UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRICA

MODELO DE APRENDIZAJE AUTOMÁTICO PARA PREDECIR EL RIESGO CARDIOVASCULAR USANDO VARIABLES CLÍNICAS Y DEL DIARIO VIVIR

POR

Rocío Arantzazú Hernández Torres

Informe Final Memoria de Titulo presentada a la Facultad de Ingeniería de la Universidad de Concepción paraoptar al grado académico de Ingeniero/a Civil Biomédica

Profesores Guía
Rosa Figueroa I.

Profesional Supervisor
Álvaro Saldaña V.
Miguel Figueroa T.

Agosto
Concepción
(Chile)

Agradecimientos

A mi madre querida, que estuvo conmigo desde el inicio de los tiempos, me motivó a buscar una carrera que me llenara y me acompañó en los buenos, malos y muy malos momentos en mis años de estudios universitarios. Fue una inspiración y mi motivación. Gracias por alentarme a siempre seguir adelante.

A mi amiga y compañera, que me acompañó desde el día uno en la carrera, Alexandra. Nos vimos en nuestros peores momentos, pasamos una pandemia viéndonos más de 12 horas seguidas diarias, y pese a todo, seguimos alentándonos mutuamente, incluso cuando de broma pedíamos dejar la carrera. Siempre apoyándome y obligándome a ir a clases cuando no quería levantarme de la cama.

A mi mejor amiga el mundo mundial, Dani, que lograba calmarme cuando veía todo negro y lo lograba entrar en razón, dándome palabras de aliento y cocinándome panes dulces cada vez que notaba lo estresada que estaba por algún ramo.

A mi hermano Nico, por tus charlas de hermano mayor, por enseñarme a ser paciente y aceptar el lado humanista de las personas.

A mis personas favoritas de Angol, Gina y Marcia. Las dos siempre tan atentas por mí, llenando mi corazón de buenos momentos, risas y sabiduría. Acompañándome en cada cumpleaños sin falta.

A mi tía Viki y a Pepe. Su apoyo y afecto contínuo desde pequeña, siempre invitándome a comer, pendientes de mis estudios, alegrándose por mis pequeñas victorias, fueron un factor fundamental a mi motivación con mis estudios.

A mis amigos de la resi, Violeta, Zaida, Rodrigo, Coni y Ana. Gracias por ser tan pacientes conmigo, aguantando mi mal humor cuando estaba colapsada por la U. Por brindarme buenos momentos en una de la residencia más toxicas de Conce. Y por enseñarme a salir más de mi zona de confort.

A mi profesora guía Rosa Figueroa, que ha estado conmigo brindándome su apoyo y conocimientos para enfrentar esta tarea.

Gracias a todos por estar. Les estoy eternamente agradecida por tenerlos en mi vida personal, académica y profesional.

Resumen

Las enfermedades cardiovasculares (ECV) son una de las principales causas de defunción en el mundo. Solo en Chile representan el 23% de los decesos. Por esta razón que existen diversos programas mundiales y nacionales para prevenir eventos cardiovasculares, tales como Programa de Salud Cardiovascular (PSCV), los promueven un estilo de vida saludable, como llevar una dieta balanceada o ejercitar frecuentemente. Pese a esto, muchas personas cuyos hábitos de vida son considerados saludables, también pueden llegar a sufrir una ECV, lo que ha motivado a muchos investigadores a estudiar las causas de estas enfermedades, buscando métodos para medir el riesgo cardiovascular.

Por esta razón que en este trabajo de memoria de título se busca estudiar la forma de predecir el riesgo cardiovascular usando variables del diario vivir presentes en la base de datos Cardiovascular Health Study (CHS). Para lograr el objetivo propuesto, primero se hizo un análisis exploratorio dentro de la base de datos y un procesamiento de las características seleccionadas, con el fin de identificar las variables más relevantes para entrenar el modelo usando distintos métodos. Posteriormente se entrenaron los modelos con los distintos métodos seleccionados. Para la selección de características, se utilizaron métodos de filtros, de envoltura, y algoritmos específicos como FeatureWiz, SelectKBest y SelectFromModel. Además, se implementaron redes neuronales para mejorar la capacidad de predicción.

Los resultados obtenidos no fueron excelentes, pero tampoco desalentadores. El modelo con el mejor rendimiento fue el modelo entrenado con el método de envoltura, alcanzando un 61% de accuracy y un puntaje f1 de 0,59. Estos resultados, aunque no óptimos, indican que las variables seleccionadas para este proyecto no fueron suficientemente relevantes para predecir el riesgo cardiovascular. Sin embargo, este resultado no descarta la posibilidad de seguir investigando si eventos de la vida diaria pueden afectar el correcto funcionamiento del músculo cardíaco, lo que sugiere la necesidad de aplicar más técnicas de ingeniería de datos.

Abstract

Cardiovascular diseases (CVD) are one of the leading causes of death worldwide. In Chile alone, they account for 23% of deaths. For this reason, there are various global and national programs to prevent cardiovascular events, such as the Cardiovascular Health Program (PSCV), which promote a healthy lifestyle, including maintaining a balanced diet and exercising frequently. Despite this, many people with habits considered healthy can still suffer from CVD, motivating many researchers to study the causes of these diseases and seek methods to measure cardiac risk.

It is for this reason that this thesis work seeks to study the way to predict cardiac risk using daily living variables present in the Cardiovascular Health Study (CHS) database. To achieve the proposed objective, an exploratory analysis was first conducted within the database and the selected features were processed in order to identify the most relevant variables for training the model using different methods. Subsequently, the models were trained using the selected methods. For feature selection, filter methods, wrapper methods, and specific algorithms such as FeatureWiz, SelectKBest, and SelectFromModel were used. Additionally, neural networks were implemented to improve prediction capability.

The results obtained were not excellent, but neither were they discouraging. The best-performing model was logistic regression, achieving an accuracy of 61% and an F1 score of 0.59. These results, although not optimal, indicate that the variables selected for this project were not sufficiently relevant to predict cardiac risk. However, this result does not rule out the possibility of further investigating whether daily life events can affect the proper functioning of the heart muscle, suggesting the need to apply more data engineering techniques.

Tabla de contenido

Lista de Tablas

Lista de Figuras

Capítulo 1. Introducción

1.1. Introducción General

Las enfermedades cardiovasculares (ECV) son un grupo de enfermedades que afectan al corazón y vasos sanguíneos. Dentro de las enfermedades cardiovasculares más comunes están la insuficiencia cardíaca, infarto de miocardio, angina de pecho, arritmias, pericarditis, cardiopatías congénitas, entre otros [1]. En Chile, las ECV son la principal causa de muerte, con un 23% de defunción en 2020 [2]. Una de las formas que existen para medir las ECV y tratar de prevenirlas es calculando el riesgo cardiovascular, el cual mide la posibilidad de sufrir alguna de estas enfermedades. En Chile existe el Programa de Salud Cardiovascular (PSCV), en el cual se monitorea a las personas con distintas enfermedades tales como la Hipertensión, Diabetes y los que presentan otro tipo de riesgo como los consumidores de tabaco, personas con obesidad o con antecedentes de ECV [3]. Existen muchos factores que pueden ayudar a desarrollar una afección cardíaca. Están los factores de riesgo que no se pueden controlar, tales como la edad, el sexo o los antecedentes genéticos de ECV, sin embargo, ¿qué hay de los factores que sí pueden ser controlados? Estos factores se asocian al estilo de vida de las personas: si hacen o no deporte, el trabajo en el que se desempeñan y su alimentación, entre otras. Todas estas variables son factores que afectan el desempeño del corazón día a día [4].

Dentro de este contexto, es sumamente importante lograr predecir el riesgo cardiovascular en una etapa temprana para evitar complicaciones a futuro o posibles decesos. Esto denota un gran desafío, considerando la cantidad de factores que pueden afectar el sistema cardiovascular. Además de los ya mencionados, existen muchos otros factores considerados dentro de los eventos de la vida diaria de las personas, y cómo estos factores pueden contribuir a que se desarrolle alguna afección cardíaca.

