsqueda de biomarcadores de Cáncer de Mama

Javier Pérez Córdova
Máster Universitario en Ciencia de Datos
Ciencia de datos aplicada a la Salud

Tutor: José Luis Iglesias Allones
Profesora responsable: Àngels Rius Gavidia

01/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Técnicas de Machine Learning aplicadas a la búsqueda de biomarcadores de Cáncer de Mama*
Nombre del autor:	*Javier Pérez Córdova*
Nombre del consultor/a:	*Jose Luis Iglesias Allones*
Nombre del PRA:	*Ángels Rius Gaviria*
Fecha de entrega (mm/aaaa):	01/2021
Titulación:	*Máster Universitario en Ciencia de Datos*
Área del Trabajo Final:	*Ciencia de datos aplicada a la Salud*
Idioma del trabajo:	*Español*
Palabras clave	*Minería de datos, Cáncer de mama, Machine Learning*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El cáncer de mama es el cáncer más prominente entre la población femenina con una prevalencia del 16% entre los cánceres femeninos atendiendo a datos de la Organización Mundial de la Salud (1). Aunque está ligado mayoritariamente al mundo desarrollado, las mayores ratios de mortalidad se producen en países en vías de desarrollo (69% de las muertes (1)) Esto se debe a la diferente capacidad para la detección precoz. Partiendo de esta base, en la realización de este trabajo se procederá al estudio de diversas bases de datos con valores antropométricos y valores obtenidos a partir de análisis de sangre como sería el dataset Breast Cancer Coimbra (2). Durante la ejecución se aplica la metodología CRISP-DM (3) para todo el ciclo de minería de datos realizándose un estudio exhaustivo de las diferentes variables, así como una revisión minuciosa de las diferentes técnicas de aprendizaje automático existentes y aquellas ya aplicadas al cribado de cáncer de mama, para la posterior aplicación árboles de decisión, random forest y gradient boosting machines para encontrar aquellas variables que puedan servir como diana en procesos de cribado y detección precoz de cáncer de mama. Finalmente se proporciona para uso clínico una herramienta que ayude en la toma de decisiones partiendo de la aplicación de los mejores modelos obtenidos para cada algoritmo, como un modelo random forest con valor ROC

de 79.4%, buscando así la mejora en la adherencia de los facultativos al uso de los conocimientos extraídos del análisis incentivando su confianza en los resultados.

Abstract (in English, 250 words or less):

Breast cancer is the most prominent cancer in the female population with a prevalence of 16% among all female cancers according to data from the World Health Organization (1). Although it is mainly linked to the ha developed world, the highest mortality rates occur in developing countries (69% of deaths (1)). On this basis, this work will proceed to the study of various databases with anthropometric values and values obtained from blood tests such as the dataset Breast Cancer Coimbra (2). During the execution, the CRISP-DM methodology (3) is applied for the whole cycle of data mining, carrying out an exhaustive study of the different variables, as well as a thorough review of the different existing machine learning techniques and those already to breast cancer screening, for the subsequent application of decision trees, random forest and gradient boosting machines to find those variables that can serve as

targets in screening processes and early detection of breast cancer. Finally, a tool is provided for clinical use to help in decision-making based on the application of of the best models obtained for each algorithm, such as a random forest model with a ROC value of 79.4%, thus seeking to improve the adherence of doctors to the use of the knowledge extracted from the analysis and encouraging their confidence in the results.

Índice

Lista de figuras

1. Introducción

1.1 Contexto y justificación del Trabajo

Cómo se ha comentado en el abstract el proceso de detección precoz de cáncer de mama se ha mencionado como factor clave a la hora de aumentar las tasas de supervivencia relacionadas con el cáncer de mama. Diferentes estudios (4) presentan evidencias relacionadas con a la efectividad de esta detección precoz. Podemos afirmar que es el tipo de cáncer con mayor impacto entre las mujeres y que sus ratios de mortalidad son aún bastante altos en los países en desarrollo. El método principal de detección consiste en la realización de una mamografía. Esta prueba diagnóstica de imagen se realiza siguiendo los protocolos de detección precoz (5,6). Añadido a esta vía principal, podemos observar como el desarrollo de modelos basados en datos antropométricos, sanguíneos y en caso de que fuese posible genéticos, podrían resultar en un aumento drástico de la supervivencia y la calidad de vida de los pacientes, al poder detectar casos en estadios TNM más bajos, lo que, acorde a fuente médicas, aumenta la efectividad de los tratamientos (7). Actualmente existen estudios como el que ha dado forma al dataset Breast Cancer Coimbra (2). Este estudio fu realizado en la universidad de Coimbra con la intención de buscar marcadores sanguíneos, como la glucosa o la resistina, que pudieran asociarse a gente con tumores de mama. Su principal intención no es sustituir a las mamografías, si no complementarlas como método de cribado previo. Desde nuestra perspectiva, resulta interesante la realización de una revisión de los métodos utilizados y la actualización de estos explorando el uso de modelos de agregación. Como añadido a lo anteriormente expuesto, y basado en mi experiencia personal, se añadirá un dashboard en el que se podrán introducir los diferentes valores y marcadores y que permitan al personal sanitario la utilización de los diferentes modelos, con la intención de aumentar el uso del conocimiento extraído por los estudios relacionados, así como facilitar la tarea y apoyar en su decisión al personal sanitario.

Mi motivación de cara a realizar este trabajo se basa en dos pilares. El primero es mi interés desde pequeño en la medicina y como este ha ido evolucionando hacia el apoyo a las decisiones médicas durante el estudio de la carrera, gracias al contacto con el Laboratorio de Bioingeniería y Cronobiología que tuve durante la misma y donde aprendí el gran potencial que tenía un buen uso de los datos para el tratamiento y diagnóstico de la hipertensión arterial. Más adelante en el máster de Ingeniería

Biomédica, traté el caso del cáncer de pulmón y el estudio de diferentes predictores de situación, desde antropométricos, hasta comorbilidades, pasando por genéticos, sanguíneos e incluso situacionales (exposición a radón), de cara a poder buscar los factores más influyentes en el deterioro del paciente y su supervivencia, trabajo realizado mediante la aplicación de técnica de aprendizaje automático. El segundo pilar se basa en mi deseo por enfocar mi futuro profesional hacia el campo del análisis de datos y la aplicación de inteligencia artificial en el ámbito médico, con la finalidad de poder aplicar todos los avances que se están produciendo en nuestro campo a mejorar la calidad de vida de las personas.

1.2 Objetivos del Trabajo

El Objetivo principal es la **Aplicación de modelos de aprendizaje automático para la búsqueda de biomarcadores relacionados con el cáncer de mama**

Para la correcta consecución de este objetivo, se han definido unos objetivos secundarios que apoyarán el cumplimiento del objetivo principal:

- Análisis exhaustivo del estado del arte
- Búsqueda e integración de nuevas bases de datos con valores sanguíneos y antropomórficos
- Aplicación y contraste de los modelos partiendo de los resultados de la Universidad de Coimbra.
- Creación de dashboard para la visualización, exploración de los resultados obtenidos, introducción de nuevos datos y aplicación de los modelos existentes.

1.3 Enfoque y método seguido

Como metodología de minería de datos se hará uso de CRISP-DM (3) para realizar todo el proceso de análisis de datos, definición de los diferentes casos de uso, e iteraciones necesarias, así como el despliegue final de los modelos.

En rasgos generales, el trabajo consistirá en la definición de los casos de uso, comprensión de los datos y modelos disponibles, Análisis y preselección de los datos, el modelado, la evaluación y el despliegue de una solución final.

Durante el desarrollo del trabajo, y como paso previo a la implementación de los modelos, se realizará una introducción a la problemática del cáncer de mama, ya que es importante como científicos de datos, que conozcamos el ámbito en el que estamos trabajando.

Tras esto se realizará una revisión de diferentes fuentes de datos relacionadas con el cáncer de mama, y no solo el conjunto de datos Coimbra, así como los modelos que se han sido aplicados en los artículos asociados.
Por ello, se intentará la consecución de más fuentes de datos con el objetivo de realizar un análisis más rico debido a que el dataset del que se dispone está constituido por un número de muestra que podrían resultar en problemas de sobre entrenamiento para los modelos aplicados.

La estrategia de investigación seguida será la de definir casos de estudio y resolverlos mediante experimentación haciendo uso del lenguaje R y de la librería Shiny para todo el proceso de minería de datos y la realización del Dashboard.
Del mismo modo, se ha optado por el paquete ofimático Office para la realización de la memoria y la pertinente presentación.

Finalmente, hay que señalar que partimos de una hipótesis de partida en la que queremos comprobar si los árboles de decisión, y diferentes combinaciones de estos pueden ser métodos adecuados para búsqueda de biomarcadores.
A partir de esta hipótesis, durante el estado del arte se revisarán más técnicas suplementarias que podrían ser incorporadas en el trabajo final, como por ejemplo las redes generativas antagónicas aplicadas a clasificación, que, pese a que no dejan de ser una implementación de red neuronal, se están empezando a popularizar.

1.4 Planificación del Trabajo

De cara a la correcta organización que permita la ejecución completa de este trabajo se ha optado por una organización semanal de esprints basada en metodologías ágiles (8) en la que se irá generando documentación que podrá ser proporcionada al tutor y posteriormente integrada en los entregables PEC. La principal intención de este proceder será el poder trabajar estructuradamente y partiendo el problema final en problemas más pequeños y de más fácil ejecución.

1. **HITO PEC 2 – (28/09-18/10):**
    - HITO 1 **(28/09 – 04/10):** Revisión del estado del arte relacionado con la aplicación de aprendizaje automático a la búsqueda de biomarcadores.
    - HITO 2 **(05/10 – 11/10):** Recopilación y localización de más fuentes de datos basadas en la exploración del estado del arte.
    - HITO 3 **(11/10 – 18/10):** Redacción estado del arte a partir del conocimiento adquirido.

2. **HITO PEC 3 – (19/10-13/12):**
    - HITO 4 **(19/10 – 25/10):** Exploración de las distintas fuentes de datos (Codebook parte 1)
    - HITO 5 **(26/10 – 08/11):** Preparación, integración y análisis de los datos (Codebook parte 2)
    - HITO 6 **(09/11 – 22/11):** Creación de Modelos y Evaluación de estos.
    - HITO 7 **(23/11 – 06/12):** Creación del Dashboard e integración de los modelos.
    - HITO 8 **(07/12 – 13/12):** Redacción parte de la memoria relativa al proceso de datos CRISP-DM
3. **HITO PEC 4 – (14/12-02/01):**
    - HITO 9 **(14/12 – 20/12):** Memoria v1.0
    - HITO 10 **(26/12 – 31/12):** Memoria Final

4. **HITO PEC 5 – (03/01-10/01):**
    - HITO 11 **03/01 – 10/01):** Elaboración Presentación

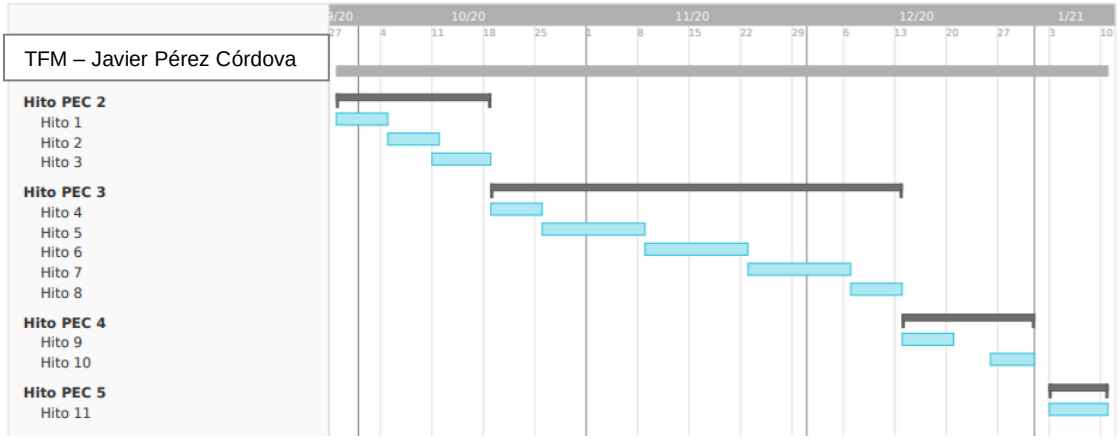

Ilustración 1: Diagrama de Gantt TFM

1.5 Breve sumario de productos obtenidos

1. Implementación de los modelos en R.
2. Dashboard operativo creado con Shiny (9).
3. Conjunto de datos curado.
4. Conclusiones y recomendaciones a partir de los análisis efectuados

1.6 Breve descripción de los otros capítulos de la memoria

En el capítulo 2 se presentará toda la información recopilada y conocimiento generado en la revisión del estado del arte. Esta consistirá en la descripción del cáncer de mama, el estudio de los diferentes biomarcadores y la revisión de métodos y técnicas.
En el capítulo 3 se englobará toda la ejecución del proyecto siguiendo la metodología CRISP-DM (3), que será explicada convenientemente en el capítulo 2. Por ello aquí encontraremos el análisis de requerimientos, la exploración y preparación de los datos, la preparación de modelos y evaluación de los mismo. Por ello y como está parte es un densa, estará repartida en subcapítulos.
En el capítulo 4 se presentarán las conclusiones extraídas de este proyecto.
Finalmente, los capítulos 5 y 6 serán los correspondientes al glosario y la bibliografía.

2. Estado del arte

2.1 Breve introducción al cáncer.

Acorde con la información de Organización Mundial de la Salud, el cáncer ha escalado durante los últimos 20 años de ser la novena causa de más muertes a la sexta en todo el mundo (10). Se estima que en 2012 ya era el causante de 8.2 millones de muertes, siendo más de la mitad en países en vías de desarrollo, y se produjeron alrededor de 14 millones de casos nuevos (11). En 2018, según el informe del SEOM de 2020 (12), los casos de cáncer ya habían aumentado hasta los 18 millones en el mundo, y el mismo informe estima que estos valores crecerán hasta los 30 millones en 2040. Para el caso de los países desarrollados, el cáncer ya es la cuarta causa de mayor mortalidad.

Antes de continuar, debemos explicar qué es el cáncer. El cáncer, según la Asociación Española contra el Cáncer (13), es una terminología que abarca más de 200 enfermedades. Pese a las diferencias que hay entre ellas, todas poseen un denominador común, y es que las células cancerosas se caracterizan por una multiplicación descontrolada y que tiende a esparcirse por el organismo.

La multiplicación celular es un proceso normal, ya que el cuerpo necesita ir renovando aquellas células al final de su ciclo de vida con el fin de asegurar el correcto funcionamiento de aquellos órganos que forman las células.

Los mecanismos encargados de asegurar la eliminación de las células viejas son típicamente muy precisos, aunque no infalibles. Acorde a varios artículos, entre los que se encuentra el escrito en Acta Sanitaria (14), diariamente desde que nacemos, tenemos células cancerosas en nuestro organismo, que se van eliminando periódicamente, aunque se puede dar el caso de que se llegue a producir un tumor.

No necesariamente la existencia de un tumor debe ser causa de preocupación, puesto que los tumores, puedes ser mayoritariamente de dos tipos (14,15):

1. *Benignos*: son aquellos cánceres que carecen de malignidad, entendiendo esta como la capacidad de invadir y destruir otros órganos.
2. *Malignos*: son aquellos cánceres con capacidad de invadir y dañar órganos, y en los que las células cancerosas pueden desprenderse y llegar a entrar en el

sistema sanguíneo, lo que se conoce típicamente como metástasis, y que suele terminar fatalmente.

Aunque los tumores benignos puedan parecer inofensivos, pueden llevar a comprometer funciones corporales, como podrían ser lesiones nerviosas en la columna vertebral por la presión ejercida en los nervios, o reducir la capacidad de las cavidades corporales, como el tórax o la cavidad cerebral.

Pese a la creencia popular, no todos los cánceres producen tumores existiendo algunos tipos de cáncer como el cáncer de células sanguíneas, el cual no produce tumores.

De cara a realizar una correcta evaluación de la situación y poder así ajustar correctamente los tratamientos, el tumor es estratificado atendiendo al tamaño, la extensión y la existencia de metástasis. Esta clasificación se conoce como TNM(16,17) y se divide en estadios del 0 al 4. Atendiendo a esta clasificación podemos decir a grandes rasgos que el estadio 0 se corresponde con la existencia de células anormales, pero sin diseminación. Los estadios 1 a 3 se corresponden con tumores malignos en los que la diferencia de gradación se basa en el tamaño y la afectación de tejidos cercanos. Finalmente tenemos el estadio 4, el cual nos indica que las células malignas se han extendido a otras partes distantes del cuerpo.

Debemos señalar la gran importancia que tiene la detección precoz (4,7) de cualquier tipo de cáncer para disminuir la prevalencia, mejorar el pronóstico, disminuir la mortalidad y evitar secuelas (6). Esta consiste y se divide en dos factores principales según la SEOM (6), la educación sanitaria y la detección selectiva.

La educación sanitaria consiste en enseñar a la población cuales podrían ser los signos indicadores de un posible tumor u enfermedad. También podríamos añadir como educación, que la gente fuese consciente de aquellos hábitos negativos y que se saben asociados al cáncer, como el tabaquismo o la obesidad.

La detección selectiva se trata de realizar pruebas a la población con el fin de detectar la enfermedad en sus fases más tempranas. Esta detección se suele hacer con cribados a poblaciones que tienen un cierto riesgo mayor de padecer una enfermedad. La división de las poblaciones se hace en base a factores antropométricos típicamente. Los cribados se efectúan siguiendo diferentes técnicas(1,18), como son

análisis de sangre en búsqueda de marcadores tumorales, análisis de fluidos corporales, diagnóstico por imagen, análisis de tejidos o búsqueda de marcadores genéticos.

De la fase de cribado/detección se hereda el diagnóstico. Una vez diagnosticado el cáncer, se procederá al tratamiento, que será ajustado en base a los resultados del seguimiento, el cual continúa típicamente durante los siguientes 5 años (supervivencia a 5 años) tras el alta médica

Una vez hemos introducido la problemática del cáncer, vamos a centrarnos en la que es objeto de este estudio, el cáncer de mama.

2.2 Cáncer de mama

El cáncer de mama es el tumor femenino más frecuente en la actualidad. En 2018, se diagnosticaron más de 2 millones de casos en todo el mundo (19). En el caso de España, el cáncer de mama representa el 30% de los tumores femeninos y se señala la población con más riesgo a las mujeres entre 45 y 65 años (19).

En términos de mortalidad, entre la población femenina española, el cáncer de mama es el que provoca mayor número de muertes, aunque bien es cierto que en los últimos años se está observando un incremento de las muertes por cáncer de pulmón.

Cómo indicamos en el apartado anterior (7), la detección precoz del cáncer de mama es clave para asegurar una elevada supervivencia, teniendo una probabilidad cercana al 100% cuando se diagnostica a tiempo.

Las asociaciones médicas se encuentran siempre en la búsqueda y optimización de las pruebas de cribado para el cáncer de mama, haciendo especial énfasis en la necesidad de prueba no invasivas cada vez más eficaces.

En lo referente a las recomendaciones para el cribado eficaz del cáncer de mama, hemos revisado las que nos proporciona la SEOM (6) y la Asociación Americana contra el Cáncer. En sus guías clínicas (5), se recomienda la mamografía a partir de los 40 años hasta los 44 como opcional, entre los 45 y 54 se recomienda anualmente, y tras los 55 años con una frecuencia bianual. En el caso de la SEOM, esta recomendación se fija a los 45 años y de los 50 a 69, mediante mamografías bianuales.

Adicionalmente al cribado periódico, se menciona, aunque la AAC (ASC) no los recomienda, la realización de exámenes clínicos a mujeres mayores de 40 años y la autoexploración, con el fin de detectar bultos anómalos y poder acudir a los profesionales sanitarios.

La asociación americana contra el cáncer también señala que estás pruebas de cribado se deben potenciar en gente que cumple ciertos factores de riesgo, como son antecedentes familiares o haber estado sometidas a algún tratamiento radiológico en la zona del tórax. Además, en (20), se nos introduce una tabla con los diferentes factores de riesgo asociados al cáncer de mama, entre los que encontramos:

- Edad
- Menopausia tardía
- Hormonas: anticonceptivos y terapias
- Genética
- Problemas benignos en las mamas
- Estilo de vida: obesidad, tabaquismo, alcoholismo y dieta
- Factores externos: alteración ritmos circadianos, radiación (radón), diabetes y contaminación.

Debemos indicar que otros muchos factores se indican como protectores o suscitan controversia, pero no son del caso de este estudio, aunque se pueden consultar en (20). También se menciona el nivel socioeconómico como factor de riesgo, pero este se ve ya mencionado en factores como el estilo de vida o el sueño, debido a las condiciones de vida típicas de la gente con menor nivel socioeconómico.

Cómo dato importante señalar el estudio (21), en el cual se compara, entre otros factores, la supervivencia de aquellos casos en lo que se hace una detección asistencial contra los que se encuentra dentro del programa de Detección Precoz del Cáncer de Mama (DPCM). En este estudio se observa una mejora dramática en la supervivencia de 8 puntos porcentuales, siendo hasta casi del 95% para casos detectados por cribado DPCM.

Hay que señalar que una vez el tumor se ha diagnosticado, los tratamientos varían desde la cirugía, donde se extirpa la parte tumoral, y que puede ser más o menos agresiva; hasta los tratamientos quimioterápicos, pasando por tratamientos radiológicos y los tratamientos más modernos basado en medicina de precisión e inmunoterapia.

Por lo tanto, hemos podido observar los diferentes métodos de cribado existentes, que van desde análisis sanguíneos, hasta datos antropomórficos, y pasando por técnicas de imagen y marcadores genéticos, lo que nos posibilitará la tarea de realizar una revisión del estado del arte exhaustiva centrada en la aplicación de la inteligencia artificial.

2.3 Aportaciones de la Inteligencia Artificial al cáncer de mama.

Según un reciente artículo publicado (22) en el que se encuestó a 500 líderes de atención médica de EE. UU, la mayoría coincidieron en la importancia de la aplicación e investigación de los métodos de Inteligencia Artificial aplicados al ámbito médico, reduciendo costes haciéndola más accesible y mejorando la experiencia del paciente, así como los resultados sanitarios. Del mismo modo, la IA puede ayudar en la automatización de los procesos a la par que ser una herramienta de apoyo a las decisiones médicas muy potente. Últimamente organizaciones como Google, mediante DeepLearning.ai, han intentado dar un empujón hacía el campo medico mediante la creación de seminarios y especializaciones resaltando la gran utilidad que pueden proveer los profesionales técnicos que se especialicen en el campo médico.

Para la revisión del estado del arte se han fijado dos pilares:

- IA centrada en detección y diagnóstico.
- Búsqueda de datos

Con el primero esperamos poder acotar la búsqueda, y con el segundo, poder satisfacer las necesidades de los métodos de aprendizaje automático encontrados, ya que la cantidad de los datos, así como su integridad, corrección y heterogeneidad son claves para conseguir sacar el máximo provecho de las técnicas y permitir que los resultados generados sean de utilidad médica real y no los debamos acotar solo a la muestra que tenemos.

2.3.1 Enfoque en los datos tipo numérico y texto (Num-Txt)

Para realizar la revisión bibliográfica siguiendo este enfoque, lo primero ha sido detectar diferentes conjuntos de datos relacionados con el cáncer de mama disponibles en la red y que se caracterizan por ser medidas o valores de atributos, con la intención de a partir de ellos poder acceder a los artículos que los citan y poder acceder así al conocimiento generado en cada uno. A continuación, se enumerarán y presentarán las conclusiones sacadas a partir de cada uno de los datasets.

Dataset Coimbra (2,24): este conjunto de datos se encuentra en el repositorio UCI y consiste en 116 registros de pacientes con información sobre valores antropométricos y análisis de sangre. Data del 2018 y corresponde con el paper análogo de la universidad de Coimbra. Por lo tanto, con este dataset se podrán enfrentar problemas relacionados con la detección y el diagnóstico, a la par que comparar diferentes técnicas predictivas.

En el paper de referencia sobre este dataset (2), los autores proponen el uso de técnicas como support vector machines (SVM), regresión lineal (RL) y random forest (RF) para ver que atributos del dataset podrían ser usados complementariamente a una mamografía para la detección del cáncer de mama, obteniendo una sensibilidad entre el 82% y 88% y una especificidad entre el 85 y el 90% con una confianza del 95% para los predictores Glucosa, Resistina, Edad e IMC del dataset.

Otros autores como (25), basan su experimentación los modelos combinados (ensemble Machine Learning) para aumentar la capacidad predictiva, obteniendo resultados muy cercanos al 100% para modelos basados en árboles de decisión (DT) y KNN. Otros artículos como (26), también hacen uso de diferentes configuraciones de árboles de decisión.

En (27) se plantea el experimento en el que comparar 8 técnicas de machine learning con el objetivo de ver si estas se comportan diferentemente a la hora de trabajar sobre el dataset. Las técnicas que se proponen comparar son regresión logística (LR), SVM, KNN, y una serie de método basados en árboles como DT, RF, y tes métodos basados en el boosting de árboles, tanto adaptativo como por gradiente. Este artículo se centra simplemente en proponer un experimento, pero no llega a ejecutarlo.

En (28) se aplican redes neuronales (NN) con intención predictiva comparándolo con Naïve Bayes y obteniendo precisiones cercanas al 87% para las redes neuronales y al 84% para el segundo método

En (29) se aplican redes neuronales multicapa con unos valores de especificidad y sensibilidad superiores al 95% tanto para 4 como 9 predictores y con un AUC del 96%.

En (30) comparan 5 técnicas, DT, SVM, RF, LR y NN aunque con unos resultados muy pobres entornos al 70% de precisión y de valor F

Finalmente, y para terminar la revisión en lo referente a este dataset, en (31) se comparan diferentes técnicas entre las que hay variantes de la regresión lineal, de SVM, métodos basados en árboles y modelos estadísticos como KNN y Naïve Bayes, obteniendo para todos los casos valores por debajo del 72% de precisión, menos en el caso de KNN y Naïve Bayes, donde el rendimiento se encuentra entorno al 58% y el 62 % respectivamente.

Por lo tanto, de la revisión del dataset Coimbra se extraen como métodos más prometedores para la clasificación los basados en redes neuronales y los basados en árboles de decisión, así como los métodos basados en agregación de modelos.
También se observa como el rendimiento mejora al tener en cuenta solo los 4 primeros predictores (glucosa, resistina, IMC y Edad).

Dataset CancerSEEK (32): en este caso se nos presenta un conjunto de datos de 1005 pacientes que padecen algún tipo de cáncer con información referente a mutaciones y valores de proteínas circulantes en sangre asociados a los diferentes tipos de tumor y los estadios en los que se encuentran los pacientes. Para ello hacen uso de esta información con el objetivo de encontrar marcadores tumorales. Se debe indicar que este dataset no tiene únicamente información sobre pacientes con cáncer de mama, si no que se presentan otros tipos como cáncer de ovario o cáncer de estómago. En el paper en cuestión se comenta que las sensibilidades varían entre el 69 y el 90% para los tipos de cáncer estudiados y se menciona el uso de regresión logística.
En (33) podemos ver otra aproximación al modelado de estos datos con resultados similares haciendo uso NN.
En (34) se propone otra forma de actuar que promete aumentar la sensibilidad hasta el 70%, con una especificidad del 99% para etapas tempranas de cáncer de mama. En el estudio se realiza la comparación de modelos de Deep learning, Naïve Bayes, DT, el original y los modelos CancerA1DE (35) y Cancer A2DE (36), que son una combinación de clasificadores basados en Naïve Bayes y árboles, y que obtienen los mejores resultados con mucha diferencia.
En este caso volvemos a observar el gran potencial de los modelos de árbol agregados y Naïve Bayes, así como las NN y SVM.

Dataset Ljubljana (37): este conjunto de datos también forma parte de la página web UCI y data del 1988. En el podemos encontrar valores antropométricos y tumorales relacionados con tamaño y nodos de 286 pacientes.

En (38), se presenta el estudio de combinación de diferentes modelos basados en redes neuronales desde un punto de vista matemático con la intención de probar si los métodos de aprendizaje de Correlación negativa son eficaces para a la agregación de modelos; mientras que en (39) se prueban diferentes configuraciones de DT.

Dataset Wisconsin (40,41): en este dataset tenemos información sobre 699 con información citológica del tumor, es decir, información de la forma y tamaño de las células tumorales, con lo que se esperaba poder discernir si un tumor era benigno o maligno. En este caso estaríamos más ante un problema de diagnóstico en base a los valores numéricos obtenidos con el fin de, una vez localizado el tumor, poder discernir entre si es benigno o maligno.
En (30) se comparan 5 técnicas, DT, SVM, RF, LR y NN con unos muy buenos resultados en torno al 95% de precisión y con unos valores de F similares

2.3.2 Enfoque en los datos de imagen

En este apartado revisitaremos algunos de los conjuntos de datos de mamografías disponibles, ya que es interesante revisar el campo de la imagen médica, ya que la mamografía es la principal prueba utilizada para el diagnóstico del cáncer de mama. Los tres datasets que presentaremos están pensados para ayudar en la mejora del diagnóstico mediante técnicas de procesado de imagen.

Dataset DDSM (42–43): este conjunto de datos consta de más de 160 GB de fotos imágenes en formato DICOM. En (42), se nos presenta la explicación de uso del dataset en cuestión.

Dataset MIAS(45): este dataset consta de 420 MB de información sobre imágenes médicas. En el artículo (46), se muestran dos aproximaciones de segmentación computacional comparadas con técnicas basadas en árboles e histogramas, llegando a la conclusión de la que la segmentación tumoral mejora levemente hasta el 91% de precisión

Dataset OPTIMAM (47): este conjunto de datos se ha producido gracias a un esfuerzo del Cancer Research UK, y se compone de imágenes de mamografía.

En conjunción con el DDSM se realizó el estudio (48) en el que se compitió entre radiólogos de US y UK y algoritmos de Deep Learning para la segmentación. Al final del estudio se observó como la máquina fue capaz de mejorar en ambos datasets las ratios de falsos positivos y negativos, siendo está mejora enorme (5.7 y 9.4%) para el caso americano. Esto nos muestra el enorme potencial que tiene el procesamiento mediante computador aplicado al diagnóstico de cáncer.

2.3.3 Comparativa y análisis de las fuentes de datos

Una vez vistos los diferentes conjuntos de datos podemos realizar una comparativa de los mismos con el objetivo de poder aislar cuáles serán las soluciones que más nos podrían interesar.

Dataset	Tipo de datos	Antropométrico	Sanguíneo	Genético	Tumoral
Coimbra	Num-Txt	✓	✓	✗	✗
CancerSEEK	Num-Txt	✓	✓	✓	✗
Ljubljana	Num-Txt	✓	✓	✗	✓
Wisconsin	Num-Txt	✗	✗	✗	✓
MIAS	Imagen	✗	✗	✗	✓
DDSM	Imagen	✗	✗	✗	✓
OPTIMAM	Imagen	✗	✗	✗	✓

Tabla 1: Comparativa uso de datos estudiados

Cómo podemos observar en la Tabla 1, los datasets de Coimbra y CancerSEEK podrían ser utilizados para nuestro objetivo, que se centra en la detección precoz del cáncer de mama, aportando dos enfoques distintos.

Por el contrario, Ljubljana y Wisconsin están más centrados en la mejora del diagnóstico debido a que se centran en características tumorales.

Finalmente, tendríamos los datasets MIAS, DDSM y OPTIMAM, centrados en el análisis mediante visión por computador para la detección de los tumores en mamografías.

Cabe señalar la poca profundidad de los conjuntos encontrados para el análisis sanguíneo y antropométrico, lo que resalta las dificultades de cooperación que se encuentran sanitarios e ingenieros a la hora de unir fuerzas.

En este trabajo partiremos del estudio de los datos de Coimbra, aunque es urgente y necesario la adquisición de más conjuntos de datos con el fin de realizar unos análisis más profundos.
La experiencia nos indica que las colaboraciones con hospitales son complicadas a la hora de materializarse y necesitan de mucho tiempo hasta su final ejecución, por lo que, en caso de no encontrar más datos, valoraremos la inclusión del dataset CancerSEEK en nuestro estudio.

2.4 Metodología y modelos propuestos para este trabajo.

A partir de la revisión bibliográfica hemos podido entrar en contacto con diferentes modelos aplicados a cáncer de mama, lo que nos ha permitido seleccionar los que van a ser utilizados para el posterior apartado de experimentación, así como metodología que vamos a seguir.

Hemos observado en la revisión bibliográfica como se comparaban diferentes técnicas como SVM, RF, NN o KNN (31), y como en (27) se proponen hacer uso de algoritmos basados en Arboles de decisión con las variantes derivadas de la aplicación de técnicas de agregación, por lo que nuestro principal objetivo será la aplicación de estos métodos para búsqueda de biomarcadores, y comparar su rendimiento con lo ya presente en la literatura.

Por ello serán introducidas en este apartado y referenciadas al libro base que se seguirá para la aplicación de los modelos (49,50), en caso de que el lector quiera conocer más sobre las mismas

2.4.1 Árboles de decisión (DT)

Esta familia de algoritmos es muy interesante debido a la interpretabilidad de la división que nos aporta en base a los datos que se usan. El objetivo de esta técnica es dividir y clasificar la información en grupos homogéneos.

Para realizar esta división y hacer crecer el árbol se podría usar el error de clasificación, pero acorde a la literatura, el rendimiento de este se muestra ineficiente para aumentar la homogeneidad por lo que se recomiendan dos técnicas, el Índice de Gini y la Entropía.

El índice de Gini se encarga de medir la varianza entre las clases, es decir, la pureza de un nodo, que es mayor cuanta mayor sea la diferencia entre el número de muestras de cada clase.

Análogamente, la entropía funciona de una manera muy similar numéricamente, mostrando también la pureza de los nodos.

Hay que indicar que cuanto menor sea el valor de estos dos marcadores, mayor será la pureza.

Si nuestro objetivo fuese mejorar el desempeño de la predicción final, lo mejor sería optar por optimizar el error de clasificación.

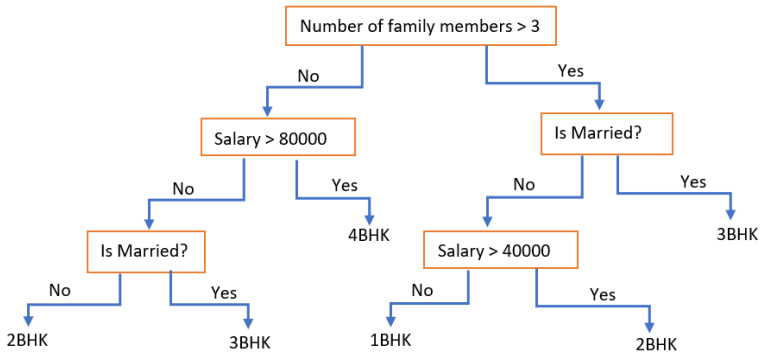

Ilustración 2: Ejemplo de árbol de decisión

$$\sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Ecuación 1: Entropía

$$1 - \sum_{i=1}^{C} (p_i)^2$$

Ecuación 2: Gini Index

2.4.2 Ensemble Learning - Random Forest (RF)

Los random forest son una implementación mejorada de los métodos de bagging para Árboles de decisión.

Primeramente, procedemos a definir en que consiste el bagging. Esta es una técnica, cuya intención reside en reducir la varianza del método de aprendizaje estadístico. Se basa en el concepto de promediado de observaciones, el cual lleva a una reducción de la varianza. Para ello los métodos de bagging toman repetidamente muestras del conjunto de datos original, generando pequeños conjuntos que luego son entregados al modelo. Seguidamente se entrena el modelo, en este caso tantos modelos como conjuntos, con estos conjuntos y se promedian las predicciones. Estos árboles se dejan crecen hasta el final, no realizando un proceso de poda, el cual consiste en definir una longitud máxima de los árboles.

A partir de lo comentado anteriormente, debemos señalar que esta técnica es muy prometedora al usar pequeños estimadores de bajo rendimiento con el fin de crear un modelo final que tenga una baja varianza y con una capacidad predictiva mucho mayor que la obtenida haciendo uso de árboles de decisión individuales.

En problemas de clasificación, el bagging tomará como la clase a dar como predicción aquella que haya obtenido el mayor número de votos u apariciones entre los pequeños árboles que confeccionan el modelo final.

Random Forest introduce una mejora basada en que, en cada división de los árboles, obliga al estimador a usar solo un subconjunto de los predictores, haciendo que haya casos en los que el predictor más potente no sea tomado en cuenta. De esta forma se puede reducir la correlación entre los árboles obtenidos, reduciendo la variabilidad de los árboles y obteniendo muchos mejores resultados cuando estamos haciendo uso de predictores correlacionados entre sí.

El algoritmo de funcionamiento es el siguiente:

- El modelo selecciona una muestra aleatoria de los datos (conjunto bootstrap) y construye un árbol de decisión siguiendo los criterios de división de los árboles de decisión.
    - Para cada nodo de división seleccionar aleatoriamente el subconjunto de variables explicativas, que son usadas en exclusiva para decidir la división
    - Se generan nodos hasta llegar al tamaño mínimo de nodo.

- El proceso se repite n veces hasta construir el forest (bosque) que será utilizado para generar la predicción

2.4.3 Ensemble Learning - Gradient Boosting Machines (GBM)

En este caso nos encontramos antes una implementación de los métodos de boosting. Pese a ser bastante similares a los métodos de bagging, en este caso, cada árbol se crea secuencialmente usando la información obtenida de los residuos de los árboles anteriores, y se usa un conjunto de datos modificado para cada árbol.
Los métodos de boosting tienen, a parte de los parámetros de número de árboles como los Random Forest, el parámetro shrinkage que se encarga de proporcionar la ratio de entrenamiento.
Los GBM, se basan en estos conceptos y usan la técnica de descenso del gradiente con la intención de optimizar una función de coste conforme se van creando los árboles, con la intención de saber cómo de bueno es el modelo.

El algoritmo de funcionamiento es el siguiente:

- El modelo selecciona una muestra aleatoria de los datos (conjunto bootstrap) y construye un árbol de decisión siguiendo los criterios de división de los árboles de decisión hasta el número mínimo de tamaño del nodo.
- El proceso se repite n veces de forma secuencial hasta minimizar la función de coste

2.4.4 Métricas de evaluación y técnicas de entrenamiento.

- Sensibilidad (Sens): indica la probabilidad de clasificar correctamente un verdadero positivo. También conocida como ratio de verdaderos positivos o TPR.

$$\text{Sensibilidad} = \frac{VP}{VP + FN}$$

Ecuación 3: Sensibilidad

- Especificidad (Espec): indica la probabilidad de clasificar correctamente un falso negativo. También conocida como ratio de falsos positivos o FPR.

$$\text{Especificidad} = \frac{VN}{VN + FP}$$

Ecuación 4: Especificidad

- ROC/AUC: métrica de evaluación de modelos clasificadores binarios que tiene en consideración la sensibilidad y la especificidad. Valores cercanos a 0.5 indican un modelo clasificador muy malo, mientras valores cercanos a 1 indican que el clasificador tiene un desempeño muy bueno. La ROC es la representación gráfica de la AUC, mientras que la AUC nos proporciona valores numéricos mediante el uso de la sensibilidad y (1-especificidad), definiendo así los valores del eje Y y del eje X.