La base de datos Cardiovascular Health Study (CHS) es una base de datos que recopiló información de personas en relación al estudio de ECV desde el año 1987 y tuvo 10 año de seguimiento clínico y 7 años de seguimiento telefónico [5]. Los datos recopilados son variables que muestran si las personas dentro del estudio tienen alguna enfermedad cardíaca, si son fumadores, si tienen alguna otra enfermedad y se estudia además los eventos de la vida diaria, con el fin de ver cómo distintos factores pueden generar una ECV, o agravarla [6].

Para esto, se propone desarrollar un modelo de aprendizaje automático, que utilice los datos presentes en la base de datos CHS y entrenarlos para poder predecir el riesgo cardiovascular mediante el análisis de variables relacionadas a eventos de la vida diaria. Así se busca brindar a las personas una detección oportuna de posibles enfermedades relacionadas a las ECV, aunque no tengan factores de riesgo convencionales mencionados.

1.2. Objetivos

1.2.1 Objetivo General

Desarrollar un modelo de aprendizaje automático capaz de predecir el riesgo cardiovacular utilizando variables clínicas y de la vida diaria de pacientes presentes en la base de datos Cardiovascular Health Study (CHS).

1.2.2 Objetivos Específicos

- Identificar variables que reflejen la ocurrencia de eventos de la vida diaria disponibles en la base de datos CHS.
- Explorar distintos métodos de selección de características para obtener el mejor rendimiento.
- Entrenar modelos de aprendizaje automático que sean capaces de predecir la ocurrencia de un evento cardiovascular utilizando variables sociodemográficas y de eventos de la vida diaria.
- Evaluar y comparar los modelos entrenados en términos de su exactitud, precisión, sensibilidad, etc.

1.3. Alcances y Limitaciones

i. Alcance 1

- Para el desarrollo del modelo se usarán los archivos baseline de la base de datos Cardiovascular Health Study (CHS), que abarca los dos primeros años del estudio.

- Para el uso de la base de datos se requiere autorización de uso, lo cual fue tramitado por la académica Rosa Figueroa en el marco del proyecto Multimorbidities Project autorización RMDA 13032, el que también fue revisado por el comité ético científico de la Universidad de Concepción, con fecha enero 2019 y cuenta con una extensión de plazo otorgado por el banco de datos BioLincc (custodio de CHS) hasta el 2026.

ii. Limitación 1

- La selección de un algoritmo adecuado, debido a la complejidad de la base de datos y a la cantidad de algoritmos disponibles.
- El gran tamaño de la base de datos CHS, que abarca varios años de recolección de datos requirió un proceso adicional para comprimir y asegurar la coherencia de la información.

1.4. Metodología

Para el trabajo de esta memoria de título se decidió la metodología en cuatro partes específicas: análisis exploratorio de la base de datos Cardiovascular Heath Study (CHS), definición de métricas de evaluación, procesamiento de la base de datos para terminar con el desarrollo y evaluación del modelo de aprendizaje (Figura 1).

En la primera parte se realizó un análisis exploratorio de la base de datos Cardiovascular Heath Study con la idea de identificar y definir claramente las variables de entrada y la variable objetivo, la cual indica si un paciente experimentó o no un evento cardiovascular. La variable objetivo es binaria, con valores de 1 para aquellos pacientes que sí tuvieron un evento y 0 para aquellos que no. De acuerdo con el análisis realizado, las variables se clasificaron en categóricas o numéricas, lo que ayuda a definir los métodos de selección de características para la clasificación.

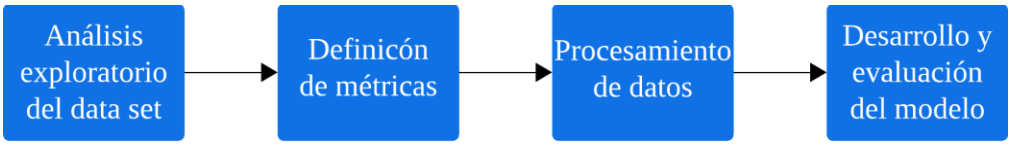


Figura 1: Esquema de la metodología que se siguió para el desarrollo del modelo de aprendizaje.

Posteriormente, se estudiaron las métricas más relevantes para poder hacer una evaluación completa de los modelos entrenados. La métrica principal fue exactitud, pues define la cantidad de aciertos del modelo. Después, se decidió evaluar los modelos con el puntaje f1, el cual es la media armónica de la precisión y la sensibilidad. Esta última métrica indica si el modelo está sesgado o no y qué tan consistente son las predicciones que hace el modelo.

Durante el preprocesamiento de la base de datos CHS, se detectó un desbalance en el número de pacientes entre las bases de datos presentes en los tres archivos baseline que fueron utilizados. Esto requirió la eliminación de 687 pacientes presentes en uno de los archivos y ausentes en los otros dos. Después, para mejorar la salida, se trabajó el desbalance de clases equilibrando las clases del modelo, obteniendo finalmente 3096 pacientes. Por último, se trataron los datos faltantes del conjunto de datos rellenándolos con la moda.

Para la selección de características, se utilizaron varios métodos: el método de filtro, métodos envolventes, y el uso de tres algoritmos se selección de características. Para el método de filtro fue necesario separar las variables en variables categóricas y numéricas para hacer la correlación de chi cuadrado y LDA respectivamente. Para el método de envoltura, que requiere de un modelo para hacer la selección, simplemente se decidió por usar el modelo de regresión logística. Y en cuanto a los tres algoritmos de selección se utilizaron SelectKBest, FeatureWiz y SelectFromModel. Este proceso permitió seleccionar las variables más relevantes para entrenar el modelo.

Por último, para la selección de características con le método de filtros y los tres algoritmos de selección se usó los modelos Decision Tree, Random Forest, Naive Bayes y Support Vector Machine (SVM). Para el método de envoltura se utilizó el modelo Logistic Regression. Además, se entrenaron los modelos Decision Tree, Random Forest, Naive Bayes y SVM con el total del conjunto de datos para tener una referencia. Posteriormente, se implementó validación cruzada a todos estos modelos menos Logistic Regression. Finalmente se implementó una red neuronal

1.5. Temario

(i) **Capítulo 1:** Se introduce al tema, resumiendo lo que será el proyecto, se describen el objetivo general, los objetivos específicos, mencionando los alcances y limitaciones, y un resumen de la metodología seguida para el desarrollo de este informe.

(ii) **Capítulo 2:** Conceptualización de los temas más importantes a considerar para el desarrollo de este proyecto. Incluye una revisión bibliográfica detallada con los conceptos clave con la cual se fue construyendo un marco teórico del tema en estudio y comparando los estudios de otros autores con el tema principal de este proyecto.

(iii) **Capítulo 3:** Descripción de la metodología empleada a lo largo del proyecto, explicando desde la revisión bibliográfica hasta la construcción de la matriz que se usará para el desarrollo del modelo de predicción.

(iv) **Capítulo 4:** Se explican los resultados obtenidos por el desarrollo del modelo de aprendizaje. Se detallan los algoritmos usados y el rendimiento de cada uno y las variables usadas para entrenarlos.

(v) **Capítulo 5:** Se presentan las conclusiones del proyecto, evaluando si se cumplieron los objetivos o no. Se describe el trabajo a futuro de este proyecto.

(vi) **Capítulo 6:** Se define la nomenclatura usada en el informe.

(vii) **Capítulo 7:** Se muestran las referencias bibliográficas utilizadas a lo largo de este informe.

Capítulo 2. Marco Teórico

2.1. Introducción

En el siguiente capítulo se revisarán las temáticas que permitirán enmarcar teóricamente este trabajo de memoria de título, como por ejemplo la conceptualización de enfermedades cardiovasculares, riesgo cardiovascular, selección de características, modelos de aprendizaje automático y los algunos de los algoritmos usados en la memoria. Además, se introducen los modelos de predicción de riesgo cardiovascular y se interioriza en la base de datos Cardiovascular Health Study (CHS). Todo para comprender el funcionamiento de los modelos de aprendizaje automático de predicción.

2.2. Enfermedades cardiovasculares

La OMS define salud como "un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades" [7]. Aunque es cierto, esta definición entró en vigor en 1948, aún la usan y aceptan muchos. Al mismo tiempo, también se define la enfermedad como una "alteración del estado fisiológico en una o varias partes del cuerpo (...)" [8]. Aun cuando la OMS utiliza los términos enfermedad y afección como diferentes, en este informe se usarán como sinónimos.