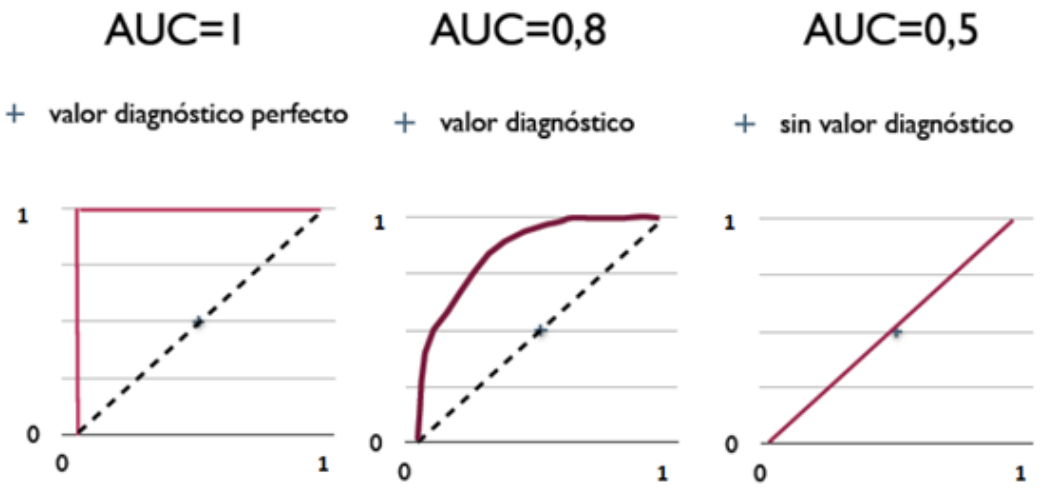

Ilustración 3: Curvas ROC y valor AUC

- Accuracy (ACC): también conocida como exactitud en castellano usada para medir la calidad de una clasificación binaria.

$$\text{exactitud} = \frac{VP + VN}{VP + FP + FN + VN}$$

Ecuación 5: Accuracy

- Youden Index: Medida realizada a partir de la sensibilidad y la especificidad

$$\text{Especificidad} + \text{sensibilidad} - 1$$

Ecuación 6: Youden Index

- Seed: valor que debe ser definido en el entorno de programación cuando nos encontremos ante la ejecución de modelos que contengan aleatoriedad en sus decisiones intermedias, como en este caso ocurre con GBM y RF.

- Validación Cruzada (49,51) (CV): Las técnicas de validación cruzada para el entrenamiento son esenciales en cualquier proceso de aprendizaje automático ya que permiten evaluar el entramiento de los modelos cuando tenemos una muestra limitada. De cara a la realización de este trabajo han sido tenidas en consideración dos técnicas de validación cruzada.

    - K-Fold CV: es un tipo de validación cruzada (CV) que consiste en dividir los datos de entrenamiento estratificada en k, siendo k el número de particiones en los datos que serán utilizadas para entrenar y testar los modelos y por lo tanto el número de iteraciones que hará nuestro algoritmo. De esta forma se realiza el entrenamiento de una forma más robusta y con una menor dependencia de los datos.
    - LOOCV: Debemos señalar que existen otras técnicas para realizar la validación cruzada que se ajustarían más al problema como la presente y que consiste en ir dejando una muestra fuera del conjunto en cada iteración. Esto nos hace tener n-1 iteraciones, siendo n la cantidad de registros disponibles en el conjunto de entrenamiento.

2.4.5 Metodología de Enfoque del Trabajo - CRISP-DM

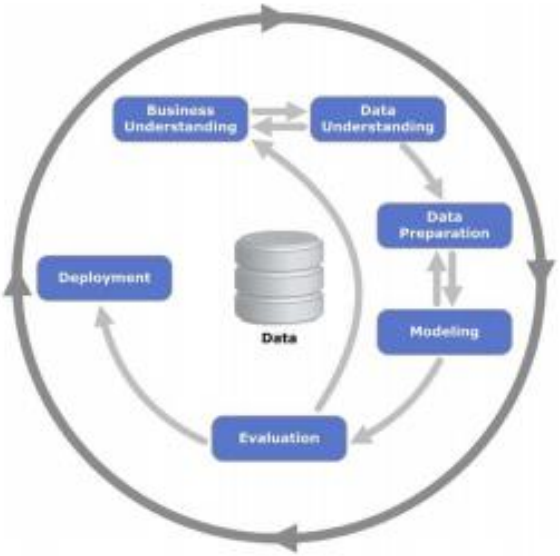

Ilustración 4: Proceso CRISP-DM

Estamos ante una metodología iterativa, y la más extendida en el sector, la cual consiste en 6 pasos (3).

1. **Entendimiento del negocio**: en esta fase se analiza el problema y los objetivos para ser traducidos en objetivos de negocio. En nuestro caso serían similares a los objetivos fijados para este trabajo
2. **Entendimiento de los datos**: fase en la cual se analizan las fuentes de los datos, la calidad y la integridad. También es necesario familiarizarse con los datos y conocer si van a ser suficientes para cumplir con los objetivos, y en el caso necesario y que sea posible, buscar nuevas fuentes de datos.
3. **Preparación de los datos**: aquí nos debemos encargar de la corriente limpieza y preparación de los conjuntos de datos que será introducidos a los diferentes modelos.
4. **Modelado**: fase en la que se preparan los modelos y se ajustan conforme a los resultados de evaluación para poder cumplir con los objetivos de una forma eficientes.
5. **Evaluación**: dónde se comprueba la calidad de los modelos y se revisa que cumplan con los objetivos
6. **Despliegue**: donde se proporciona la solución final.

Debemos indicar que las fases 3 a 5 son iterativas entre sí. Del mismo modo adoptaremos un proceder Agile para las fases 4 y 5, con el fin de acelerar la generación y evaluación de los diferentes modelos (8)

2.5 Conclusiones para el plan de trabajo

Como resultado de la revisión del estado del arte, se ha podido observar como el ámbito de estudio de la detección y diagnóstico del cáncer de mama están a la orden del día.

Han sido observados los diferentes enfoques propuestos, los cuales hemos dividido en análisis de imagen para la detección de tumores, análisis de características tumorales con el fin de determinar su naturaleza y el análisis de datos antropomórficos, genéticos y de valores en sangre.

Se observa del mismo modo, como la cantidad de datos disponible de forma abierta es bastante escueta, dejando en evidencia las dificultades de colaboración entre los diferentes actores involucrados en un proceso de minería de datos médicos.

Pese a que nuestra intención sea la de poder ampliar el conjunto de datos disponible, la situación actual que estamos viviendo nos dificulta, aún más si cabe, las tareas de búsqueda de nueva información. Debido a esto nos centraremos en el conjunto de datos Coimbra. Esto nos permitirá comparar nuestras modelos con los modelos del artículo original y con las mejores soluciones que hemos observado basadas en modelos agregados (ensemble learning) y redes neuronales.

Adicionalmente, y si el tiempo lo permite, buscaremos aplicar los mismos modelos disponibles al dataset cancerSEEK centrado en las proteínas sanguíneas y las mutaciones genéticas, ya que los resultados observados en los diferentes estudios revisados se muestran prometedores, al nivel de los revisados en el dataset Coimbra.

Por lo tanto, en este trabajo se intentará profundizar en los siguientes aspectos:

1. Estudio y tratamiento de los datos disponibles
2. Preparación de modelos basados en árboles de decisión.
3. Comparación con las técnicas más prometedoras arrojadas por el estudio del estado del arte.

4. Optimización de los modelos y conjuntos de datos propuestos.

En este capítulo ya hemos realizado la primera etapa de la metodología CRISP-DM. En los siguientes capítulos ejecutaremos las restantes fases que culminaran en la presentación de los modelos obtenidos y su puesta en desarrollo mediante un dashboard ejecutado con shiny-R (9) que permitirá la introducción de nuevos datos y la revisión de los originales, con la intención de poder atajar la brecha que hemos observado entre el personal sanitario y la gente de nuestro ámbito de acción, mediante el acercamiento de estas herramientas.

3. Experimentación

En este apartado se procederá a explicar el problema ante el que nos encontramos y como prevemos afrontarlo, siguiendo siempre los principios de la metodología CRISP-DM.

Del mismo modo se hará un análisis de los datos disponibles con la intención de realizar una breve aproximación descriptiva.

Acto seguido se preparan los conjuntos de datos que van a ser utilizados para entrenar los modelos.

El siguiente paso consiste en poner a punto los modelos de referencia y realizar la optimización de estos.

Una vez obtenidos los modelos, se presenta el dashboard mediante el cual los profesionales puedan interactuar con el mejor modelo y añadir nuevos datos, hecho que se detallará más es su respectivo apartado.

Finalmente se compara el desempeño de nuestros modelos con los observados en el estado del arte.

Toda la experimentación presentada ha sido realizada mediante el uso del lenguaje estadístico R mediante una libreta de rmarkdown, el cual es el lenguaje de marcado para la realización de informes dentro del lenguaje R, la cual se adjunta a esta memoria estando disponible en un repositorio de github, el cual tiene un readme donde se explican los diferentes ficheros.

Repositorio: https://github.com/Javipercor/TFM_UOC

3.1 Comprensión del problema

Nuestro problema consiste en encontrar patrones que permitan clasificar personas en afectadas por cáncer de mamá o no atendiendo a datos antropométrico y valores observados en analíticas.

Para la resolución de este problema contamos con datos recogidos en el hospital de Coimbra entre el 2009 y el 2013 y los cuales hacen referencia a la edad y los valores del índice de masa corporal, glucosa, resistina, insulina, HOMA, leptina, adiponectina y MCP-1.

El problema por analizar se divide en 3 partes:

- Estudiar el poder predictivo de los atributos disponibles con la intención de apoyar el diagnóstico de cáncer de mama.
- Derivado de la extensa bibliografía y pruebas que existen entorno al modelado de estos datos, la utilización de métodos de aprendizaje automático aplicados

con bajo desempeño o simplemente planteados, con el fin de continuar con el estudio del campo que algunos autores dejan abierta.

- Analizar y desarrollar una herramienta primitiva que facilite el uso de los modelos generados por este tipo de iniciativas por parte del personal médico, así como la introducción de nuevos datos.

A continuación, y siempre siguiendo la metodología CRISP-DM, se procede a listar las conclusiones obtenidas a partir de la recolección de objetivos de negocio, u objetivos de trabajo definidos en el apartado 1.2, con el fin de obtener los propósitos de data mining finales que serán respondidos durante la elaboración de este trabajo:

- **ON 1 – Estudio de los biomarcadores aplicables a cáncer de mama**, el cual afrontaremos mediante el **DM-1: estudiar las relaciones entre los datos disponibles**
- **ON 2 – Creación y comparación de algoritmos aplicados a la búsqueda de patrones**, para el cual definimos el **DM-2: Aplicación de modelos de clasificación basados en árboles** y el **DM-3: Comparación de algoritmos propios y presentes en la bibliografía.**
- **ON 3 – Proporcionar al personal sanitario de una herramienta para la explotación de los modelos creados**, el cual trataremos en el **DM-4: Creación de dashboard para la consulta e introducción de datos médicos**.

En la siguiente tabla se presentan los objetivos para facilitar su revisión.

Objetivo de Negocio	Objetivo de Minería de Datos
ON 1 – Estudio de los biomarcadores aplicables a cancer de mama	**DM-1: estudiar las relaciones entre los datos disponibles**
ON 2 – Creación y comparación de algoritmos aplicados a la búsqueda de patrones	**DM-2: Aplicación de modelos de clasificación basados en árboles**
	DM-3: Comparación de algoritmos propios y presentes en la bibliografía.
ON 3 – Proporcionar al personal sanitario de una herramienta para la explotación de los modelos creados	**DM-4: Creación de dashboard para la consulta e introducción de datos médicos**

Tabla 2: Tabla resumen Objetivos de Negocio – Objetivos de Minería de Datos

3.2 Comprensión de los datos

3.2.1 Adquisición de datos

Previo a la comprensión de los datos que se han podido obtener para la realización de este trabajo se deben repasar las diferentes fuentes de datos disponibles. Dos casos se presentan a partir de esta hipótesis, repositorios online que ya han sido tratados y estudiados en el apartado de estado del arte, y de los cuales se ha seleccionado el Dataset de Coimbra. Otra fuente de datos son los hospitales. A través de los tutores de este TFM se han mantenido contactos con organizaciones y hospitales para la consecucioón de datos. Por mi parte, me he puesto en contacto con el Complejo Hospitalario Universitario de Santiago de Compostela, con el área de cardiotoxicidad, con el fin de poder conseguir datos relacionados con pacientes afectadas por el cáncer de mama. El problema en este caso ha sido que no se estaban recogiendo los mismos biomarcadores sanguíneos que en el estudio de la universidad de Coimbra, por lo que no nos ha sido posible cerrar una colaboración con ellos para aumentar el conjunto de datos disponibles.

3.2.2 Procedencia y exploración de datos

A continuación, se pasa a definir el dataset de coimbra. Este Dataset está integrado por datos referentes a 116 pacientes mujer, con información sobre edad y los valores del índice de masa corporal, glucosa, resistina, insulina, HOMA, leptina, adiponectina y MCP-1. A parte de estos 9 atributos numéricos descriptivos, también contiene un atributo más, Clasification, que nos indica si se trata de alguien enfermo de cáncer de mama o no. Estos datos han sido recogidos en el Centro Hospitalar e Universitário de Coimbra durante los años 2009 a 2013. La distribución entre clases es de 52 pacientes sanos por 64 pacientes con cáncer. En el caso de los pacientes afectados con cáncer, han sido diagnosticados mediante mamografía y se han introducido en el estudio aquellas voluntarias que no habían sido sometidas a ningún tratamiento ni cirugía, y mujeres sanas que serán parte del grupo de control. Debemos indicar también, que los autores del dataset han eliminado del estudio original de la Universidad de Coimbra, y por lo tanto del conjunto de datos del que se dispone, a pacientes con un índice de masa corporal superior a 40kg/m2 y aquellos con variables faltantes.

En lo referente a los datos obtenidos a parte de las analíticas de sangre, se ha de señalar que estas fueron realizadas en ayunas. Esta sangre luego fue centrifugada 2.5 kg a 4°C y almacenada a -80 °C. La leptina, adiponectina, resistina y MCP-1 fueron

obtenidas mediante kits ELISA específico para cada una. Mediante HOMA se ha calculado usando los valores de insulina y glucosa (2)

$$\text{logarithm}\ ((\text{If}) \times (\text{Gf})) / 22.5$$

Ecuación 7: Cálculo de HOMA

3.2.3 Descripción atributos

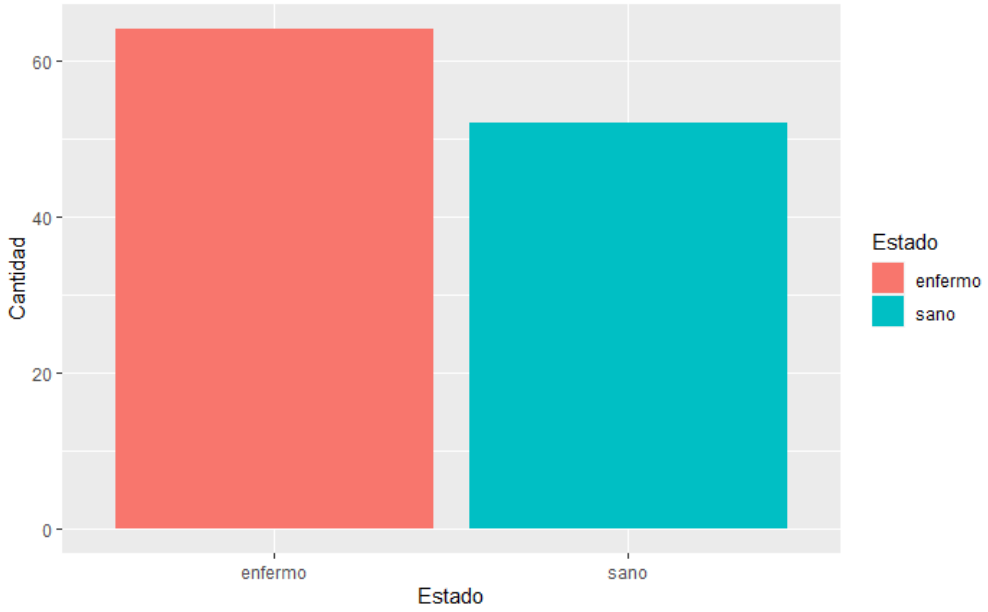

Ilustración 5: Distribución variable Objetivo Classification

Como se ha comentado contamos con una distribución de registros que no sufre un gran desbalanceo entre clases en referencia a la variable objetivo (Classification) ya que el 55% de los registros corresponden a pacientes enfermos, mientras el 45% a pacientes sanos

A continuación, se procede al estudio de los 9 atributos presentes en este conjunto de datos, divididos en sanos y enfermos. Para ello se hará uso de gráficas boxplot con el fin observar la distribución de los datos, así como detectar posibles valores anómalos, y de gráficos de densidad con los que poder observar grupos poblacionales más afectados.

3.2.3.1 Age (Edad)

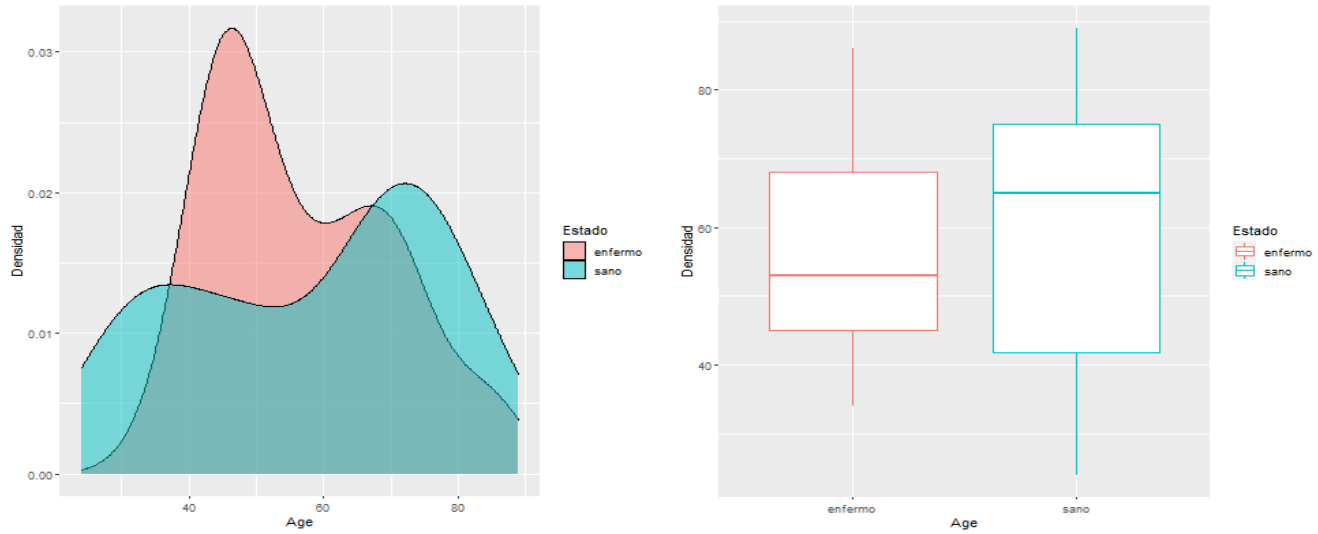

Ilustración 6: Análisis descriptivo variable Age

Esta variable viene a caracterizar la edad del paciente al momento de la toma de la muestra. Tras el estudio de la variable Age, podemos observar como hay un rango de edad entorno a los 40-55 años, donde se puede observar un agrupamiento de los casos de afectación por cáncer de mama. A través del análisis de diagramas de caja confirmamos la distribución de los datos, y se constata la no existencia de valores que podrían ser considerados anómalos.

3.2.3.2 BMI (Índice de masa corporal)

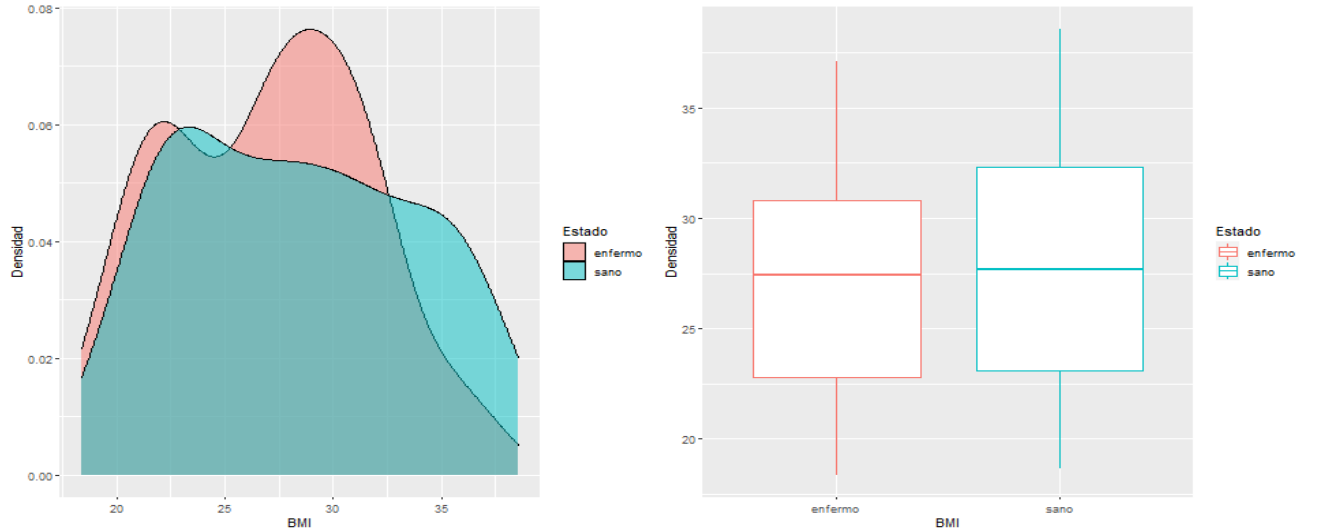

Ilustración 7: Análisis descriptivo variable BMI

La variable BMI se encarga de recoger la relación entre el peso y la altura del paciente. En nuestro conjunto de datos no se ven diferencias destacables entre las dos poblaciones ni se observa la existencia de valores anómalos.

3.2.3.3 Glucose (Glucosa)

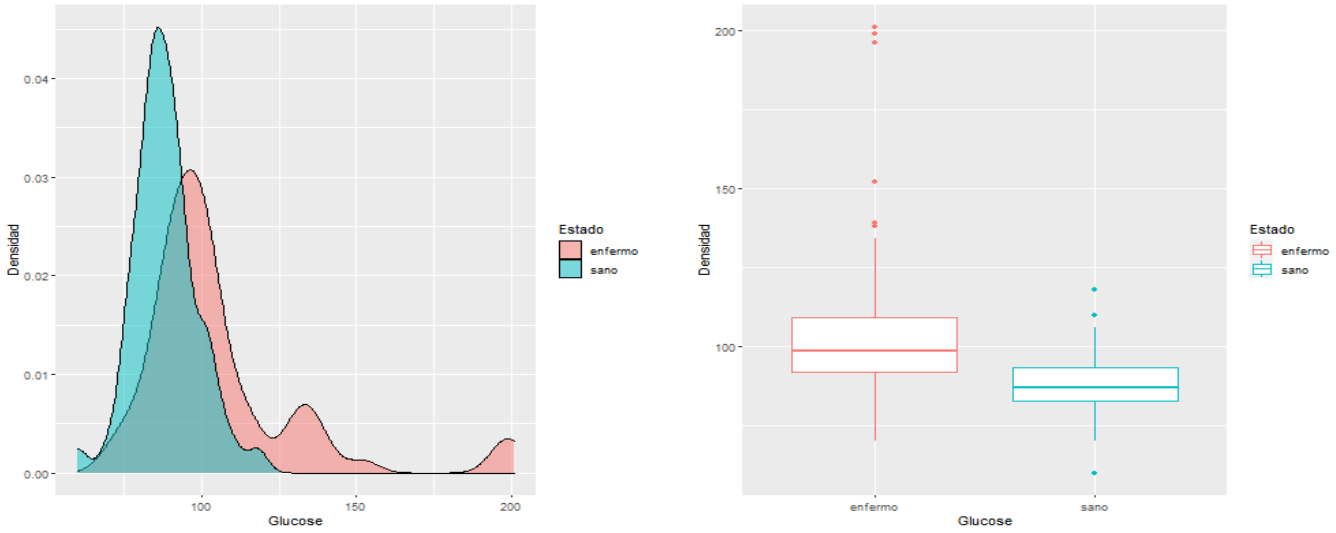

Ilustración 8: Análisis descriptivo variable Glucose

La glucosa es una variable cuyos valores normales oscilan entre 70 y 100 mg/dL. En las muestras contenidas en los datos, se observa como los pacientes enfermos tienden a tener la glucosa algo más elevada, mientras que los pacientes sanos se mantienen en ese margen en la mayoría de los casos. En los diagramas de cajas se ven valores anómalos para ambas poblaciones, aunque esta desviación es mucho más llamativa en el caso de los pacientes enfermos

3.2.3.4 Insulin (Insulina)

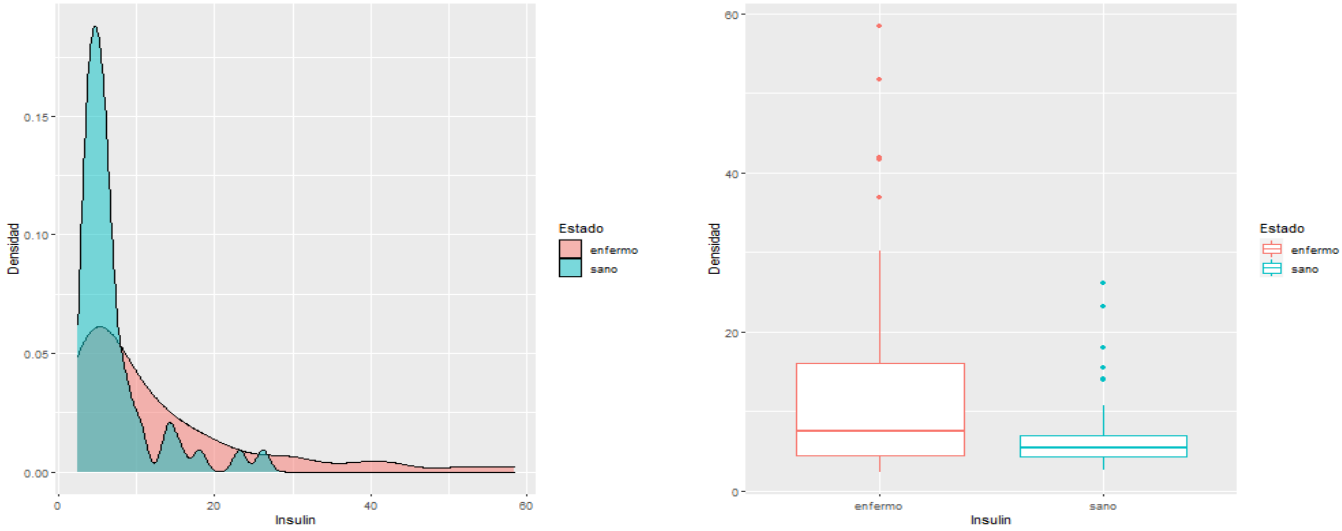

Ilustración 9: Análisis descriptivo variable Insulin

La insulina se trata de una hormona que regula los niveles de azúcar en sangre y que típicamente distribuye sus valores normales entre comidas entre 60 y 100 mg/dL
Del análisis descriptivo no se pueden sacar conclusiones que hagan pensar que nos encontramos ante valores peligrosos, aunque para el caso de pacientes vemos como su distribución es más suave que en los pacientes sanos, teniendo un valor medio superior

3.2.3.5 HOMA

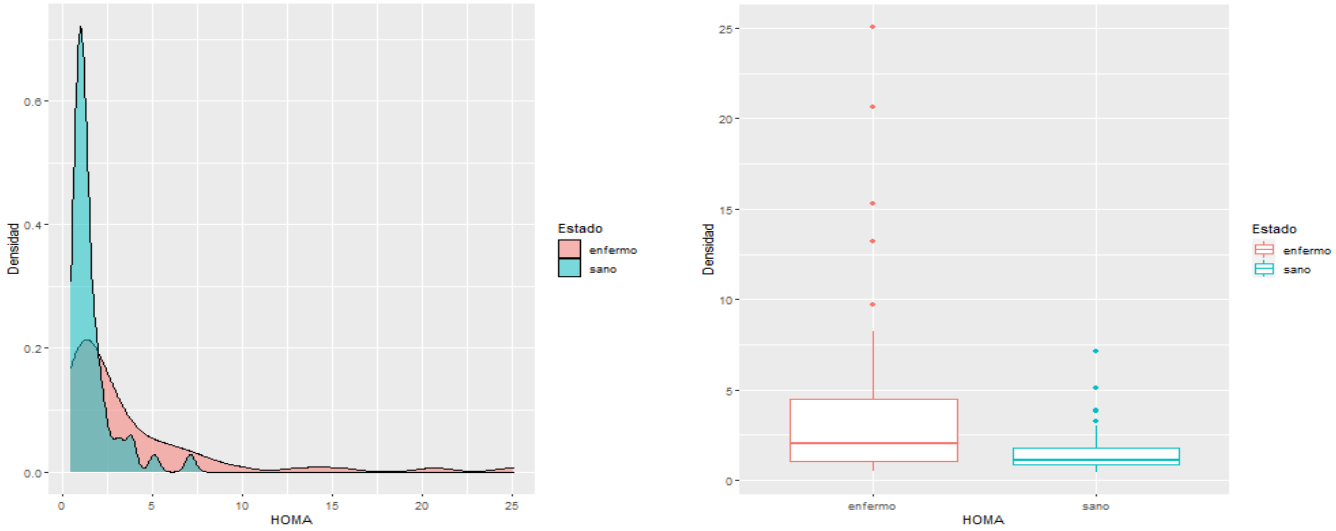

Ilustración 10: Análisis descriptivo variable HOMA

Como se ha indicado con anterioridad está variable proviene de la combinación de Insulina y Glucosa. Tiene una distribución similar a insulina y se observan algunos valores anómalos, al igual que para el caso de Insulina.

3.2.3.6 Leptin (Leptina)

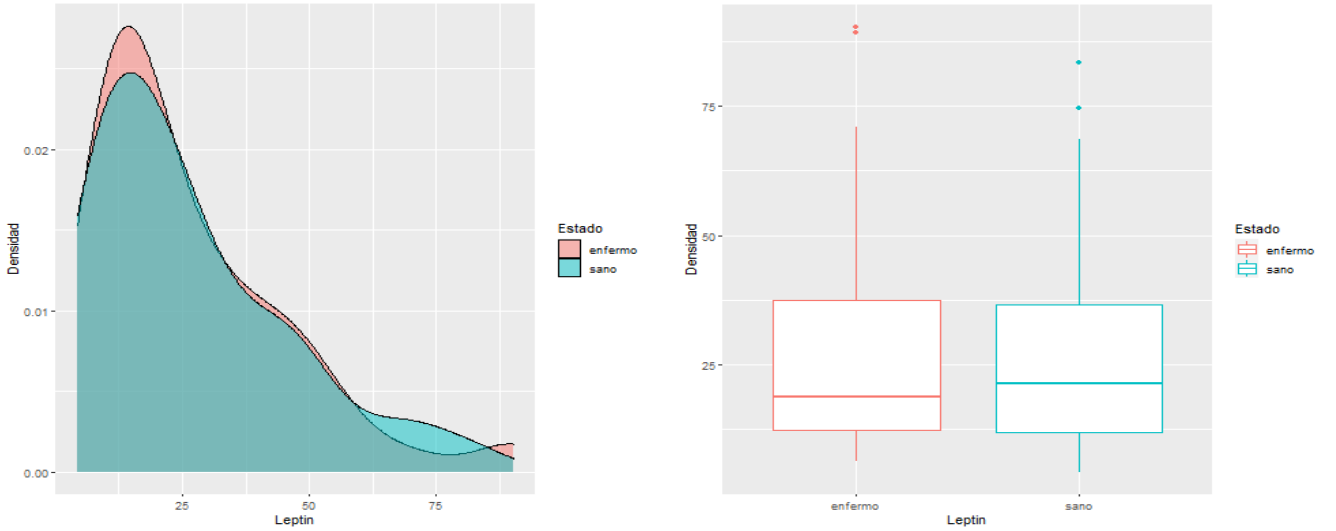

Ilustración 11: Análisis descriptivo variable Leptin

La leptina es un indicador relacionado con la grasa acumulada por un paciente y que le hace regular el apetito, y se creía asociada al cáncer de mama por su relación con el nivel de estrógenos (52). En nuestro caso, se advierte como no hay diferencia entre ambas poblaciones al observar esta variable, y que se presentan algunos valores anómalos

3.2.3.7 Adiponectin (Adiponectina)

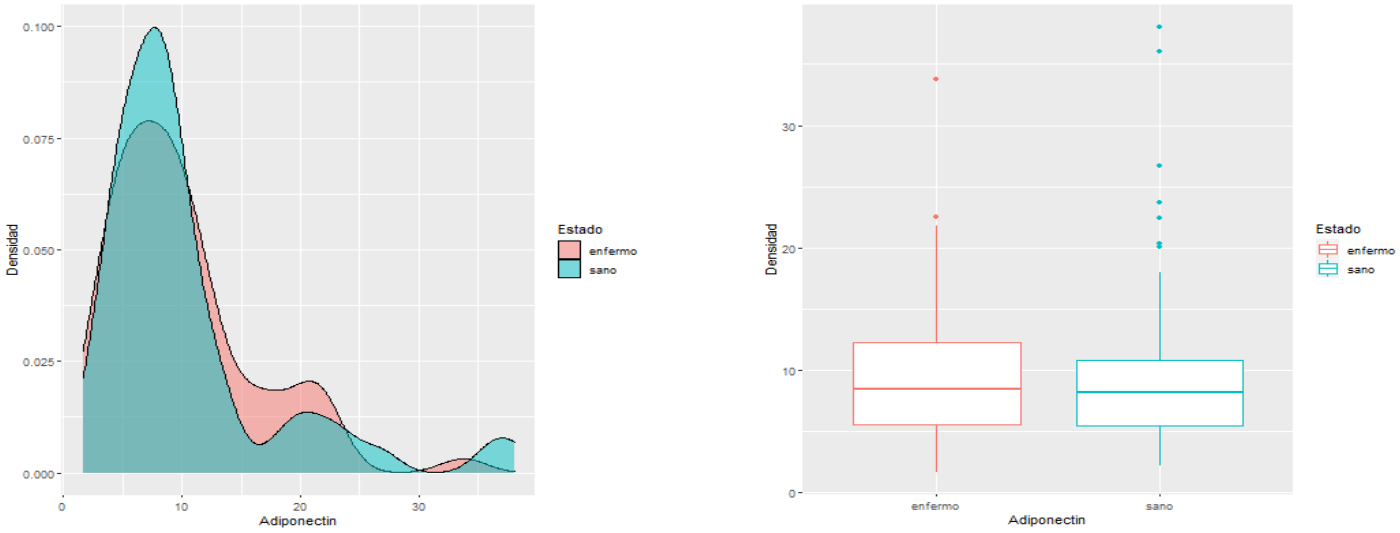

Ilustración 12: Análisis descriptivo variable Adiponectin

La adiponectina tiene influencia en la regulación de la sensibilidad a la insulina y homeostasis de la glucosa y está asociada a la obesidad. Tras el análisis exploratorio no se observan diferencias en la distribución de valores de ambas poblaciones, aunque se debe señalar la existencia de mayor número de anómalos en los casos sanos, debido al agrupamiento de la distribución

3.2.3.8 Resistin (Resistina)

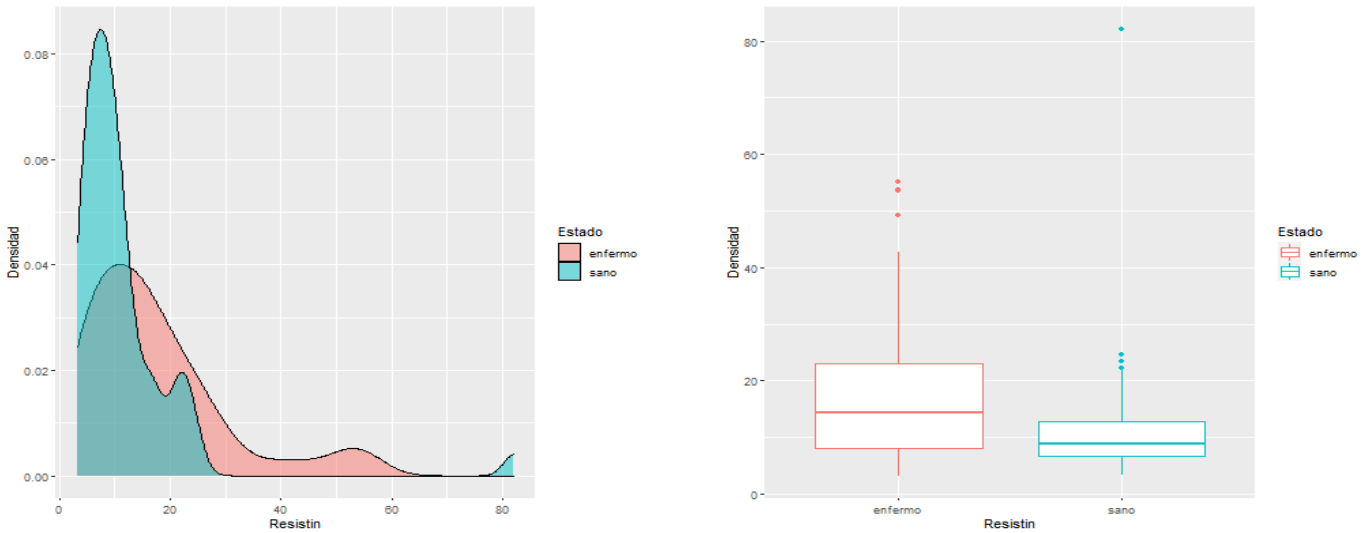

Ilustración 13: Análisis descriptivo variable Resistin

La resistina es una hormona que se produce en el tejido adiposo y se asocia a la obesidad y respuesta inflamatoria del organismo. En el análisis exploratorio se ve una distribución más suave de los valores para la población enferma, mientras que está se agrupa mucho hacia valores pequeños en la población sana, lo que nos produce la existencia de valores anómalos, aunque relativamente bajos, para el caso de los sanos. También se observan valores anómalos en el caso de los pacientes enfermos.