El sistema cardiovascular lo forman el corazón, la sangre y los vasos sanguíneos, que incluyen venas, arterias y capilares que ayudan con la irrigación de distintos órganos dentro del cuerpo, usando la sangre como transporte (Figura 2). El corazón funciona como una bomba la cual hace circular la sangre a través de los vasos sanguíneos transportando distintas sustancias fundamentales para el organismo como oxígeno, hormonas, dióxido de carbono y nutrientes [9].

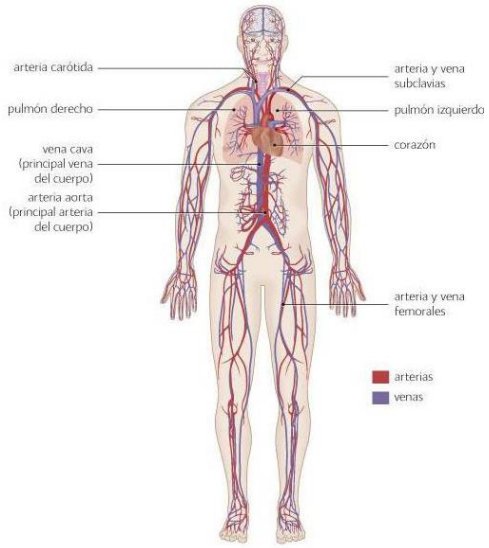


Figura 2: Sistema cardiovascular constituido por arterias, venas y el músculo cardiovascular [9].

Las alteraciones que afectan al sistema cardiovascular se denominan Enfermedades Cardiovasculares (ECV). Existen numerosas enfermedades cardiovasculares, algunas de estas afecciones impactan los vasos sanguíneos, produciendo un déficit de irrigación en distintas partes del cuerpo, a modo de ejemplo tenemos la cardiopatía coronaria, donde se ven afectados los vasos sanguíneos que irrigan el músculo cardiaco; enfermedades cerebrovasculares, que afectan los vasos que irrigan el cerebro; y las arteriopatías, cuyos vasos afectados irrigan miembros periféricos. También hay ECV que afectan directamente al músculo cardíaco, como la cardiopatía reumática, o afecciones que están presentes desde el nacimiento como las cardiopatías congénitas. También existen ECV que afectan a la sangre, como trombosis venosas profundas y embolias pulmonares.

Las enfermedades cardiovasculares se desarrollan a través de varios procesos biológicos, por ejemplo, los accidentes cardiovasculares (ACV) se pueden producir por una acumulación de grasa en las paredes de los vasos sanguíneos (aterosclerosis), lo que reduce o bloquea el flujo de sangre. También pueden producirse por la formación de coágulos (trombosis) o por la rotura de un vaso sanguíneo (hemorragia).

Pese a conocer el funcionamiento biológico de estas enfermedades, determinar la razón exacta que causa estas ECV no es posible, sin embargo, sí podemos decir que varios factores pueden favorecer la aparición de estas. Entre estos factores están algunos hábitos de vida poco saludable como lo son

el tabaquismo, una dieta deficiente, sedentarismo, así como condiciones médicas preexistentes como la hipertensión, diabetes o hiperlipidemia [10].

2.3. Riesgo cardiovascular

El riesgo cardiovascular se define como la probabilidad de sufrir una ECV. Para categorizar el riesgo cardiovascular, se deben considerar varios factores de la vida del paciente, los que incluyen factores de riesgo no modificables y los modificables.

Los factores de riesgo cardiovascular no modificables son aquellos propios de cada individuo y generalmente son condiciones que no se pueden evitar, tales como el sexo, la herencia, la edad o si ha sufrido alguna enfermedad coronaria previa [11].

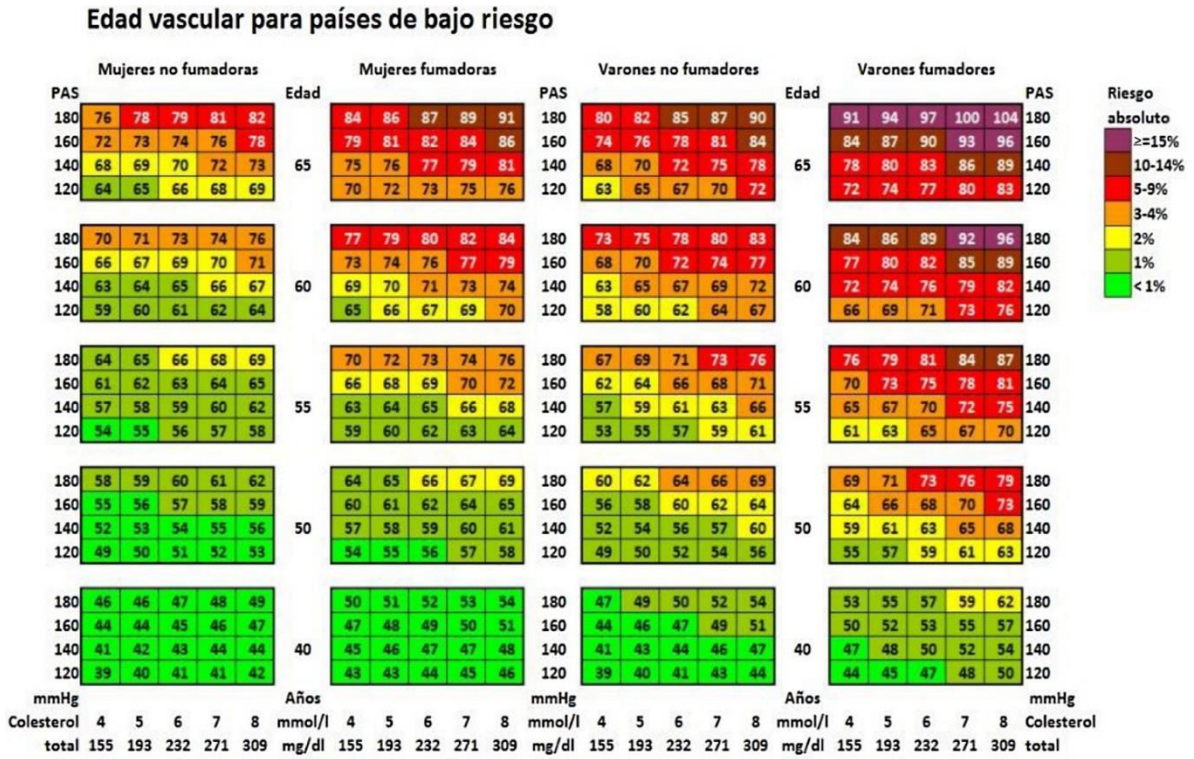


Figura 3: Estratificación de riesgo cardiovascular para países de bajo riesgo por edad y sexo entre fumadores y no fumadores. Los colores muestran la probabilidad de riesgo cardiovascular, siendo morado lo más probable y verde claro lo menos probable [12].

Los hombres tienen más probabilidad de contraer afecciones cardiovasculares primero que las mujeres, tal y como se muestra en la Figura 3, en donde además se observa que el riesgo cardiovascular aumenta con la edad.

Los factores de riesgo cardiovascular modificables son aquellos en los que tenemos control, ya que se pueden corregir e incluso pueden ser eliminados con el tiempo. Es en estos factores de riesgo en los que las autoridades de salud suelen enfocarse a la hora de generar políticas públicas tendientes a la prevención y manejo de las ECV. Los factores de riesgo modificables más comunes son tener la presión arterial alta, obesidad, tabaquismo, sedentarismo, niveles altos de colesterol, consumo elevado de alcohol y estrés. Todos estos factores de riesgo pueden ser reducidos con una buena dienta, actividad física, dormir bien y controles periódicos de sangre [11]. La figura 3 ilustra la diferencia en el riesgo cardiovascular entre las personas que consumen tabaco con las que no. De esta figura se puede observar que el consumo de esta sustancia aumenta aproximadamente un 1% en la probabilidad de desarrollar una ECV.

2.4. Selección de características

La selección de características es el proceso de elegir las mejores variables dentro de un conjunto de datos para entrenar un modelo. Este paso es fundamental en el aprendizaje automático, ya que no todas las características disponibles en una base de datos son útiles para el entrenamiento. En ocasiones, cuando los conjuntos de datos son grandes, algunos datos pueden no mostrar información importante, lo que introduce ruido y afecta negativamente el rendimiento del modelo. "Si entra basura, saldrá basura", entendiendo basura como el ruido en el entrenamiento [13]. Para evitar el ruido y mejorar el rendimiento del modelo, es necesario realizar una correcta selección de características.

Existen diferentes métodos para la selección de características. La elección del método adecuado dependerá de las necesidades específicas y del tipo de datos disponibles. Entre estos métodos se incluyen el método de filtro, los métodos envolventes, o el uso de algoritmos preexistentes para seleccionar las mejores características.

2.4.1 Método de filtro

El método de filtro se usa como preprocesamiento, independientemente de los algoritmos de modelos de aprendizaje automáticos que se puedan emplear (Figura 4). Este método sirve para identificar características constantes, repetidas y correlacionadas. Es en muchos casos considerado el primer paso para la selección de características. Con este método, las variables se seleccionan según métricas estadísticas, puntuadas con su correlación entre variables de entrada y la de salida.

Figura 4: Esquema del funcionamiento el método de filtro para selección de características [14].

Dentro de este contexto, existen diversos coeficientes de correlación, y el uso de cada uno dependerá del tipo de variable, ya sea categórica o numérica. Una variable es de tipo categórica cuando describe datos cualitativos, es decir, datos que no se pueden cuantificar, como el sexo de una persona, el tipo de dieta que sigue, o si fuma, entre otros. Una variable es numérica cuando describe algo que sí se puede cuantificar, como la edad de una persona, la presión arterial o el peso. Dentro de las variables numéricas existen más subdivisiones, lo que significa que una variable puede ser discreta, continua, binaria, etc. [15]. Para simplificar el informe, se hablará de variables numéricas o continuas; sin embargo, es importante tener en cuenta que no son sinónimos. En la siguiente tabla se muestra qué coeficientes de correlación es adecuado para cada tipo de variable [16].

Tabla 1: Coeficientes de correlación según tipo de variable [16].

Característica\Predicción	**Continuo**	**Categórico**
Continuo	Correlación de Pearson	LDA
Categórico	ANOVA	Chi-cuadrado

Si la variable de salida, o variable objetivo es de tipo continua puede ser correlacionada con variables de entrada continuas a través de la correlación de Pearson o con variables de entrada categóricas a usando de ANOVA (Análisis de varianza). Por otro lado, si la variable de salida es de tipo categórica, puede ser correlacionada con variables de entrada tipo continuas usando el análisis de discriminante lineal (LDA) o con variables de entrada categóricas usando Chi-cuadrado.

- **Correlación de Pearson**: Mide la dependencia lineal entre variables continuas.
- **ANOVA**: El análisis de la varianza compara las medias de las variables mediante el estudio de las varianzas, calculando las varianzas de las medias [17].
- **LDA**: Modela la distribución de datos para cada clase, utilizando técnicas como el teorema de Bayes [18].
- **Chi-cuadrado**: Evalúa la independencia entre dos o más variables categóricas. Mide si la diferencia entre datos se debe al azar o existe alguna relación [19].

2.4.2 Método de envoltura

El método de envoltura suele ser usando cuando se trabaja con muchos datos. Debido a esto, el uso de este método requiere un alto costo computacional. Este método selecciona un subconjunto de características para entrenar modelos, evaluando su desempeño. Luego, se agregan o quitan características según el rendimiento de los modelos. Este proceso se repite hasta encontrar el mejor subconjunto de características cuyo modelo haya tenido el mejor rendimiento (Figura 5). Además, este método necesita de un algoritmo de aprendizaje automático, y usa el rendimiento del modelo como criterio de evaluación [16].

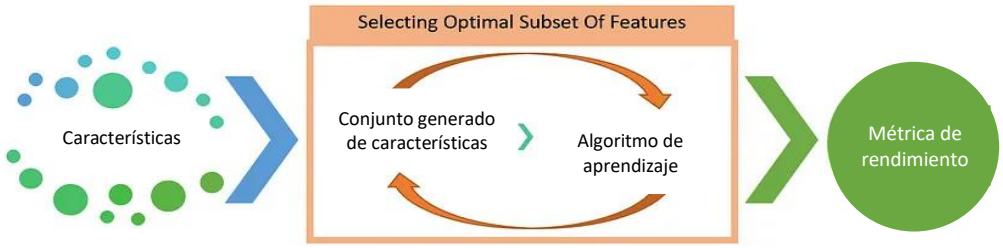


Figura 5: Esquema básico del funcionamiento el método envolvente para selección de características [14].

Algunos de los métodos envolventes más frecuentes son Selección de Características Hacia Delante (en inglés *Forward Feature Selection*), Eliminación Hacia Atrás (en inglés *Backward Elimination*) y Eliminación de características recursivas (en inglés *Recursive Feature Elimination*) [16].

Selección hacia delante. Inicia sin ninguna característica y en cada iteración va agregando una nueva característica hasta encontrar las mejores [20].

Eliminación hacia atrás. Inicia con el conjunto total de características y elimina la característica que mejore el rendimiento del modelo. Se repite este proceso hasta encontrar el mejor conjunto de características [20].

Eliminación de características recursivas. Inicia con un conjunto de características, evalúa la importancia de cada característica en función del rendimiento del modelo entrenado. Luego descarta las características con menos importancia. Este proceso se repite hasta llegar a la mejor combinación de características [20].

Además de estos métodos existe un método complementario llamado Análisis de Componentes Principales (PCA), que se utiliza para reducir la dimensionalidad con el fin de identificar las características más relevantes. Este método transforma un conjunto de datos y lo reduce a componentes principales [21].

2.4.3 Algoritmos de selección de características

Aunque es cierto, los métodos mencionados son algoritmos para la selección de características, existen otros que no entran en la clasificación de ninguno de los métodos descritos. En este informe, se utilizaron dos de estos algoritmos: FeaturesWiz, SelectKBest y SelectFromModel.

FeatureWiz.

Es un algoritmo relativamente nuevo, usado para la selección de características. Este algoritmo utiliza dos métodos: SULOV (Searching for Uncorrelated List Of Varibales) y XGBoost. FeatureWiz primero utiliza el método de SULOV, que busca los pares de características que estén más correlacionadas que exceda un umbral predefinido, generalmente 0,7. Cada par es puntuado con su puntaje de información mutua (MIS score) respecto de la variable objetivo; el par con la puntuación MIS más baja es eliminado. Luego el algoritmo emplea XGBoost, que toma las características restantes que el método SULOV no seleccionó y las divide en conjuntos de entrenamiento y de prueba. Así, XGBoost encuentra las X mejores características de uno conjunto. Luego repite el proceso con otro grupo de características. Este proceso se repite cinco veces, generando cinco modelos de XGBoost. Finalmente combinan las mejores características obtenidas por SULOV y XGBoost [22].

SelectKBest.

Este es otro algoritmo que sirve para seleccionar las características más relevantes. Es de la librería Scikit-Learn. Este algoritmo es uno de los muchos de Scikit-Learn que selecciona las mejores características usando pruebas estadísticas, en donde SelectKBest va descartando variables y dejando las *k* variables con mejor puntuación. Esto lo puede hacer especificando la función de puntuación y las *k* mejores variables que se desee dejar. La función de puntuación puede ser de regresión o clasificación. Para las de regresión están las funciones *r_regression*, *f_regression* y *mutual_info_regression*; para las de clasificación están las funciones *chi2*, *f_classif* y *mutual_info_classif* [23].

Validación cruzada.

La validación cruzada es un método que sirve para evaluar la capacidad de un modelo para hacer predicciones precisas. Esto lo hace dividiendo un conjunto de datos en k pliegues y crea un modelo para cada subconjunto, evaluando cada uno de ellos. Al tener la precisión de cada modelo entrenado se puede concluir la precisión. [24]

SelectFromModel.

SelectFromModel es un algoritmo que, como dice su nombre en inglés, selecciona las características más relevantes de un modelo de aprendizaje automático previamente entrenado. Para usarlo se requiere de un modelo como un Árbol de Decisión o un Random Forest, el cual es entrenado,

y luego, usando SelectFromModel selecciona las mejores características del modelo, usando el nivel de importancia de cada variable otorgado por el modelo [25].