3.2.3.9 MCP.1 (MCP.1)

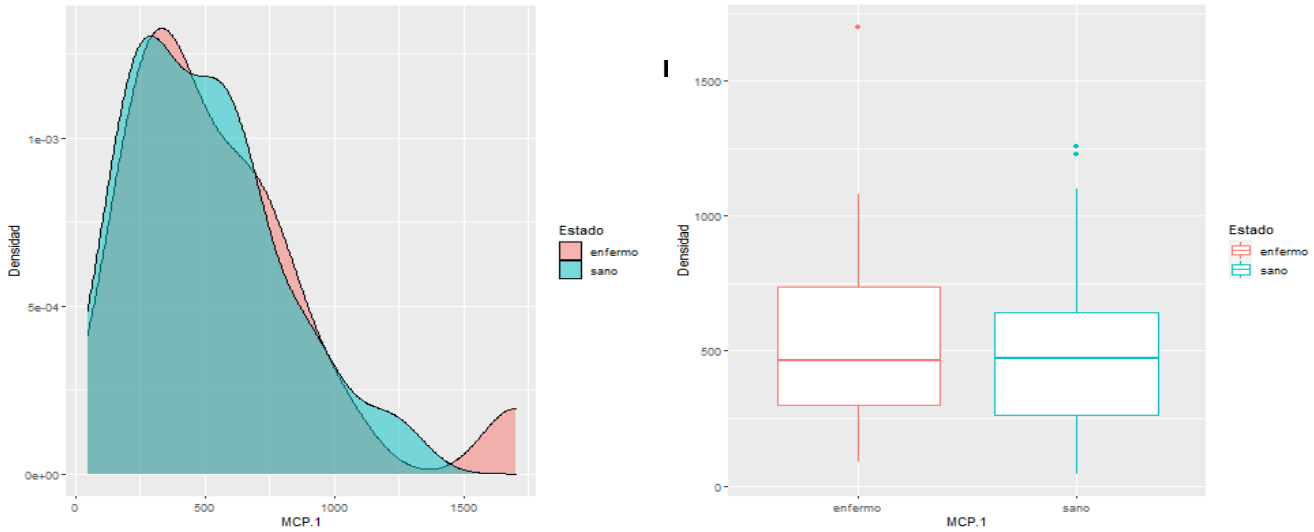

Ilustración 14: Análisis descriptivo variable MCP.1

La MCP.1 es una hormona relacionada con enfermedades inflamatorias del organismo. Se extrae cierta homogeneidad de distribución entre ambas poblaciones y, presentado ambos valores anómalos, en caso de enfermedad son más extremos.

3.2.4 Relaciones entre variables

Con el ánimo de estudiar las relaciones entre variables para una posterior eliminación de aquellas que no aporten información adicional, se ha realizado un estudio de correlación como el que se muestra en la ilustración 14

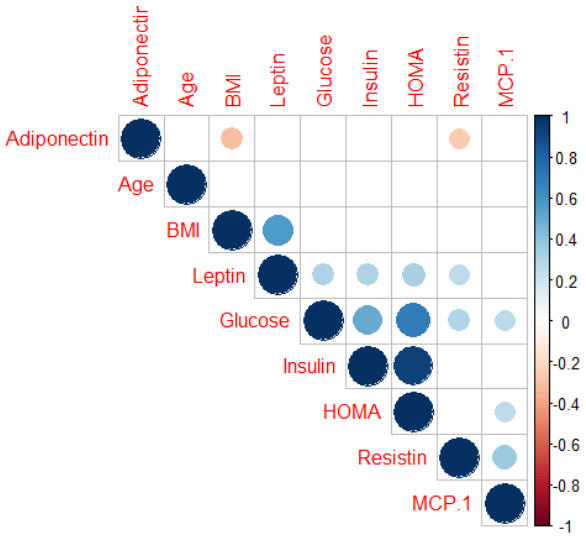

Ilustración 15: Relación entre variables

A partir de este estudio confirmamos nuestras sospechas de gran relación entre las variables HOMA, Insulina y Glucosa. También se observan relaciones de menor intensidad entre las variables relacionadas con la grasa y obesidad, como son el BMI, adiponectina, Resistina, Leptina.

3.2.5 Conclusiones sobre la exploración inicial de los datos

Una vez han sido estudiadas las variables a nuestra disposición se pueden obtener las siguientes conclusiones:

- **Poca longitud del conjunto de datos**: cómo se puede advertir, solo tenemos a nuestra disposición 116 registros. Este hecho dificultará la tarea de entrenamiento, ya que es necesario dividir el conjunto en dos subconjuntos, para el que en este caso se seguirá la proporción 70-30 (train-test), con la intención de tener un conjunto de prueba lo suficientemente grande con el que poder comprobar el desempeño de los modelos entrenados. Esto provocará que entrenemos solo con 80 casos aproximadamente lo cual, a priori puede ser insuficiente con el fin de obtener mejores resultados.
- **Homogeneidad de los datos**: Esto se debe a que todos los datos provienen de la misma fuente y siguiendo el mismo procedimiento.

- **Diferencias entre poblaciones**: Se debe indicar que a partir del análisis exploratorio no se han conseguido encontrar diferencias significativas por cada variable con relación a desarrollar o no la enfermedad, por lo que se tendrá que proceder a un análisis multifactorial.
- **Correlación entre variables**: A partir de este estudio se concluye que no veríamos un efecto negativo de la eliminación de la variable HOMA, o de las variables Insulina y Glucosa, ya que están fuertemente relacionadas. Las demás variables, pese a existir pequeñas relaciones entre ellas, estás no son lo suficientemente grandes como para poder plantearnos la eliminación de alguna de las mismas.

Los dos primeros hechos nos inclinan a pensar que los modelos producto del entrenamiento con este conjunto de datos tendrán una fuerte tendencia al sobre entrenamiento y, por lo tanto, toda conclusión deberá ser extrapolada a otros problemas con mucha cautela, en el caso de que los datos provengan de fuentes distintas a la disponible para la realización de este trabajo y que sigan procedimiento similares pero no iguales, con la intención de minimizar el impacto de los sesgos que introduzcan los procedimientos y pruebas actuales.
Por ello, el caso ideal habría sido la consecución de más conjuntos de datos de otras organizaciones u hospitales.

En el apartado del Dashboard (3.5) se introducirá una iniciativa que pueda permitir el crecimiento de estos conjuntos de datos tanto en tamaño como en heterogeneidad mediante una herramienta que ayude a la explotación de los modelos y a su vez la introducción de nuevos datos que podrán ser tenidos en cuenta en futuras versiones de los modelos.

3.3 Preparación de los conjuntos de datos

Una vez realizado el estudio de las variables disponibles para la realización de este trabajo se procede al tratamiento del conjunto inicial con el fin de limpiar los datos y generar los conjuntos que posteriormente serán utilizados en la parte de modelado.
Previo al tratamiento de las anomalías encontradas, hay que señalar que este conjunto de datos no tiene valores en blanco o perdidos, por lo que no será necesario tomar ninguna decisión ni aplicar ninguna técnica en este aspecto.

3.3.1 Tratamiento de valores anómalos

Han sido observados valores anómalos en las variables MCP.1, Resistin, Adiponectin, Leptina, Insulina, Glucosa y HOMA. Estas variables son aquellas obtenidas a partir de los análisis de sangre y, aunque pudieran deberse a algún error de medida ya que se encuentran muy alejadas de los valores considerados normales, especialmente para la población enferma, se opta por mantener los registros referentes a estas variables inalterados, ya que cualquier imputación de variables podría tener una influencia no deseada en la distribución de nuestros datos.
Unido a lo anterior, debemos señalar el factor más importante a la hora de tomar esta decisión en lo referente al rendimiento de los modelos, el cual se trata sobre la robusted que proporcionan los métodos basados en árboles de decisión ante el tratamiento de valores extremos, ya que las reglas de corte dividen los datos en torno a un valor, sin importar la distribución de los valores que queden a izquierda y derecha de este corte.

3.3.2 Escalado de variables

Debido a los diferentes rangos entre los que se distribuyen las variables, se ha planteado la posibilidad de realizar un escalado de las variables para evitar tener problemas con la varianza. Esta medida sería muy interesante de aplicar en otros tipos de modelos de clasificación como SVM o NN, ya que ayudaría en la convergencia hacia una solución óptima de estos algoritmos. Para nuestro caso, en el cual se hace uso de algoritmos basados en árboles de decisión, este escalado no es necesario, ya que estos métodos no son sensibles a la varianza contenida en los datos y entre diferentes variables

3.3.3 Categorización de variables

Se ha decidido no realizar ninguna categorización debido a que hacerlo conllevaría introducir nuestro propio sesgo a la hora de agrupar en rangos los valores que toman las variables y que los métodos basados en árboles de decisión no presentan problemas a la hora de fijar umbrales en los que dividir los datos numéricos, por lo que los mismos métodos se encargan de agrupar por los valores de las variables.

3.3.4 Creación de conjuntos

Una vez limpios los datos se procede a la creación de los conjuntos de datos.
Previo a la presentación de los conjuntos de datos, se debe señalar que se ha realizado la transformación de la variable objetivo original, Classification, de la cual han sido transformados sus valores 1 y 2 a sano y enfermo, posteriormente convirtiéndola en un factor que permita el uso de clasificadores.
A continuación, se presentan los diferentes conjuntos.

3.3.4.1 Conjuntos 9 predictores

Se realiza la división del conjunto inicial con las modificaciones ya planteadas, siguiendo una proporción 70-30% para los conjuntos de entrenamiento y prueba y realizando una división estratificada, es decir, intentando mantener la proporción enfermo-sano del conjunto original.
Producto de esta operación obtenemos:

- **Coimbra_train_9**: conjunto formado por 82 registros sobre las variables Age, BMI, Leptin, Resistin, Insulin, Glucose, HOMA, MCP.1 y Adiponectin.
  Se presentan 37 casos sanos por 45 afectados por cáncer
- **Coimbra_test_9**: conjunto formado por 34 registros sobre las variables Age, BMI, Leptin, Resistin, Insulin, Glucose, HOMA, MCP.1 y Adiponectin.
  Se presentan 15 casos sanos por 19 afectados por cáncer

3.3.4.2 Conjuntos 4 predictores

En el paper original (2), se presenta una experimentación haciendo uso de únicamente 4 predictores, por lo que generaremos esos conjuntos de datos con fines comparativos de resultados entre nuestros modelos y con los modelos propuestos en el artículo original de la Universidad de Coimbra. Se realiza la división del conjunto inicial con las modificaciones ya planteadas, teniendo en cuenta solo las variables BMI, Age, Resistin y Glucose, siguiendo una proporción 70-30% para los conjuntos de entrenamiento y

prueba y realizando una división estratificada, es decir, intentando mantener la proporción enfermo-sano del conjunto original.

Producto de esta operación obtenemos:

- **Coimbra_train_4**: conjunto formado por 82 registros sobre las variables Age, BMI, Resistin, Glucose.

  Se presentan 37 casos sanos por 45 afectados por cáncer
- **Coimbra_test_4**: conjunto formado por 34 registros sobre las variables Age, BMI, Resistin, Glucose.

  Se presentan 15 casos sanos por 19 afectados por cáncer

3.3.4.3 Conjuntos PCA

Debido a las conclusiones extraídas en el estudio de correlación y el estudio descriptivo, sabemos de la existencia de variables en nuestro conjunto de datos, las cuales nos pueden estar aportando una información similar de cara entrenar el modelo, por lo que resulta interesante agrupar la mayor cantidad de variabilidad explicable en un conjunto de predictores menores con el fin de comprobar si el rendimiento aumenta o disminuye. Antes de realizar este paso, tenemos que señalar que los modelos basados en árboles de decisión ya proporcionan de forma nativa la importancia de los predictores que se han utilizado para realizar la partición de los datos. Pese a este hecho, exploraremos un nuevo conjunto que pueda ser extrapolable en un futuro a otras aproximaciones estadísticas del problema.

Para realizar esta agrupación de variabilidad se ha empleado la técnica de análisis de componentes principales con el ánimo de obtener estas nuevas variables eliminando la correlación entre ellas.

Fijamos como objetivos conseguir al menos un 90% de la varianza explicada, prescindiendo de aquellos predictores que expliquen menos de un 5% de varianza

A continuación, se presentan una tabla con los valores obtenidos:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Proporción varianza	0.3398	0.1691	0.1297	0.1229	0.08028	0.07303	0.04906	0.03251	0.00356
Proporción acumulada	0.3398	0.5090	0.6387	0.7615	0.84184	0.91487	0.96393	0.99644	1.00000

Tabla 3: Resultados aplicación PCA

Una vez generadas las componentes principales, se seleccionan aquellas variables que han cumplido con los criterios que habíamos fijado al inicio de la aplicación e la técnica. Se realiza la división del conjunto inicial con las modificaciones ya planteadas, teniendo en cuenta solo las componentes principales PC1, PC2, PC3, PC4, PC5, PC6; siguiendo una proporción 70-30% para los conjuntos de entrenamiento y prueba y realizando una división estratificada, es decir, intentando mantener la proporción enfermo-sano del conjunto original.

Producto de esta operación obtenemos:

Coimbra_train_PCA: conjunto formado por 82 registros sobre las componentes principales PC1, PC2, PC3, PC4, PC5, PC6.

Se presentan 37 casos sanos por 45 afectados por cáncer

Coimbra_test_PCA: conjunto formado por 34 registros sobre las componentes principales PC1, PC2, PC3, PC4, PC5, PC6.

Se presentan 15 casos sanos por 19 afectados por cáncer.

3.4 Modelado y evaluación

Común a todos los modelos que han sido entrenados, presentamos la metodología de trabajo de cara a la creación y optimización de los modelos. Los tres algoritmos en los que se basan los modelos, y de los cuales referenciamos los apartados donde se presentan los resultados de la experimentación, han sido árboles de decisión (apartado 3.4.1), random forest (apartado 3.4.2) y gradient boosting machines (apartado 3.4.3), siendo esta última una implementación de GBM para la clasificación.

Se ha comenzado creando un modelo línea base con el que comparar la tarea de optimización, y para el cual se han seleccionado los valores por defecto existentes en los paquetes utilizados para su implementación y que serán mencionados en los apartados 3.4.1, 3.4.2 y 3.4.3. Esta creación de modelo base nos permite realizar una primera aproximación al algoritmo y su comportamiento.

Tras la creación del modelo base, se ha hecho uso del paquete Caret (53), que está pensado para la implementación de modelos y optimización de los mismos mediante una prueba masiva de diferentes valores para los parámetros modificables de cada algoritmo. Este paquete nos permite seleccionar la métrica a tener en cuenta para la selección del mejor modelo, así como realizar el entrenamiento mediante validación cruzada.

Por lo anteriormente comentado, a la hora de optimizar los modelos, haremos uso de Validación cruzada (CV). En nuestro caso, y debido al tamaño del conjunto aplicaremos validación cruzada con 5 iteraciones, y la haremos en 10 ciclos, lo que resulta en una técnica de Validación cruzada repetida de 5 pliegues o K-fold repeated Cross-Validation.

Se han tenido en consideración otras técnicas de Validación cruzada como LOOCV. Pese a parecer que pueda ser perfecta para nuestro problema en el que se tienen pocas muestras, se corre mucho riesgo de producir sobreajuste al conjunto de entrenamiento, y como estamos usando unos modelos que son proclives a sobreentrenar debido su naturaleza, seguiremos las recomendaciones pautadas en (49), por lo que finalmente se ha seleccionado como técnica de CV la señalada anteriormente.

De cara a realizar la comparación entre modelos del proceso de optimización, la elección de modelo subóptimo se realizará mediante el valor ROC.

Posteriormente a este proceso, con el fin de realizar una comparación, probaremos los modelos sobre el conjunto reservado para el test observando las métricas Accuracy (ACC), Especificidad (Espec) y Sensibilidad (Sens) mediante la aplicación de la función ConfusionMatrix de R, y el índice Youden, que relaciona Sensibilidad y

Especificidad y se seleccionarán las implementaciones más prometedoras de cada algoritmo con el fin de tener nuestro algoritmo final óptimo.

3.4.1 Árboles de decisión (DT)

3.4.1.1 Modelos Base

La implementación de los árboles de decisión se ha realizado mediante el paquete C50 de R (54).

Las variables que han sido modificadas a lo largo de nuestra ejecución son:

- Trials: variable que se encarga de la operación de boosting. Para nuestro caso está fijada en 1, ya que reservamos las operaciones de boosting para GBM/GBT
- Rules/Model: Variable encargada de indicar si se desea que el modelo en árbol sea transformado en un modelo de reglas True/'rules' o que se mantenga como False/'tree'
- Winnow: Variable que indica si se desea realizar selección de variables previa por poder predictivo

De cara a la realización del modelo base, se han utilizado los siguientes valores por defecto:

- Trials:1
- Rules: False
- Winnow:Tree

En la tabla 4 se muestra el desempeño de los modelos base tras haber sido entrenados haciendo uso del conjunto de entrenamiento, sin validación cruzada y con los valores por defecto y posteriormente aplicados al conjunto de datos de test.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	DT_base_9	0.7647	(0.5883, 0.8925)	0.8421	0.6667	0.5088
Coimbra_test_9						
Coimbra_train_4	DT_ base _4	0.7941	(0.621, 0.913)	0.7895	0.8000	0.5895
Coimbra_test_4						
Coimbra_train_PCA	DT_base_PCA	0.6176	(0.4356, 0.7783)	0.5789	0.6667	0.2456
Coimbra_test_PCA						

Tabla 4: Modelos Base DT aplicados a conjuntos de test

Se puede observar que el modelo que mejor se ha comportado es el que utiliza las 4 variables propuestas para estudio en (2).

3.4.1.2 Modelos Óptimos

De cara a realizar la optimización se ha hecho uso del paquete Caret de R y los siguientes rangos de valores:

- Trials: 1
- Rules/Model: tree, rules
- Winnow: False, True
- Métrica de evaluación optimización: ROC
- Validación cruzada: RepeatedCV
- Repeticiones CV: 10
- K-Folds: 5

En la tabla 5 se presentan los resultados del proceso de optimización haciendo uso del paquete Caret de R. Se proporcionan los valores de simulación que han dado lugar a los mejores modelos y las métricas de desempeño del proceso de selección del modelo óptimo con el conjunto de entrenamiento y haciendo uso de validación cruzada.

Modelo	Trials Óptimo	Model Óptimo	Winnow Óptimo	ROC	SD	Sens	Espec	Youden
DT_Optim_9	1	tree	False	0.6937	0.1129	0.7800	0.5586	0.3386
DT_ Optim _4	1	tree	False	0.7054	0.1216	0.7511	0.6154	0.3665
DT_Optim_PCA	1	tree	False	0.6552	0.1128	0.7222	0.5486	0.2708

Tabla 5: Variables y métricas de los modelos óptimos DT a partir del conjunto de entrenamiento

La tabla 6 muestra el resultado de la aplicación de los modelos óptimos al conjunto de test, con las métricas de desempeño, y mediante la cual se puede realizar la comparación con los modelos base. Se debe indicar que, para el caso de los árboles de decisión, los valores por defecto del paquete C5.0 se han mostrado como los mejores valores de cara a preparar los modelos aplicados al conjunto de datos a estudio, obteniendo como mejor experimento, el derivado de la aplicación de los modelos al conjunto de 4 predictores.
Debido a esto, el modelo seleccionado para la comparación posterior es DT_Optim_4.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	DT_Optim_9	0.7647	(0.5883, 0.8925)	0.8421	0.6667	0.5088
Coimbra_test_9						
Coimbra_train_4	DT_ Optim _4	0.7941	(0.621, 0.913)	0.7895	0.8000	0.5895
Coimbra_test_4						
Coimbra_train_PCA	DT_Optim_PCA	0.6176	(0.4356, 0.7783)	0.5789	0.6667	0.2456
Coimbra_test_PCA						

Tabla 6: Modelos óptimos DT aplicados a conjuntos de test

3.4.2 Random Forest (RF)

3.4.2.1 Modelos Base

La implementación de los árboles de decisión se ha realizado mediante el paquete RandomForest de R (55).

Las variables que han sido modificadas a lo largo de nuestra ejecución son:

- ntree: Indica el número de árboles débiles que se generarán para su posterior promediado
- mtry: Número de predictores a tener en cuenta para realizar cada división
- seed: valor general del entorno para asegurar la repetibilidad de los resultados

De cara a la realización del modelo base, se han utilizado los siguientes valores por defecto, que son los que entrega el paquete:

- ntree: 500
- mtry: 3
- seed:400

En la tabla 7 se muestra el desempeño de los modelos base tras haber sido entrenados haciendo uso del conjunto de entrenamiento, sin validación cruzada y con los valores por defecto y posteriormente aplicados al conjunto de datos de test.