2.5. Modelo de aprendizaje automático

Un modelo de aprendizaje automático o machine learning (ML) es, como lo dice su nombre en inglés, una máquina o computador que es capaz de aprender a tomar decisiones por sí solo. Para ello, requiere de un conjunto de datos de entrada que, mediante el reconocimiento de patrones permita construir un modelo de aprendizaje automático. Luego, el modelo entrenado se prueba realizando predicciones en un set de prueba. Este proceso se ilustra en la Figura 6.

Los conjuntos de datos de aprendizaje deben ser adquiridos de una base de datos. En caso de ser necesario, pasan por un preprocesamiento, en el que se realizan acciones para normalizar los datos, filtrar de acuerdo con la importancia del dato, verificar si existen valores perdidos, entre otros. Seguido de ello, se inicia el proceso de extracción de características, en el que analizamos el set de datos y sus variables en busca de aquellas variables o características que representen la información y que permita ejecutar la tarea de aprendizaje. Estas características se miden por su poder predictivo y se seleccionan para obtener el mejor conjunto de características para discriminar entre distintas clases [26]..

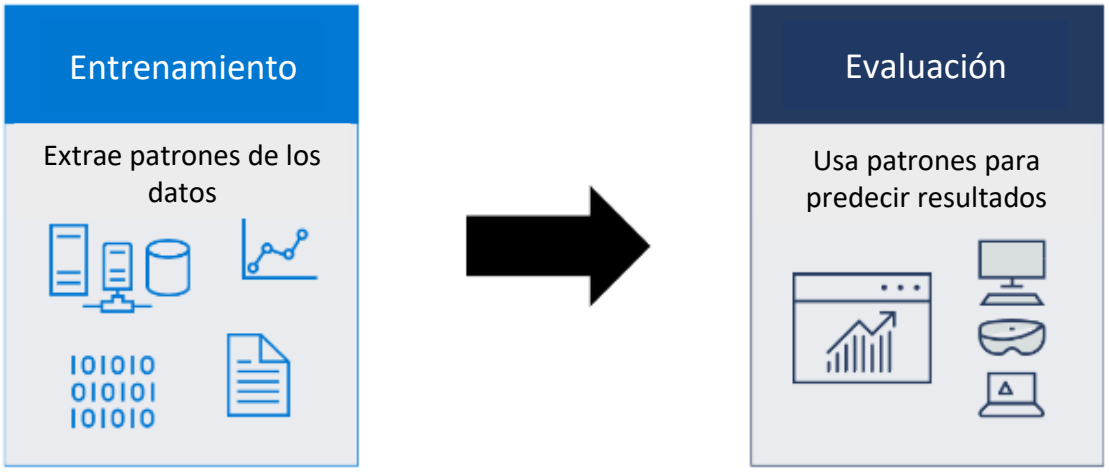


Figura 6: Entrenamiento de un modelo. Conjunto de entrenamiento extrae patrones de una base de datos para que el conjunto de prueba use los patrones extraídos para predecir resultados [26].

Existen distintos tipos de algoritmos de aprendizaje, entre los más conocidos destacan el aprendizaje supervisado, aprendizaje no supervisado y el aprendizaje por refuerzo.

2.5.1 Aprendizaje automático supervisado

Este aprendizaje utiliza datos previamente etiquetados para poder construir los modelos. Son estas etiquetas las que se desean predecir y, a la vez, se usan para entrenar el modelo, comparando las predicciones iniciales con las reales, ajustando sus parámetros cuando haga una predicción incorrecta [27]. Una explicación más visual de esto se muestra en la figura 7. Generalmente, una vez entrenado el modelo, se somete a una evaluación de rendimiento, en la que se mide la exactitud y la precisión, para asegurar que el modelo prediga correctamente y no memorizando las etiquetas, para evitar sobre ajuste.

La base de datos CHS, como ya se mencionó, es un conjunto de datos sobre las ECV ya etiquetadas, por lo que es conveniente para este proyecto utilizar un modelo de aprendizaje supervisado. Más adelante se describirán los usado en este proyecto.

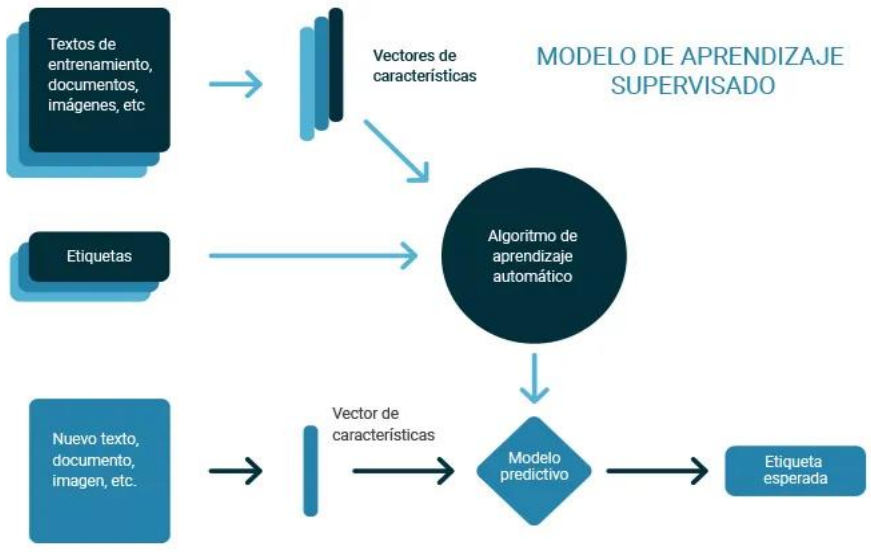


Figura 7: Diagrama de flujo de Aprendizaje Supervisado. Se utilizan los datos de entrada: texto de entrenamiento, y de salida: las etiquetas reales, para entrenar el modelo predictivo y obtener la etiqueta esperada [27].

2.5.2 Modelo Árbol de decisión

Los modelos de árboles de decisión son de los más usuales de aprendizaje automático. Son un método de aprendizaje supervisado utilizado para tareas de clasificación y regresión. Este tipo de algoritmo generalmente no requiere de normalización de datos y su objetivo es predecir una variable.

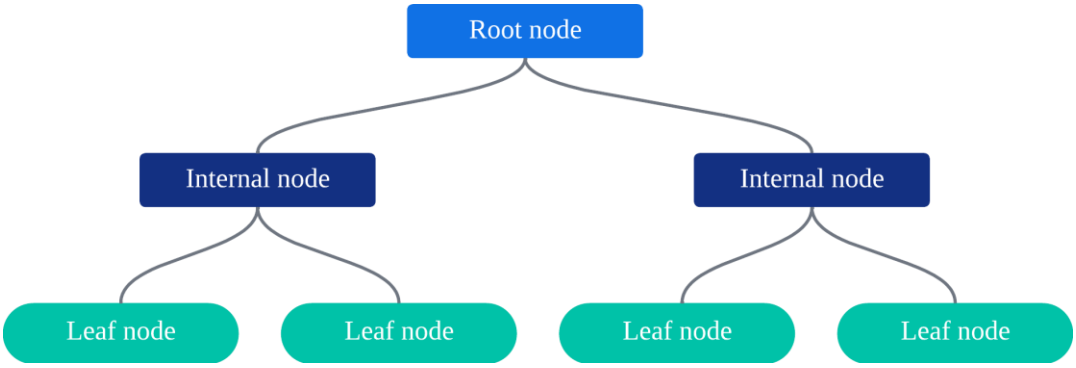


Figura 8: Diagrama general de un árbol de decisión con su estructura básica: nodo raíz, nodos internos y nodos hoja [28].

Los árboles de decisión funcionan gracias a la presencia de nodos raíz, nodos internos, ramas y hojas, tal como se muestra en la Figura 8, en donde las ramas vendrían siendo las líneas que conectan cada nodo y hoja [28]. Esto lo hace a través de un proceso recursivo, en el que, en cada nodo se selecciona la mejor característica para dividir los datos. Este proceso se repite hasta que se cumpla el criterio de parada, es decir, que no haya más características por dividir, será entonces que el árbol habrá terminado de predecir.

Usar árboles de decisión puede ser útil para interpretar datos, pues se puede entender mucho más visualmente, y el concepto es más fácil de digerir, pues es un modelo White box, lo que significa que se puede ver cómo predice. Sin embargo, es fácil caer en sobreajuste pues es sensible al ruido [29]. Además, para las bases de datos extensas puede llegar a ser más engorroso y poco visual.

2.5.3 Modelo Random Forest

El algoritmo de random forest es otro método de aprendizaje automático supervisado que sirve para entrenar modelos de clasificación y de regresión. Este se va construyendo a partir de muchos arboles de decisión.

Como se mencionó anteriormente, los árboles de decisión toman todo el conjunto de datos y consideran todas las particiones posibles hasta llegar a la mejor [28]. Random forest en contraste, toma una muestra del conjunto de datos de manera aleatoria. Así se construirán muchos árboles de decisión. Cada uno de los árboles de decisión tendrá entonces su propio subconjunto en donde se separarán nuevamente de forma aleatoria en dato de prueba o "muestra fuera de bolsa" y datos de entrenamiento. La aleatoriedad presente en el random forest agrega diversidad y reduce la correlación entre los árboles. Luego, una vez que cada árbol haya entrenado su modelo, este llegará a un resultado final, en donde si se entrenó como un modelo de clasificación, este elegirá la variable categórica más frecuente; si se entrena un modelo de regresión, se promediarán todos los árboles de decisión para llegar a un valor continuo. Una forma fácil de visualizar el random forest es como se muestra en la Figura 9.

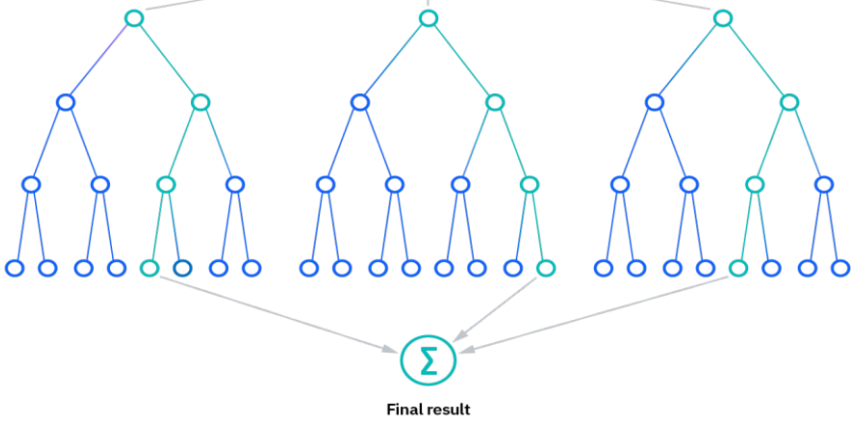


Figura 9: Diagrama general de un random forest. Conjunto de árboles de decisión predicen un resultado final, ya sea categórico o continuo [30].

Una de las ventajas de este algoritmo es que tiene baja probabilidad de entrenar un modelo con sobreajuste, pues, al tener muchos árboles de decisión con baja correlación entre ellos ayuda a reducir la varianza. Además, dada la forma de agrupar características hace que sea una herramienta eficaz y

precisa cuando existen datos faltantes. Sin embargo, el ejecutar este algoritmo puede llegar a ser bastante lento, pues entrena cada árbol de manera individual, sumándole el hecho de que necesita más recursos para almacenar datos [30].

2.5.4 Modelo Naïve Bayes

El modelo Naive Bayes es un algoritmo de aprendizaje automático en donde se asume que las variables predictoras son independientes entre sí. Este modelo de basa en el teorema de Bayes que describe la siguiente ecuación:

$$P(A|B) = \frac{P(B|A)*P(A)}{P(B)}, \text{ [31]}$$

Este modelo explica la probabilidad posterior de una clase A dada la característica B. es decir, la probabilidad de que un suceso A ocurra dado un suceso B. P(B|A) es la probabilidad de la característica B dada la clase A, P(A) es la probabilidad de la clase A y P(B) es la probabilidad previa de la característica B [32]. El teorema de Bayes puede ser utilizado para varias disciplinas, tales como de diagnóstico, clasificación de texto, entre otros [33].

Para entrenar un modelo utilizando Naive Bayes es necesario contar con un conjunto de datos de entrenamiento y con las clases que se desean predecir, de esta forma, la ecuación calculará la probabilidad futura de cada clase.

Este modelo se destaca su rapidez y la facilidad de predecir las clases, incluso para los multiclases. Este modelo funciona mejor que otros modelos de clasificación, como la regresión logística, si se asume la independencia de las variables predictoras y sin necesitar una gran cantidad de datos de entrenamiento. Por otro lado, si el conjunto de prueba no reconoce una característica, el modelo le asignará el valor de cero, sin poder realizar la predicción correctamente. Otra desventaja es que se asume la independencia de las variables, lo que no es necesariamente posible en la vida real [34].

2.5.5 Modelo SVM

El modelo Support Vector Machine (SVM) es otro modelo comúnmente usado en aprendizaje supervisado. Este algoritmo busca clasificar las variables en diferentes categorías correlacionando los datos en un espacio de las características. Este espacio de características es un espacio de las características originales transformado con la función kernel y es en este espacio donde se genera una 'línea' de separación (hiperplano). Posteriormente, con los datos ya clasificados, el modelo puede comenzar a hacer predicciones [35].

SVM es muy efectivo en espacios de grandes dimensiones. Además, existen diferentes funciones de kernel que se pueden implementar para determinar el hiperplano, estas pueden ser *linear*, *rbf* (por sus siglas en ingles función de base radial), *poly* y sigmoid [36]. La elección del kernel es una parte primordial a la hora de implementar SVM pues con este se define el rendimiento del modelo. Como se muestra en la Figura 10, dependiendo de la función elegida, la clasificación de características puede estar más cercana a la realidad o no.

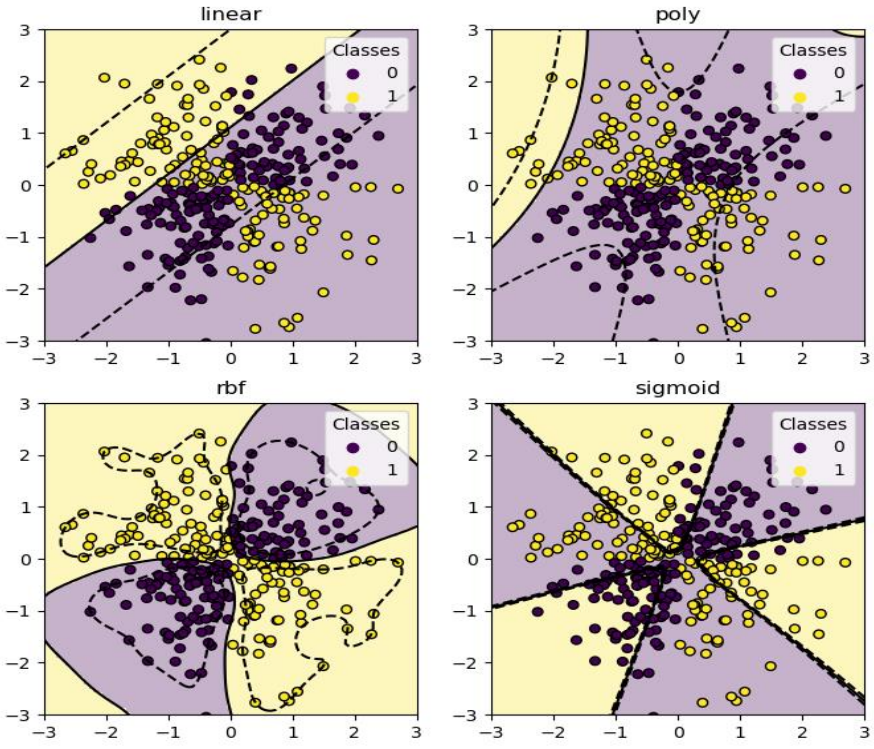


Figura 10: Representación gráfica de las funciones kernel 'linear', 'poly', 'rbf' y 'sigmoid' [37].