Se observa, que para este caso el modelo base que mejor se ha comportado es aquel que contiene los 9 predictores, tanto en Accuracy como en el Índice de Youden.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	RF_base_9	0.8235	(0.6547, 0.9324)	0.8421	0.8000	0.6421
Coimbra_test_9						
Coimbra_train_4	RF_ base _4	0.7059	(0.5252, 0.849)	0.6842	0.7333	0.4175
Coimbra_test_4						
Coimbra_train_PCA	RF_base_PCA	0.6765	(0.4947, 0.8261)	0.6316	0.7333	0.3649
Coimbra_test_PCA						

Tabla 7: Modelos Base RF aplicados a conjuntos de test

3.4.2.2 Modelos Óptimos

De cara a realizar la optimización se ha hecho uso del paquete Caret de R y los siguientes rangos de valores:

- ntree: 500,1000,1500,2000,2500
- mtry:
  - Para RF_Optim_9: 1,2,3,4,5,6,7,8,9
  - Para RF_Optim_4: 1,2,3,4
  - Para RF_Optim_PCA -> 1,2,3,4,5,6
- Métrica de evaluación optimización: ROC
- Validación cruzada: RepeatedCV
- Repeticiones CV: 10
- K-Folds: 5
- Seed:400

En la tabla 8 se presentan los resultados del proceso de optimización haciendo uso del paquete Caret de R. Se proporcionan los valores de simulación que han dado lugar a los mejores modelos y las métricas de desempeño del proceso de selección del modelo óptimo con el conjunto de entrenamiento y haciendo uso de validación cruzada.

Del proceso de optimización, se puede extraer que el experimento que mejor desempeño proporciona es aquel aplicado al conjunto de 4 predictores.

Modelo	ntree Óptimo	mtry Óptimo	ROC	SD	Sens	Espec	Youden
RF_Optim_9	500	4	0.7944	0.1042	0.7489	0.6400	0.3889
RF_ Optim _4	1500	1	0.8620	0.0794	0.8000	0.7436	0.5436
RF_Optim_PCA	2000	6	0.8068	0.0965	0.7556	0.6671	0.4227

Tabla 8: Variables y métricas de los modelos óptimos RF

La tabla 9 nos presenta los resultados de la aplicación de los modelos óptimos generados al conjunto de test. En este caso se puede observar una mejora notable en comparación con los resultados de los modelos base (tabla 7).

Del mismo modo, se observa como el modelo derivado del experimento con 9 predictores muestra unos mejores resultados pese a ser el que peor tras el proceso de optimización. Esto se debe a que los modelos de los experimentos para 4 y PCA predictores, tienden a sobrentrenarse durante la optimización debido al pequeño número de predictores disponible en cada conjunto, 4 y 6, unido al corto número de registros disponibles, resultando en una mejor generalización por parte del modelo RF_Optim_9, aplicado a los 9 predictores disponibles.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	RF_Optim_9	0.8529	(0.6894, 0.9505)	0.8421	0.8667	0.7088
Coimbra_test_9						
Coimbra_train_4	RF_ Optim _4	0.7353	(0.5564, 0.8712)	0.7895	0.6667	0.4502
Coimbra_test_4						
Coimbra_train_PCA	RF_Optim_PCA	0.6765	(0.4947, 0.8261)	0.6316	0.7333	0.3649
Coimbra_test_PCA						

Tabla 9: Modelos óptimos RF aplicados a conjuntos de test

3.4.3 Gradient Boosting Machines (GBM)

3.4.3.1 Modelos Base

La implementación de los árboles de decisión se ha realizado mediante el paquete GBM (56) y Caret para adaptarlo al problema de clasificación (53).
Las variables que han sido modificadas a lo largo de nuestra ejecución son:

- interaction.depth: Valor que indica la profundidad máxima del árbol
- n.trees: Indica el número de árboles débiles que se generarán de forma secuencial para ir mejorando el modelo
- shrinkage: factor de aprendizaje para cada árbol
- n.minobsinnode: Número mínimo de registros en el nodo terminal/final
- seed: valor general del entorno para asegurar la repetitibilidad de los resultados

De cara a la realización del modelo base, se han utilizado los siguientes valores por defecto:

- interaction.depth: 1
- n.trees: 100
- shrinkage: 0.1
- n.minobsinnode: 10
- seed: 400

En la tabla 10 se muestra el desempeño de los modelos base tras haber sido entrenados haciendo uso del conjunto de entrenamiento, sin validación cruzada y con los valores por defecto y posteriormente aplicados al conjunto de datos de test.
Se observa, que para este caso el modelo base que mejor se ha comportado es aquel que contiene los predictores PCA, tanto en Accuracy como en el Índice de Youden.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	GBM_base_9	0.7353	(0.5564, 0.8712)	0.7895	0.6667	0.4562
Coimbra_test_9						
Coimbra_train_4	GBM_ base _4	0.6765	(0.4947, 0.8261)	0.5263	0.8667	0.3930
Coimbra_test_4						
Coimbra_train_PCA	GBM_base_PCA	0.7647	(0.5564, 0.8712)	0.6842	0.8667	0.5509
Coimbra_test_PCA						

Tabla 10: Modelos Base GBM aplicados a conjuntos de test

3.4.3.2 Modelos Óptimos

De cara a realizar la optimización se ha hecho uso del paquete Caret de R y los siguientes rangos de valores:

- interaction.depth(i_depth): 1,5,9
- n.trees (nt): 100,200,300,400,500
- shrinkage(s): 0.1, 0.3, 0.5, 0.7, 0.9
- n.minobsinnode(minnode): 3,6,10
- Métrica de evaluación optimización: ROC
- Validación cruzada: RepeatedCV
- Repeticiones CV: 10
- K-Folds: 5
- Seed:400

En la tabla 11 se presentan los resultados del proceso de optimización haciendo uso del paquete Caret de R. Se proporcionan los valores de simulación que han dado lugar a los mejores modelos y las métricas de desempeño del proceso de selección del modelo óptimo con el conjunto de entrenamiento y haciendo uso de validación cruzada.
Del proceso de optimización, se puede extraer que el experimento que mejor desempeño proporciona es aquel aplicado al conjunto de 4 predictores.

Modelo	i_depth Óptimo	n.trees Óptimo	shrinkage Óptimo	minnode Óptimo	ROC	SD	Sens	Espec	Youden
GBM_Optim_9	5	100	0.1	3	0.7957	0.0941	0.7733	0.6546	0.4279
GBM_ Optim _4	5	100	0.1	6	0.8412	0.0910	0.7800	0.7654	0.5454
GBM_Optim_PCA	5	400	0.1	3	0.8359	0.0950	0.7867	0.7261	0.5128

Tabla 11: Variables y métricas de los modelos óptimos GBM

La tabla 12 nos presenta los resultados de la aplicación de los modelos óptimos generados al conjunto de test. En este caso se puede observar una mejora notable en comparación con los resultados de los modelos base (tabla 10) excepto para el caso del modelo PCA. Esto se debe a que las métricas de optimización empleadas para la selección de modelos resultan en un modelo que se ajusta bien a los datos de entrenamiento, pero tiene un desempeño inferior una vez aplicado al conjunto de test.
Como se ha comentado anteriormente, esto se debe a que los modelos GBM y RF, tienden a sobre ajustarse al conjunto de entrenamiento a la hora de realizar las

simulaciones de optimización. El modelo con mejor desempeño tras ser aplicado al conjunto de test es, en este caso, GBM_Optim_9.

Datasets	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
Coimbra_train_9	GBM_Optim_9	0.8235	(0.6547, 0.9324)	0.8421	0.8000	0.6421
Coimbra_test_9						
Coimbra_train_4	GBM_ Optim _4	0.7647	(0.5883, 0.8925)	0.7368	0.8000	0.5368
Coimbra_test_4						
Coimbra_train_PCA	GBM_Optim_PCA	0.7059	(0.4947, 0.8261)	0.6842	0.7333	0.4175
Coimbra_test_PCA						

Tabla 12: Modelos óptimos GBM aplicados a conjuntos de test

3.4.4 Comparación modelos optimizados

A lo largo de este apartado revisaremos el rendimiento de los modelos óptimos obtenidos tras la utilización del paquete Caret, siendo la primera tarea revisar el rendimiento de entrenamiento de los modelos, para posteriormente comparar su desempeño sobre el conjunto de prueba.

Tipo	Modelo	ROC	SD	Sens	Espec	Youden
DT	DT_Optim_9	0.6937	0.1129	0.7800	0.5586	0.3386
DT	DT_ Optim _4	0.7054	0.1216	0.7511	0.6154	0.3665
DT	DT_Optim_PCA	0.6552	0.1128	0.7222	0.5486	0.2708
RF	RF_Optim_9	0.7944	0.1042	0.7489	0.6400	0.3889
RF	RF_ Optim _4	0.8620	0.0794	0.8000	0.7436	0.5436
RF	RF_Optim_PCA	0.8068	0.0965	0.7556	0.6671	0.4227
GBM	GBM_Optim_9	0.7957	0.0941	0.7733	0.6546	0.4279
GBM	GBM_ Optim _4	0.8412	0.0910	0.7800	0.7654	0.5454
GBM	GBM_Optim_PCA	0.8359	0.0950	0.7867	0.7261	0.5128

Tabla 13: Comparación modelos entrenados

En la Tabla 13, se presenta el desempeño de todos los modelos optimizados con sus métricas, remarcando los mejores para cada conjunto de entrenamiento.

Podemos observar cómo encontramos una gran diferencia en términos de valores ROC y Youden entre los modelos vitaminados (Bagging y Boosting) y el modelo básico de árbol de decisión.

Atendiendo al conjunto con 9 predictores, el modelo DT_Optim_9 se muestra muy por detrás de los otros dos. Por otra parte, los modelos RF_Optim_9 y GBM_Optim_9, muestran un desempeño similar, aunque si se tiene en cuenta el índice de Youden, el modelo **GBM_Optim_9** muestra un mejor desempeño a la hora de ajustarse a los datos de entrenamiento.

En el caso del conjunto reducido con 4 predictores, se produce la misma situación que para el conjunto de 9 predictores. Los modelos RF_Optim_4 y GBM_Optim_4 son los que muestran un mejor rendimiento, alcanzando valores de índice de Youden muy similares, pero observando una pequeña diferencia para el valor ROC, siendo por lo tanto el modelo que mejor se ha ajustado a los datos, el **RF_Optim_4.**

Seguidamente se compara el rendimiento en el conjunto PCA, donde se observa un claro ganador, el modelo **GBM_OPTIM_PCA,** el cual muestra el mejor rendimiento con creces tanto para el índice de Youden como para el valor ROC, superando a RF_OPTIM_PCA por 7 puntos porcentuales y 3 puntos porcentuales respectivamente.

Por lo tanto, podemos concluir que los modelos Random Forest y Gradient Boosting Trees tiene un desempeño muy superior al de los árboles de decisión tradicionales.

Añadir finalmente, que los valores ROC para todos los modelos Random Forest y Gradient Boosting Machines están entorno al 80-85%, pero observamos como punto débil, que los modelos no son capaces de obtener unos valores elevados de especificidad, no consiguiendo superar en la mayoría de los casos el 75%. Al estar trabajando con una muestra tan pequeña, se acepta como normal que estos valores fluctúen y sean próximos, debido a que un error penaliza mucho al porcentaje total de la métrica.

Una vez estudiados los modelos tras el entrenamiento, se compara su desempeño al ser aplicados al conjunto de prueba, con la finalidad de comprobar el nivel de sobreajuste que han sufrido los modelos y si son capaces de generalizar. Esta afirmación debe ser tomada con cautela, ya que el hecho de solo poseer 34 registros para realizar la prueba produce que un error de clasificación penalice mucho al porcentaje con el que se presentan las métricas de evaluación.

Tipo	Modelo	ACC	ACC 95%CI	Sens	Espec	Youden
DT	DT_Optim_9	0.7647	(0.5883, 0.8925)	0.8421	0.6667	0.5088
DT	DT_ Optim _4	0.7941	(0.621, 0.913)	0.7895	0.8000	0.5895
DT	DT_Optim_PCA	0.6176	(0.4356, 0.7783)	0.5789	0.6667	0.2456
RF	RF_Optim_9	0.8529	(0.6894, 0.9505)	0.8421	0.8667	0.7088
RF	RF_ Optim _4	0.7353	(0.5564, 0.8712)	0.7895	0.6667	0.4502
RF	RF_Optim_PCA	0.6765	(0.4947, 0.8261)	0.6316	0.7333	0.3649
GBM	GBM_Optim_9	0.8235	(0.6547, 0.9324)	0.8421	0.8000	0.6421
GBM	GBM_ Optim _4	0.7647	(0.5883, 0.8925)	0.7368	0.8000	0.5368
GBM	GBM_Optim_PCA	0.7059	(0.4947, 0.8261)	0.6842	0.7333	0.4175

Tabla 14: Comparación modelos entrenados aplicados a conjunto de test

Atendiendo al conjunto con 9 predictores, observamos como el modelo **RF_Optim_9** es el que mejor resultados obtiene a la hora de clasificar el conjunto de prueba, aunque obtenemos unos resultados bastante satisfactorios con el modelo GBM_Optim_9.

En el caso del conjunto con 4 predictores, sorpresivamente observamos que, en contraposición con los resultados de entrenamiento, el modelo que mejor consigue generalizar y clasificar los datos es el **DT_Optim_4**, mientras que los modelos de Bagging y Boosting presentan un rendimiento menor al esperado, por lo que sospechamos que están sobre ajustándose al conjunto de prueba al entrenar.

Finalmente, se ha de señalar que los resultados de prueba para el conjunto PCA han sido los peores de los 3 conjuntos por una amplia diferencia, lo que nos inclina a no recomendar su uso a la hora de clasificar conjuntos nuevos debido a la pobre generalización que muestran capaces de hacer.

3.5 Despliegue

3.5.1 Motivación Dashboard Shiny

La intención para implementar un Dashboard reside en dos pilares.
El primero consiste en ofrecer al personal sanitario investigador una herramienta en la que pueda revisar los conjuntos de datos, su estadística básica y aplicar modelos preentrenados a sus conjuntos con la finalidad de apoyar sus decisiones, transformando la medicina tradicional en una medicina basada en evidencias.
El segundo pilar se basa en la idea de incrementar los conjuntos disponibles para realizar modelos de ciencia de datos que permitan un mejor apoyo predictivo a la decisión, mediante nuevos biomarcadores o combinaciones que faciliten el trabajo de cribado.
Para la implementación que se presenta en este trabajo, se hace uso del conjunto de Coimbra Original y el conjunto de prueba para 9 predictores, así como los 3 modelos optimizados que han sido entrenados sobre el conjunto de 9 predictores.

3.5.2 Diseño Dashboard

A continuación, se presentan los esquemas de diseño realizados para resolver este apartado, el cual ha sido validado gracias a la colaboración de familiares y amigos del ámbito sanitario, que han resultado de mucha ayuda de cara a ir añadiendo características.

3.5.3.1 Pantalla inicial

Esta pantalla debería servir como entrada a la aplicación, y tendrá que contener la información básica sobre qué problema estamos enfrentando, en este caso el Dataset Coimbra, qué posibilidades de uso tiene la herramienta, realizar una presentación de las variables con las que se va a trabajar, así como una breve explicación de las métricas utilizadas a la hora de evaluar un modelo y que permitan al personal sanitario tener el grado necesario de confianza en nuestra solución.
En la ilustración 16, se presenta el esquema final tras las iteraciones realizadas.

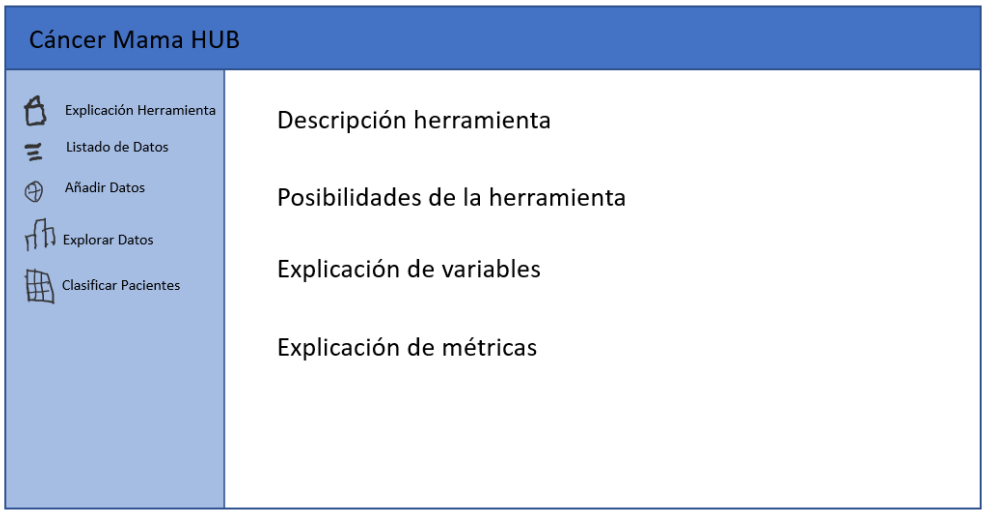

Ilustración 16: Pantalla Inicial

3.5.3.2 Pantalla listado de Datos

Esta pantalla es la encargada de presentar los datos a los usuarios. Debido a ello tendrá que presentar un listado de aquellos registros presentes en el conjunto de datos. Se ha decidido añadir un buscador que permita localizar pacientes por características, así como un filtro que permita ver un número reducido de pacientes. Tras las iteraciones con el personal sanitario, se ha añadido un botón que permita la descarga de los datos almacenados.

En la ilustración 17 se presenta el esquema final.

Ilustración 17: Pantalla listado de datos

3.5.3.3 Pantalla introducción de nuevos datos

La idea inicial tras esta pantalla consistía en permitir al personal sanitario añadir datos nuevos manualmente. Se diferencia entre datos validados que pueden pasar a formar parte del dataset de preparación de modelos y datos para clasificar, que son aquellos que el personal sanitario quiere someter a los modelos.

Tras varias iteraciones, se ha recopilado una gran necesidad del personal sanitario. Esta es la posibilidad de cargar ficheros directamente en el sistema para realizar las pruebas o para añadir nuevos datos.

Por ello, en la ilustración 18 se presenta el esquema que da respuesta a estas necesidades. También se nos ha comentado que sería interesante añadir a futuro una característica que permita borrar valores o listas, en caso de que se haya cometido un error.

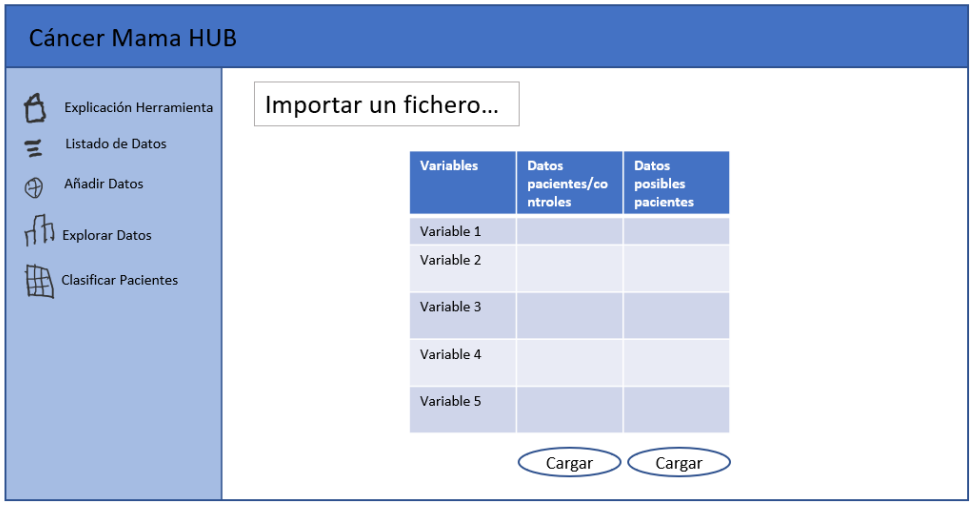

Ilustración 18: Pantalla introducción de datos

3.5.3.4 Pantalla exploración de datos

El cometido de esta pantalla es presentar la distribución de los datos con el fin de conocer en profundidad las variables que conforman el conjunto de datos.