2.6. Redes Neuronales

Las redes neuronales han ganado popularidad estos últimos años. Se utiliza para muchas áreas, una de ellas es en modelos de predicción. Computacionalmente hablando, las redes neuronales computacionales tratan de imitar las redes neuronales cerebrales con una estructura relativamente sencilla para entender. Como se muestra en la Figura 11, una red neuronal, en su geometría más sencilla, consta con una neurona inicial, que interactúa con múltiples neuronas, hasta llegar a una neurona final, realizando la tarea deseada. En un mundo más complejo, no se tendrá una sola neurona, sino muchas, que interactuarán con más neuronas, y esas neuronas interactúan con múltiples neuronas más, hasta llegar a la neurona deseada que realizará el trabajo que se desea. Cada conjunto de neurona es llamado capa, y a cada línea que se traza entre neuronas se le otorgará un peso; este peso le da prioridad o le resta prioridad a cada neurona, o característica. Además, cada vez que una neurona le pasa información a otra, cuando entra a esta nueva neurona, necesitará una función de activación, la cual irá guiando y clasificando la información entregada para que se logre llegar al objetivo [38].

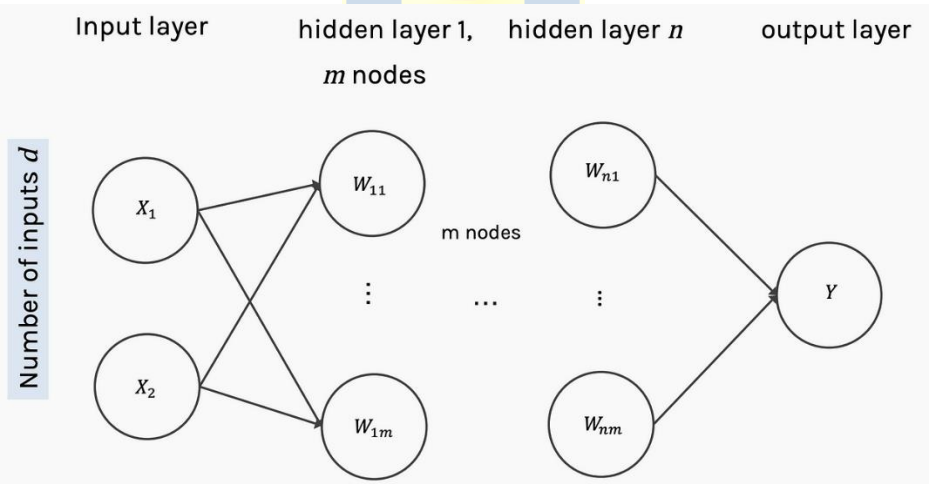


Figura 11: Geometría básica de una red neuronal [38].

Todas las redes neuronales necesitan un grupo de entrada y otro de salida para hacer predicciones, en donde por lo menos se necesitan tres: una capa de entrada, otra intermedia, y una capa de salida.

Existen diferentes funciones de activación, la más utilizada actualmente es la función ReLU, por sus siglas en inglés: Unidad Lineal Rectificada. Es una de las funciones no lineales más simples pues, si la entrada de la función es inferior a cero, la salida devuelve cero, en caso contrario, si la entrada es superior a cero, la salida devuelve ese valor. Esta función funciona bien en la mayoría de los casos, sin embargo, no se recomienda su uso en capas de salida, solo en capas densas. Para las capas de salida, existen otro tipo de función. La Sigmoidea, por ejemplo, es una buena función para ocupar en una capa de salida, pues sirve bastante para modelos cuya clase sea binaria. Sin embargo, esta función no funciona tan bien en las capas ocultas, pues uno de sus problemas es que no tocan el cero [38].

2.7. Modelos de predicción de riesgo cardiovascular

Los modelos de predicción de riesgo cardiovascular se han convertido en una herramienta muy útil en los últimos años, pues como ya se ha mencionado antes, una de las principales causas de decesos a nivel mundial son las enfermedades cardiovasculares (ECV). Solo en Chile se registran un 23% de casos fatales anuales debido a estas enfermedades [3]. Para predecir estos eventos, se debe tener en claro cómo manejar la base de datos que se usará para entrenar el modelo. Para efectos de este proyecto se utilizará la base de datos CHS, la cuál es un estudio retrospectivo, lo que significa que compara las personas sanas y no sanas, en el caso partículas de este informe se refiere a personas que sufrieron una afección cardíaca o que tienen un mayor riesgo de sufrirla y así, se busca una relación entre ambos grupos de personas [39]. A continuación, se presentan algunos estudios en donde se busca predecir de riesgo cardiovascular usando diferentes métodos.

2.7.1 Modelos de regresión

Los modelos de regresión son métodos estadísticos que evalúan la relación entre una variable dependiente con una independiente [40]. En el contexto de ECV, se estima la relación entre riesgo cardiovascular y los factores de riesgo cardiovascular-

Existen numerosos modelos de predicción de riesgo cardiovascular en los que usa regresión para entrenar los datos, pues los modelos de regresión predicen un valor continuo. En este contexto, el estudio titulado "*Mejora de la predicción de riesgo cardiovascular mediante modelos de aprendizaje automático de registros médicos electrónicos repetidos de forma irregular*" habla precisamente de un modelo para predecir el riesgo cardiovascular en una población en China, utilizando registros electrónicos de salud. Este estudio propone una comparación entre modelos tradicionales como el modelo CHINA-PAR con modelos de aprendizaje automático, usando modelos de regresión para predecir el riesgo cardiovascular a cinco años. Los resultados tras entrenar el modelo con regresión LASSO fueron significativamente mejor que el modelo CHINA-PAR [41].

2.7.2 Comparación Framingham con machine learning

En general, el estudio de Framingham es uno de los más conocidos y reconocidos a nivel mundial, por eso muchos de los modelos de predicción de riesgo cardiovascular se comparan con este. Sin embargo, este ha tenido deficiencias en el área de pronóstico.

El artículo titulado "*Perspectivas del uso de métodos de aprendizaje automático para predecir enfermedades cardiovasculares*" usó 2236 pacientes para entrenar un modelo de aprendizaje y los resultados mostraron que los indicadores encargados de medir el desempeño de cada modelo fueron significativamente mejores comparados al modelo con Framingham [42].

2.7.3 Modelos de predicción de riesgo cardiovascular y eventos de la vida diaria

Además de la evaluación de riesgo cardiovascular con las variables tradicionales, como los de regresión y comparar modelos con el estudio de Framingham, otros estudios proponen la integración de estos modelos usando eventos de la vida diaria como variables.

Un ejemplo de esto se encuentra en el estudio "*Aplicación de algoritmos conjuntos de aprendizaje automático sobre factores de estilo de vida y dispositivos portátiles para la predicción*

del riesgo cardiovascular", aunque este se fue comparando con la variable de riesgo de Framingham para tener una referencia. Este estudio consideró los datos de 600 personas del sudeste asiático sin enfermedad cardiovascular. Este estudio predijo el riesgo cardiovascular usando un algoritmo de aprendizaje automático, el cual mostró que el uso variado de variables de eventos cotidiano hizo que mejorara la predicción del modelo. Este estudio concluyó que los algoritmos de ML tienen un mejor rendimiento que Framingham cuando se agregan múltiples grupos de riesgo [43].

2.8. Base de datos

2.8.1 Cardiovascular Health Study dataset

Con todo lo anterior, la base de datos Cardiovascular Health Study (CHS) es un estudio que contiene información de enfermedades coronarias e infartos en adultos mayores de 65 años. Su objetivo es evaluar los factores de riesgo cardiovascular haciendo un seguimiento a las personas que participen del estudio para identificar factores relacionados con la aparición y/o evolución de Cardiopatías Coronarias (CC) o infartos. Este estudio fue realizado con hombres y mujeres estadounidenses, los cuales se sometieron a pruebas físicas y de laboratorio para identificar el estado de las personas, si estas tenían una ECV y otra condición que podría aumentar el riesgo cardiovascular [6].

La importancia de los estudios en factores de riesgo en adultos mayores radica en los decesos por ECV en este grupo etario, sumándole que cada año este grupo aumenta rápidamente. Además, la evaluación de ECV puede variar con la edad [6].

Entrando más en detalle, esta base de datos reclutó ia 5,888 hombres y mujeres estadounidenses mayores de 65 años. La selección de participantes fue llevada a cabo a partir de las listas de elegibilidad de Medicare de Health Care Financing Administration (HCFA), en donde se mostró que las personas elegibles eran personas que vivían en el hogar de cada individuo muestreado del marco de muestreo de la HCFA, excluyendo a las personas que estaban en silla de ruedas [6]. Entre 1989-1990 se recluto una cohorte de 5.201 participantes, y entre 1992 y 1993 se agregó una cohorte suplementaria de 687 participantes. Durante diez años el estudio realizó exámenes clínicos

anuales, con contactos telefónicos semestrales intermedios. Los seguimientos telefónicos continuaron durante 7 años aproximadamente.