Se implementan dos tipos de gráficas, boxplot y densidad, con el objetivo de facilitar la tarea de comprensión.

Se añade en la parte superior un selector de variable, que permita una mejor gestión del espacio.

En la ilustración 19 se presenta la propuesta de pantalla.

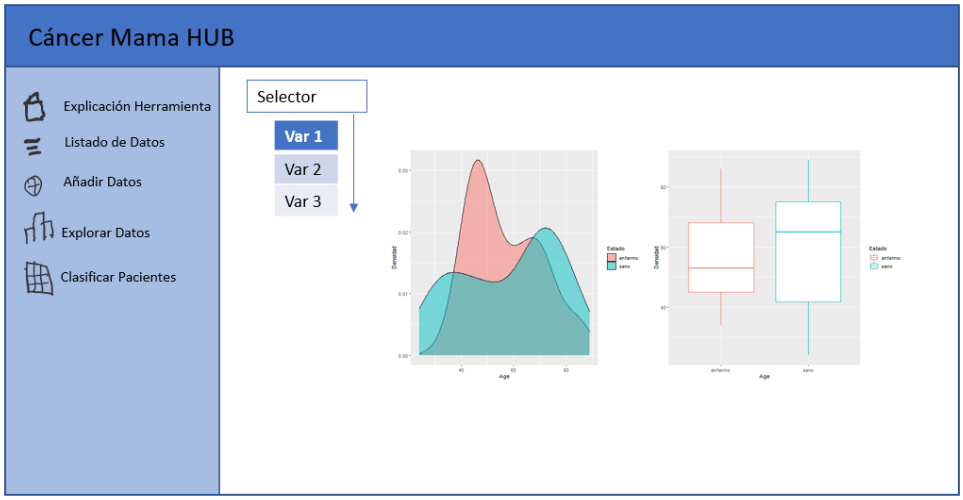

Ilustración 19: Pantalla exploración de datos

3.5.3.5 Pantalla clasificación de datos

La última pantalla de nuestro sistema debe servir como punto de prueba de los diferentes modelos entrenados. Para este caso se han seleccionado los mejores modelos de cada algoritmo para 9 predictores.

Se ha de permitir la selección del conjunto de datos a clasificar.

A partir de las iteraciones con los potenciales usuarios, se ha acordado la introducción de las métricas que indiquen la confianza en los modelos, así como un botón para permitir la descarga del conjunto clasificado. En la ilustración 20 se presenta la propuesta final de esquema para esta pantalla.

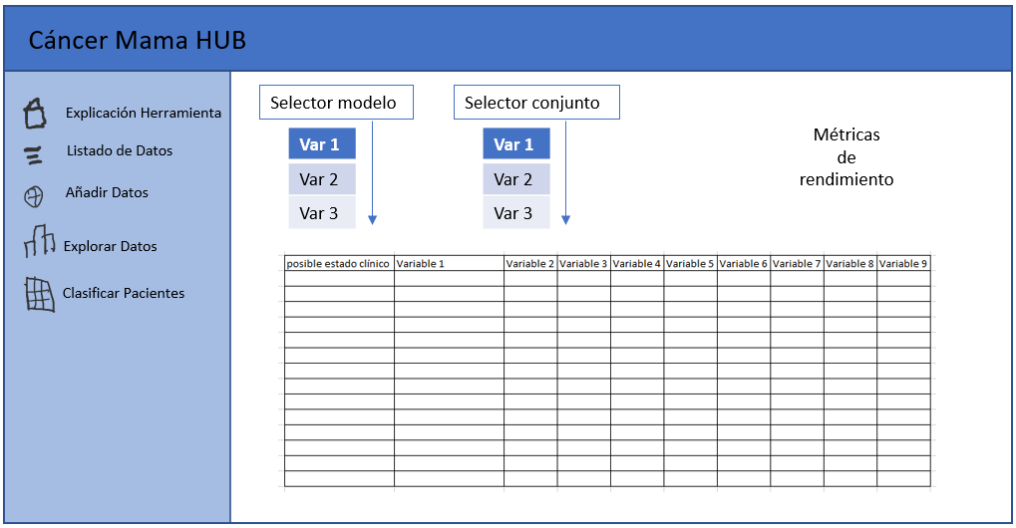

Ilustración 20: Pantalla clasificación de datos

3.5.3 Implementación

A partir de los esquemas generados en la fase de diseño, se ha realizado la implementación del dashboard mediante el uso del paquete Shiny de R (9)
Este paquete permite unir el ambiente de experimentación con el consumo y explotación de los datos y conclusiones extraídos de este.
En la fase de implementación se han llevado a cabo rediseños estéticos para adaptar el sistema a dispositivos móviles o con pantallas de dimensiones reducidas.
En el repositorio del TFM está disponible el script de R (app_shiny.R) mediante el cual se ha implementado el dashboard que procedemos a presentar.

3.5.3.1 Pantalla inicial

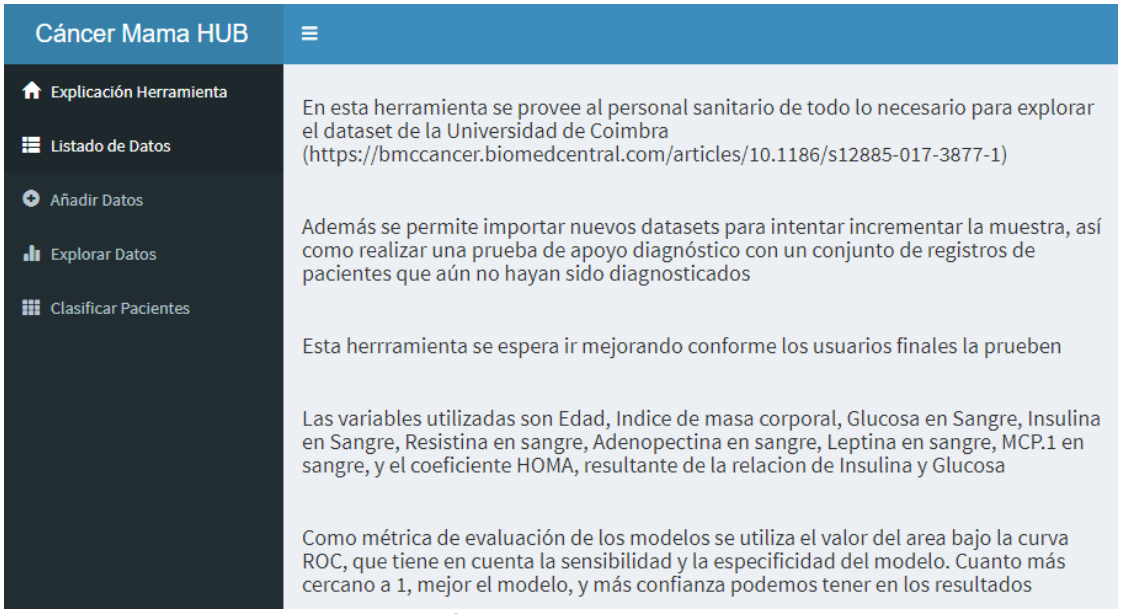

Ilustración 21: Pantalla inicial final

Se presenta en la ilustración siguiente la implementación que consideramos como final de esta pantalla, aunque esté sujeta a pequeñas modificaciones conforme se vaya obteniendo retorno de las opiniones de los usuarios

3.5.3.2 Pantalla listado de Datos

Presentamos la implementación de esta pantalla en la ilustración 22. A partir del diseño esquemático se ha modificado la localización del buscador y del filtro de selección de registros. Se debe indicar que, a parte del buscador inicial, existe a pie de pantalla un buscador por variable nativo de Shiny, que funciona correctamente en caso de introducir el valor exacto, pero que se ha observado como el rendimiento es peor de cara a filtrar por valores.

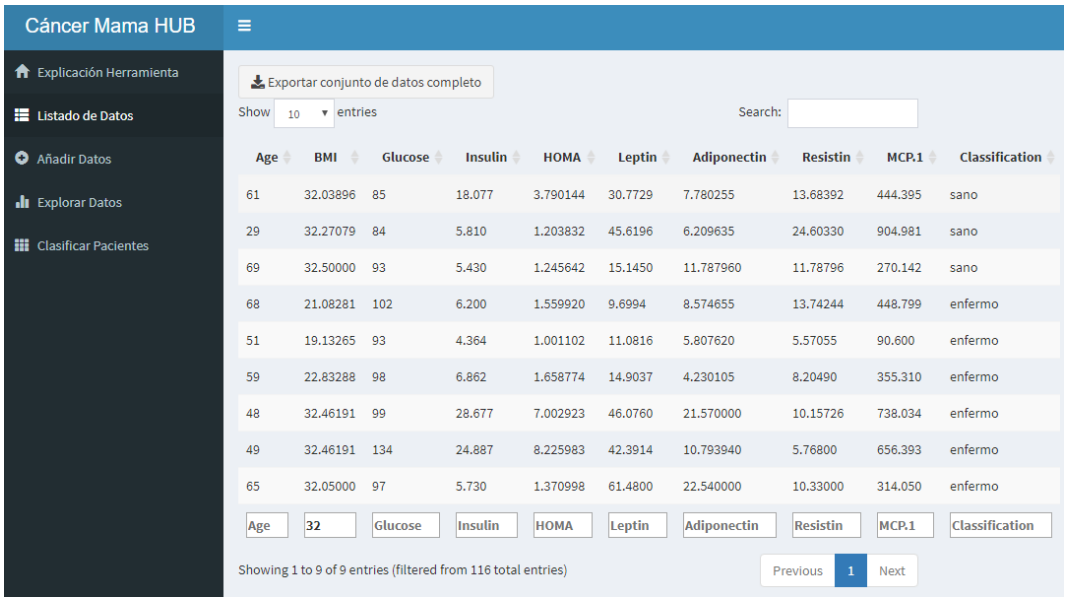

Ilustración 22: Pantalla listado de datos final

3.5.3.3 Pantalla introducción de nuevos datos

De cara a la implementación final de esta pantalla se ha diferenciado entre nuevos registros introducidos manualmente que contienen el estado del paciente, y aquellos que no, considerando los primeros como un añadido del conjunto de datos Coimbra, y el segundo como un conjunto dedicado a la clasificación, con el que apoyar la decisión del personal sanitario.

En caso de importar un conjunto de datos, debe ser en formato csv (comma separated value) y estos conjuntos podrán ser considerados para ambas tareas, por lo que se espera que los nuevos registros introducidos que contengan el estado del paciente, cumplan con los requisitos señalados en el artículo original (2).

En la ilustración 23 se presenta el aspecto final de la implementación de esta pantalla en el dashboard.

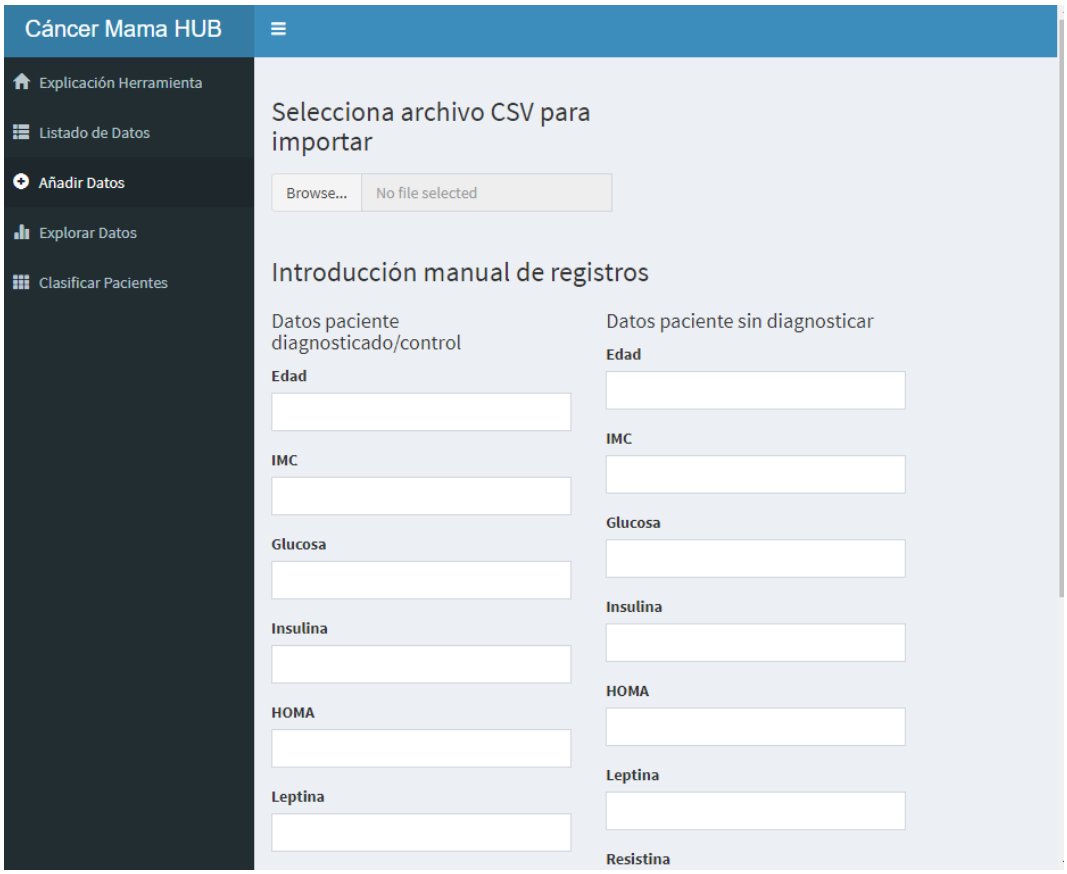

Ilustración 23: Pantalla introduccion de nuevos datos final

3.5.3.4 Pantalla exploración de datos

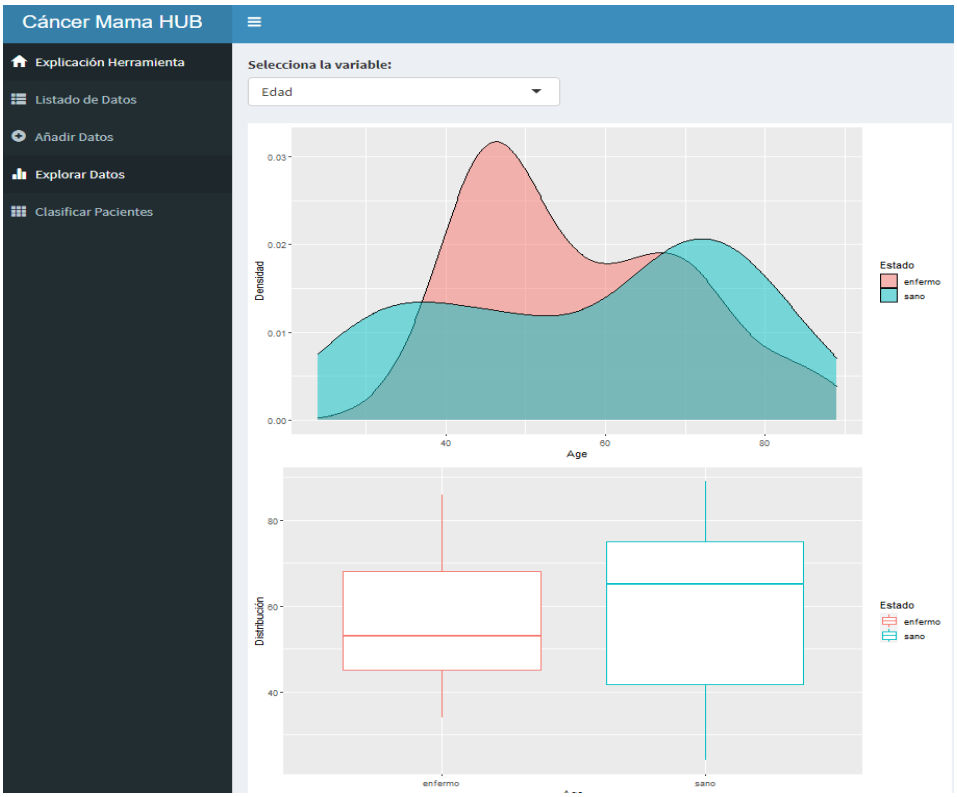

Ilustración 24: Pantalla exploración datos final

En la ilustración 24 se presenta la pantalla de explotación de datos que ha sufrido una remodelación de cara a la implementación final, y que consiste en la relocalización de las gráficas para ser presentadas en vertical, lo que facilita la consulta del dashboard desde dispositivos móviles, cuya pantalla se caracteriza por ser más alta que ancha.

3.5.3.5 Pantalla clasificación de datos

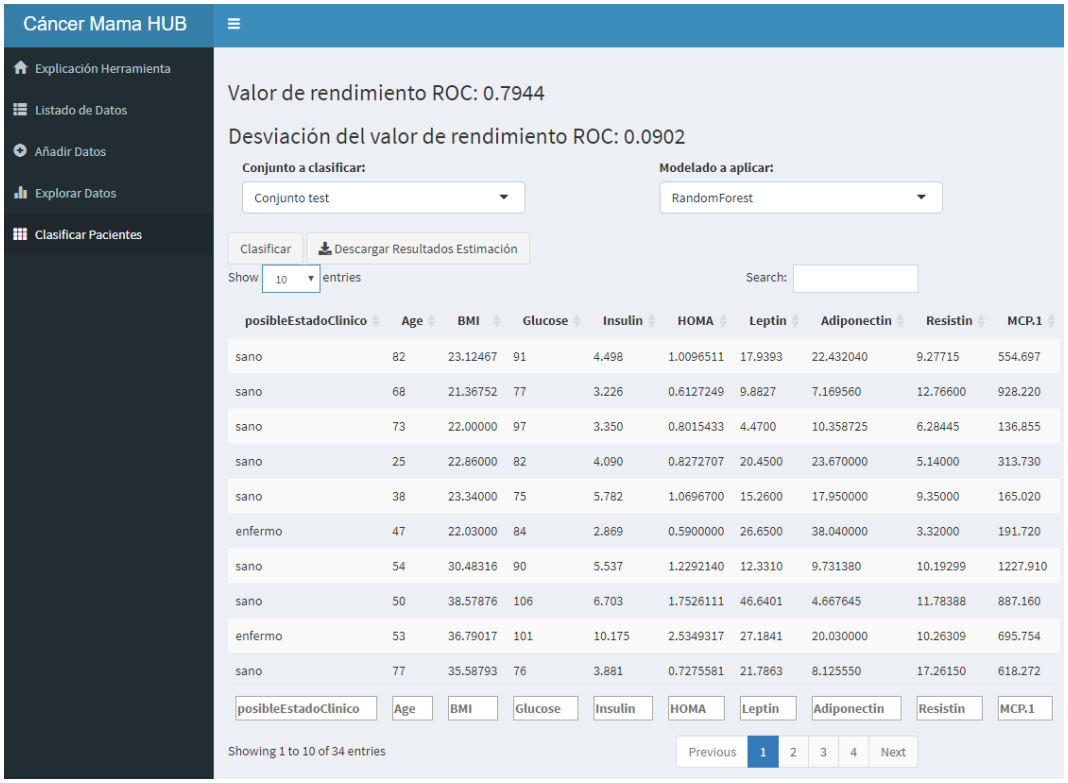

Ilustración 25: Pantalla clasificación de datos final

En la ilustración 25 se presenta la implementación de la pantalla de clasificación, donde se observan los dos selectores, de modelo y conjunto, así como el botón para realizar la clasificación y el botón para descargar los resultados de la clasificación en caso de que sean interesantes para el usuario.

Como se ha comentado anteriormente, se ha decidido incluir las métricas de rendimiento tras el entrenamiento, ya que el personal sanitario consultado nos ha comentado varias veces que fiabilidad aporta el método de clasificación seleccionado. Resulta interesante continuar con esta colaboración de cara a conseguir proveer al usuario de unas métricas que incentiven el uso de este sistema.

Los modelos implementados son GBM_Optim_9, RF_Optim_9 y DT_Optim_9.

En caso de querer añadir un nuevo modelo para 9 predictor bastaría con añadirlo a la carpeta modelos9 del repositorio proporcionado al inicio del capítulo 3

3.6 Comparación con modelos del estado del arte

Se presenta en el estado del arte un modelo SVM (2) de los autores del dataset, como el mejor modelo que han podido alcanzar.

La construcción de este modelo fue llevada a cabo con el conjunto de 4 predictores obteniendo unos valores de sensibilidad y especificidad en los siguientes intervalos [82%, 88%] y [84%, 90%], una vez aplicados a un conjunto de entrenamiento. Este modelo tiene un rendimiento superior sobre el conjunto de entrenamiento que el propuesto en este trabajo, con valores 0.749 y 0.640 respectivamente derivados del proceso de optimización.

En (30), se nos presentan los resultados de aplicación de diferentes modelos al conjunto de datos original, con nueve predictores. Los valores de accuracy obtenidos por estos modelos (DT=0.686, SVM=0.514, RF=0.743, LR=0.657, NN=0.600), se muestran inferiores a los valores de accuracy obtenidos por nuestro mejor modelo (RF_Optim_9) el cual obtiene un valor de accuracy de 0.8529 con un intevarlos de confianza al 95% entre (0.6894, 0.9505) y con valores de especificidad y sensibilidad alrededor del 85%, lo cual nos muestra como nuestro modelo propuesto es superior.

A continuación se procede a comparar nuestros modelos con los presentes en (31), debido a que también hacen uso de árboles de decisión, como algoritmo de clasificación. En este artículo se preparan modelos DT, RF y GBM los cuales obtienen unos valores de accuracy de 69.28%, 70.31% y 74.14%, tras aplicarlos a un conjunto de prueba con 9 predictores. Por nuestra parte, se han conseguido unos modelos que proporcionan valores de accuracy de 76.47% (DT), 85.29% (RF) y 82.35% (GBM).

Atendiendo a modelos de Redes Neuronales en (28), se presenta un modelo con un 86.95 de accuracy, lo cual está levemente por encima de nuestro mejor modelo, RF_Optim_9, aunque la diferencia no se presenta como significativa.

En la tabla 15, se presenta una comparativa de los mejores modelos de cada estudio mencionado con resultados de aplicación a un conjunto de test.

Trabajo	Modelo	ACC	Sens	Espec
Propio	RF_Optim_9	0.853	0.842	0.867
(30)	RF	0.743	-	-
(31)	GBM	0.741	-	-
(28)	Red Neuronal	0.870	-	-

Tabla 15: Comparación modelos entrenados aplicados a conjunto de test

4. Conclusiones y Líneas futuras

Para la realización de este trabajo se han tenido contactos con el Complejo Hospitalario Universitario de Santiago, pero finalmente no han fructificado en la obtención de nuevos datos.

Se ha seguido exhaustivamente la planificación del proyecto y se ha adecuado el proceso de ciencia de datos a la metodología CRISP-DM, la cual se ha mostrado eficaz a la hora de conocer, comprender, planificar y responder a los objetivos de este trabajo

En las etapas finales del proyecto se ha podido implementar el dashboard, el cual se espera que esté disponible en la red para las fechas de presentación del proyecto, permitiendo así la realización de una demostración. Pese a esto se puede ejecutar localmente con la información disponible en el repositorio de github mencionado en el comienzo del apartado 3.

Se han extraído las siguientes conclusiones de la realización del trabajo:

1. El cáncer de mama es el tipo de tumor maligno más presente entre la población femenina actualmente.
2. Se estima que la población de riesgo se encuentra entre los 45 y 65 años, ampliando este rango para nuestra población de pacientes y controles entre 38 y 65 acorde a la gráfica de edad que hemos visto.
3. La detección precoz se muestra muy eficaz de cara a reducir el impacto negativo de la enfermedad.
4. Actualmente el método más eficaz de detección es la mamografía, aunque nos encontramos ante un procedimiento invasivo para el paciente y costoso.
5. La búsqueda de nuevos métodos basados en ciencia de datos que vengan a complementar las mamografías se muestra necesario, con la intención ayudar en el diagnóstico.
6. Los modelos generados basados en árboles de decisión se han mostrado a la altura de la mayor parte de los estudios presentados en el estado del arte, con un rendimiento superior tras ser aplicados a la población de test, excepto en el caso del estudio original de la universidad de Coimbra.
7. Para el correcto aprovechamiento de estas técnicas es necesario:
    a) Impulsar la cooperación entre personal sanitario y científicos de datos.
    b) Fruto de esta colaboración, la recogida y generación de conjuntos de datos más extensos y heterogéneos que permitan extrapolar los

modelos y conclusiones producidas a otras poblaciones distintas de la de estudio en diferentes localizaciones geográficas.

8. Sin embargo, el pequeño número de registros y la homogeneidad de estos no nos permite extrapolar los resultados obtenidos a poblaciones distintas a la del estudio inicial de la universidad de Coimbra.
9. Por ello, sería necesario seguir trabajando con una población mayor y más heterogénea con el fin de poder depurar una técnica de cribado alternativa, incluso teniendo en cuenta resultados de análisis genéticos
10. Se ha creado un dashboard mediante el cual se puede consultar la información utilizada, así como cargar nuevos datos para entrenar los modelos o para procesos de inferencia en los que apoyar al personal sanitario mediante el uso de los modelos generados en este trabajo.

Por lo tanto, podemos concluir que los objetivos de este trabajo planteados en el apartado 1.2 y 3.1 han sido cumplidos.

Las líneas futuras derivadas de este trabajo son las siguientes:

1. Aplicación de modelos no lineales como método de búsqueda de importancia de los predictores con el fin de reducir la dimensionalidad.
2. Ampliar el sistema de dashboard actual añadiendo nuevas funcionalidades derivadas de los resultados de interacción del personal sanitario.
3. Ampliación del sistema a tecnologías web consistentes de una base de datos de almacenamiento y servicios de conexión que permitan el acceso desde distintos dispositivos y aplicación, mientras se asegura la persistencia de los datos.
4. Ampliación de los modelos disponibles para la clasificación de los datos introducidos en el sistema con el ánimo del apoyo a la decisión del personal médico
5. Incrementar el tamaño del conjunto de datos mediante la implicación de un mayor número de organismos sanitarios con el fin de depurar este método de cribado y dotarlo de mayor robustez
6. Inclusión de nuevos predictores como los derivados de análisis genéticos.

5. Glosario

Agile: métodos de desarrollo de software basado en el desarrollo iterativo donde los requisitos y la solución evolucionan a lo largo del tiempo de proyecto.

Deep Learning (DL): conjunto de algoritmos de aprendizaje automático que intenta modelar abstracciones de alto nivel en datos usando arquitecturas computacionales que admiten transformaciones no lineales múltiples e iterativas de datos expresados en forma matricial o tensorial.

DICOM: estándar de transmisión de imágenes médicas y datos entre hardware de propósito médico.

Inteligencia Artificial (IA)-(AI): capacidad de las máquinas de procesar cantidades ingentes de información, y llevar a cabo acciones o elecciones que aumenten las probabilidades de éxito y ayuden a la mejora continua del comportamiento.

KNN: método de clasificación supervisada que se encarga de agrupar las muestras teniendo en cuenta la distancia de estas en un espacio vectorial.

Machine Learning (ML): rama de la inteligencia artificial cuyo objetivo es desarollar técnicas estadísticas que permitan a las máquinas aprender y mejorar con la experiencia.

Naïve Bayes (NB): clasificador probabilístico fundamentado en el teorema de Bayes.

Redes Neuronales (NN): modelo computacional vagamente inspirado en el funcionamiento de las neuronas.

Regresión Lineal (RL): método estadístico que busca la relación entre dos variables, de las cuales una es objetivo y la otra predictora.

Regresión Logística (LR): método de regresión centrado en la predicción del resultado que necesita de una variable categórica como variable objetivo.

SEOM: Sociedad Española de Oncología Médica, cuyo fin es avanzar frente al cáncer y contribuir a que los pacientes reciban la mejor atención sanitaria posible.

Support Vector Machines (SVM): conjunto de algoritmos de aprendizaje supervisado utilizados para resolver problemas de clasificación y regresión.

TNM: Clasificación de tumores malignos por la UICC.

UICC: Unión Internacional Contra el Cáncer, organización no gubernamental encarga en exclusividad a controlar mundialmente el cáncer, buscando su eliminicación.

Variable Categórica: variable que puede adoptar un número limitado de categorías.

Variable Objetivo: Variable de la cual se quiere predecir el resultado. Si es continua se estaría ante un problema de regresión, mientras que, si fuese categórica, se estaría ante un problema de clasificación.

6. Bibliografía

1. OMS | Cáncer de mama: prevención y control [Internet]. WHO. World Health Organization; 2020 [citado 27 de septiembre de 2020]. Disponible en: https://www.who.int/topics/cancer/breastcancer/es/

2. Patrício M, Pereira J, Crisóstomo J, Matafome P, Gomes M, Seiça R, et al. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. BMC Cancer. diciembre de 2018;18(1):29.

3. CRISP-DM: La metodología para poner orden en los proyectos [Internet]. Sngular. 2016 [citado 27 de septiembre de 2020]. Disponible en: https://www.sngular.com/es/data-science-CRISP-dm-metodologia/

4. Yip C-H, Smith RA, Anderson BO, Miller AB, Thomas DB, Ang E-S, et al. Guideline implementation for breast healthcare in low- and middle-income countries: early detection resource allocation. Cancer. 15 de octubre de 2008;113(8 Suppl):2244-56.

5. Recomendaciones de la Sociedad Americana Contra El Cáncer para la detección temprana del cáncer de seno [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.cancer.org/es/cancer/cancer-de-seno/pruebas-de-deteccion-y-deteccion-temprana-del-cancer-de-seno/guias-de-la-sociedad-americana-contra-el-cancer-para-la-deteccion-temprana-del-cancer-de-seno.html

6. Prevención del cáncer - SEOM: Sociedad Española de Oncología Médica © 2019 [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://seom.org/informacion-sobre-el-cancer/prevencion-cancer?start=2

7. Viñes JJ. La efectividad de la detección precoz de las enfermedades. An Sist Sanit Navar. abril de 2007;30(1):11-27.

8. Scaling Data Science: How We Use CRISP-DM and Agile [Internet]. AgileThought. 2018 [citado 27 de septiembre de 2020]. Disponible en: https://agilethought.com/blogs/scaling-data-science-use-CRISP-dm-agile/

9. Shiny [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://shiny.rstudio.com/

10. Las 10 principales causas de defunción [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

11. Estadísticas del cáncer - Instituto Nacional del Cáncer [Internet]. 2015 [citado 18 de octubre de 2020]. Disponible en: https://www.cancer.gov/espanol/cancer/naturaleza/estadisticas

12. Sociedad Española de Oncología Médica. Cifras del cancer 2020 [Internet]. https://seom.org/seomcms/images/stories/recursos/Cifras_del_cancer_2020

.pdf. [citado 18 de octubre de 2020]. Disponible en: https://seom.org/seomcms/images/stories/recursos/Cifras_del_cancer_2020.pdf

13. ¿Qué es el cáncer? [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.aecc.es/es/todo-sobre-cancer/que-es-cancer

14. Todos tenemos cáncer [Internet]. Acta Sanitaria. 2015 [citado 18 de octubre de 2020]. Disponible en: https://www.actasanitaria.com/todos-tenemos-cancer/

15. Tumor. En: Wikipedia, la enciclopedia libre [Internet]. 2020 [citado 18 de octubre de 2020]. Disponible en: https://es.wikipedia.org/w/index.php?title=Tumor&oldid=129992707

16. Estadificación del cáncer [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.cancer.org/es/tratamiento/como-comprender-su-diagnostico/estadificaciondelcancer.html

17. Estadificación del cáncer - Instituto Nacional del Cáncer [Internet]. 2015 [citado 18 de octubre de 2020]. Disponible en: https://www.cancer.gov/espanol/cancer/diagnostico-estadificacion/estadificacion

18. Diagnóstico del cáncer [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.aecc.es/es/todo-sobre-cancer/que-es-cancer/diagnostico-cancer

19. Pronóstico del Cáncer de Mama: Mortalidad y Esperanza de vida [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.aecc.es/es/todo-sobre-cancer/tipos-cancer/cancer-mama/mas-informacion/evolucion-cancer-mama

20. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. Breast Cancer Targets Ther. 10 de abril de 2019;11:151-64.

21. Salvadó Usach MT, Bosch Príncep R, Navas García N, Pons Ferré L, Lejeune M, López Pablo C, et al. Estudio comparativo de la supervivencia del cáncer de mama según diagnóstico asistencial versus programa de detección precoz. Rev Senol Patol Mamar - J Breast Sci. 1 de enero de 2016;29(1):13-8.

22. La IA es Clave para el Sector HealthCare – Nemix – Artificial Intelligence [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://www.nemix.es/ai/la-ia-es-clave-para-el-sector-healthcare/

23. Kelly A. Machine Learning Resampling Techniques for Class Imbalances [Internet]. Medium. 2020 [citado 18 de octubre de 2020]. Disponible en: https://towardsdatascience.com/machine-learning-resampling-techniques-for-class-imbalances-30cbe2415867

24. Crisóstomo J, Matafome P, Santos-Silva D, Gomes AL, Gomes M, Patrício M, et al. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. Endocrine. 1 de agosto de 2016;53(2):433-42.

25. Naveen, Sharma RK, Ramachandran Nair A. Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models. En: 2019 4th International Conference on Recent Trends on Electronics, Information, Communication Technology (RTEICT). 2019. p. 100-4.

26. Kayaalp F, Başarslan M. Performance Analysis Of Filter Based Feature Selection Methods On Diagnosis Of Breast Cancer And Orthopedics. 2019.

27. Salod Z, Singh Y. Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol. J Public Health Res [Internet]. 4 de diciembre de 2019 [citado 18 de octubre de 2020];8(3). Disponible en: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6902303/

28. Saritas MM, Yasar A. Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. Int J Intell Syst Appl Eng. 30 de junio de 2019;7(2):88-91.

29. Benítez-Mata B, Castro C, Castañeda R, Vargas E, Flores D-L. Prediction of Breast Cancer Diagnosis by Blood Biomarkers Using Artificial Neural Networks. En: González Díaz CA, Chapa González C, Laciar Leber E, Vélez HA, Puente NP, Flores D-L, et al., editores. VIII Latin American Conference on Biomedical Engineering and XLII National Conference on Biomedical Engineering. Cham: Springer International Publishing; 2020. p. 47-55. (IFMBE Proceedings).

30. Li Y, Chen Z. Performance Evaluation of Machine Learning Methods for Breast Cancer Prediction. Appl Comput Math. 18 de octubre de 2018;7(4):212.

31. Austria Y, Goh M, Jr L, Lalata J-A, Goh J, Vicente H. Comparison of Machine Learning Algorithms in Breast Cancer Prediction Using the Coimbra Dataset. Int J Simul Syst Sci Technol. 28 de julio de 2019;

32. Supplementary Materials. Science [Internet]. 19 de septiembre de 2020 [citado 27 de septiembre de 2020]; Disponible en: https://science.sciencemag.org/content/suppl/2018/01/17/science.aar3247.DC1

33. Zhang F, Chen J, Wang M, Drabier R. A neural network approach to multi-biomarker panel discovery by high-throughput plasma proteomics profiling of breast cancer. BMC Proc. 20 de diciembre de 2013;7(Suppl 7):S10.

34. Wong K-C, Chen J, Zhang J, Lin J, Yan S, Zhang S, et al. Early Cancer Detection from Multianalyte Blood Test Results. iScience. mayo de 2019;15:332-41.

35. Webb GI, Boughton JR, Wang Z. Not So Naive Bayes: Aggregating One-Dependence Estimators. Mach Learn. 1 de enero de 2005;58(1):5-24.

36. Webb GI, Boughton JR, Zheng F, Ting KM, Salem H. Learning by extrapolation from marginal to full-multivariate probability distributions: decreasingly naive Bayesian classification. Mach Learn. 1 de febrero de 2012;86(2):233-72.

37. UCI Machine Learning Repository: Breast Cancer Data Set [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer

38. Brown G. Diversity in Neural Network Ensembles [Internet]. 2004 [citado 18 de octubre de 2020]. Disponible en: http://www.cs.man.ac.uk/~gbrown/publications/gbrownThesis.pdf

39. Esmeir S, Markovitch S. Lookahead-based algorithms for anytime induction of decision trees. En: Proceedings of the twenty-first international conference on Machine learning [Internet]. New York, NY, USA: Association for Computing Machinery; 2004 [citado 18 de octubre de 2020]. p. 33. (ICML '04). Disponible en: https://doi.org/10.1145/1015330.1015373

40. Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A. diciembre de 1990;87(23):9193-6.

41. UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29

42. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. Sci Data. diciembre de 2017;4(1):170177.

43. Sawyer-Lee R, Gimenez F, Hoogi A, Rubin D. Curated Breast Imaging Subset of DDSM [Internet]. The Cancer Imaging Archive; 2016 [citado 18 de octubre de 2020]. Disponible en: https://wiki.cancerimagingarchive.net/x/lZNXAQ

44. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. J Digit Imaging. diciembre de 2013;26(6):1045-57.

45. MIAS Mammography [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://kaggle.com/kmader/mias-mammography

46. Shrivastava A, Chaudhary A, Kulshreshtha D, Prakash Singh V, Srivastava R. Automated digital mammogram segmentation using Dispersed Region Growing and Sliding Window Algorithm. En: 2017 2nd International Conference on Image, Vision and Computing (ICIVC). 2017. p. 366-70.

47. Optimam mammography image database and viewing software | Commercial Partnerships | Cancer Research UK [Internet]. [citado 18 de octubre de 2020]. Disponible en: http://commercial.cancerresearchuk.org/optimam-mammography-image-database-and-viewing-software

48. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. enero de 2020;577(7788):89-94.

49. Hastie T, Tibshirani R, Friedman J. The elements of statistical Learning. Segunda. Springer International Publishing;

50. Decision Tree Ensembles- Bagging and Boosting | by Anuja Nagpal | Towards Data Science [Internet]. [citado 18 de octubre de 2020]. Disponible en: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9

51. Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping [Internet]. [citado 12 de diciembre de 2020]. Disponible en: https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap

52. La leptina: un indicador más para cáncer de mama [Internet]. CuidatePlus. 2003 [citado 12 de diciembre de 2020]. Disponible en: https://cuidateplus.marca.com/enfermedades/cancer/2003/04/07/leptina-indicador-cancer-mama-4420.html

53. Kuhn M. The caret Package [Internet]. [citado 12 de diciembre de 2020]. Disponible en: https://topepo.github.io/caret/index.html

54. C5.0 Decision Trees and Rule-Based Models [Internet]. [citado 12 de diciembre de 2020]. Disponible en: https://topepo.github.io/C5.0/

55. randomForest function | R Documentation [Internet]. [citado 12 de diciembre de 2020]. Disponible en: https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest

56. gbm function | R Documentation [Internet]. [citado 12 de diciembre de 2020]. Disponible en: https://www.rdocumentation.org/packages/gbm/versions/2.1.8/topics/gbm