En los exámenes iniciales se recopilaron los datos demográficos, de salud y de estilo de vida a través de entrevistas, cuestionarios y exámenes clínicos. Además, se hizo una evaluación de la calidad de vida, el apoyo social, la depresión, la actividad física; se tomó la presión arterial, se auscultó el corazón y los pulmones. Dentro de los exámenes físicos se hicieron exámenes de fuerza, de velocidad al caminar, levantarse de una silla. Se prosiguió con una evaluación neurológica y cognitiva, en donde se evaluó el estado mental y se sometieron a pruebas de sustitución de dígitos y símbolos. Se midió la impedancia bioeléctrica como estimación de grasa corporal y se realizaron ECG tanto en reposo como ambulatorios. Luego se midió la capacidad vital forzada (CVF) y el volumen espiratorio forzado en un segundo (FEV1) [6].

Para medir el control de calidad, se utilizó una red de microordenadores, recopilando datos en papel, en donde cada formulario ingresado fue duplicado. También se tomaron muestras de sangre, exámenes ecográficos, tomas de ECG en reposo y ambulatorio, todo esto duplicado para determinar la reproducibilidad, tomando medidas correctivas si se encontraban diferencias significativas.

Para el seguimiento de los participantes del estudio se programaron llamadas telefónicas y visitas clínicas alternas. Las llamadas telefónicas duraban 10 minutos y eran esenciales para identificar hospitalizaciones o eventos importantes que pudieran haber afectado a las personas. CHS también realizó búsquedas periódicas en los archivos de HCFA Medicare, para identificar hospitalizaciones de participantes que pudieron haber sido pasados por alto.

Finalmente, se clasificaron los eventos cardiovasculares destacando las principales ECV según CHS, las cuales son infarto al miocardio, angina de pecho, insuficiencia cardíaca congestiva, arteriopatía periférica, infarto y accidentes isquémicos transitorios. Además, ya que el estudio no excluyó a personas con alguna ECV al momento de la inscripción, se clasificó a los participantes en relación con la presencia o ausencia de una de las seis ECV más frecuentes.

2.8.2 Variables de la vida diaria en CHS

Dentro del informe de CHS no se hace mucha mención a los eventos de la vida diaria, sin embargo, sí se consideran dentro de las variables para el estudio. Tal y como se muestra en la Figura 12, se hizo seguimiento de los eventos de la vida semi-anualmente. Estos eventos incluyen tabaquismo, hábitos alimentarios, estrés, entre otros.

TABLA2. Componentes del examen inicial y repeticiones previstas durante el seguimiento

Componente	Administración Repetida prevista
FACTORES PSICOSOCIALES	
Salud percivida/Calidad de vidad (13)	Anualmente
Apoyo social (14)	
Redes sociales (15)	Anualmente
Eventos de vida (16)	Semi-anualmente
Depresión (17)	Anualmente

Figura 12: Fragmento de tabla con el seguimiento de las mediciones de los componentes del examen inicial [6].

2.9. Discusión

El estudio del riesgo cardiovascular es, sin duda, un campo de estudio muy interesante, no solo por el hecho de estudiar la bomba del cuerpo humano, sino también porque contribuye a entender por qué se producen las enfermedades cardiovasculares como los infartos al miocardio o anginas de pecho. Comprender y lograr reducir los factores de riesgo cardiovascular es sumamente importante para mantener una buena la salud.

Ahora bien, con la información recopilada en el marco teórico sabemos que uno de los factores de riesgo más comunes es la edad, siendo este un factor no modificable. El estilo de vida de cada persona puede impactar positivamente reduciendo el riesgo cardiovascular. Para este informe, la base de datos CHS es un recurso de información para estudiar el comportamiento de las enfermedades cardiovasculares en adultos mayores, ayuda al desarrollo de un modelo de aprendizaje que tome las variables de eventos de la vida diaria para estudiar el riesgo cardiovascular de estas y tratar de mitigar ese factor. Por otro lado, pese a que la base de datos CHS tiene saltos temporales en donde no se registraron datos para ciertas variables, y que el abordaje fue complejo, existen suficientes datos para considerar en periodos de tiempo acotados.

Si bien es cierto, el desarrollo de un modelo de aprendizaje requiere un alto manejo de conocimientos, que van desde el preprocesamiento de la información hasta la evaluación del modelo, vale la pena instruirse y aprender todos los conceptos necesarios para estudiar cómo actividades cotidianas pueden llegar a alterar nuestro corazón, y aún más si es posible desarrollar un modelo capaz de predecir un evento cardiovascular con alguna de estas actividades.

Capítulo 3. Materiales y Métodos

3.1. Introducción

Ahora que ya se describió lo fundamental para poder entender los conceptos de esta memoria, en el siguiente capítulo se describe en detalle la metodología empleada en este proyecto, además de los materiales y técnicas empleadas para el desarrollo y análisis del modelo de aprendizaje.

3.2. Materiales

En el marco de este proyecto se utiliza la base de datos Cardiovascular Health Study disponible en BioLincc[1] como material principal para el desarrollo del modelo [44]. La muestra seleccionada para esta memoria de título corresponde a la cohorte original de 5.201 pacientes y 86 variables que describen eventos de la vida diaria. La información tanto de los pacientes como de las variables se obtuvieron procesando los archivos '*base1final.csv*', '*base2final.csv*' y '*basebothfinal.csv*' que corresponden a los dos primeros años de seguimiento.

El archivo '*base1final.csv*' contiene la información de 5201 pacientes y se describen 642 variables. El archivo '*base2final.csv*' contiene la información de estos 5201 pacientes que en *base1final.csv*, pero se describen 454 variables. Por último, el archivo '*basebothfinal.csv*' contiene la información de los 5201 paciente de '*base1final.csv*' más otros 687 nuevos pacientes, sumando así un total de 5888 pacientes. En '*basebothfinal.csv*' se describen 322 variables. Para efectos del procesamiento se eliminaron los 687 nuevos pacientes para solo trabajar con la cohorte original.

Cabe destacar que, para utilizar este set de datos, se cuenta con permiso de BioLINCC, que son los custodios de los datos, a través de la firma de un acuerdo de distribución de materiales para la investigación, que a su vez posee la aprobación del comité ético científico de la Universidad de Concepción.

[1] https://biolincc.nhlbi.nih.gov/home/

3.3. Metodología

3.3.1 Análisis exploratorio de base de datos

Tras comprender los conceptos clave para desarrollar este proyecto, se examinó en las bases ya mencionadas. Se estudiaron de forma manual variable por variable en el archivo "*Data Dictionary.pdf*", en donde se encuentran todas las variables del estudio y en qué archivo se pueden encontrar. De esta forma se pudieron seleccionar variables de los archivos '*base1final.csv*', '*base2final.csv*' y '*basebothfinal.csv*', con la cuales se pudo formar la base de datos a utilizar en este trabajo de memoria de título.

Primero se seleccionaron las variables de entrada, en donde de '*base1final.csv*' se seleccionaron 36 variables, las que describen si los pacientes son fumadores, su ocupación laboral, las relaciones que tienen con otras personas, educación, si tiene actividades al aire libre, los ingresos, entre otros. De '*base2final.csv*' se seleccionaron 30 variables, de las cuales describen las relaciones que los pacientes tienen con otras personas, los ingreso, el tipo de dieta y si alguna vez ha sido asaltados. Y por último se seleccionaron 20 variables de '*basebothfinal.csv*', las que describen el tipo de dieta de los pacientes, si son fumadores, la educación, si tienen problemas de visión o audición, entre otros.

Luego se seleccionaron 11 variables clínicas necesarias para definir la variable de salida que posteriormente se usan para predecir el riesgo cardiovascular. Estas variables fueron seleccionadas de la base '*basebothfinal.csv*', las que describen eventos cardiovasculares pasados como angina, infarto al miocardio insuficiencia cardiaca, strokes, accidente isquémico transitorio, anormalidad en ECG y arritmias.

Tras tener todas las variables necesarias para el desarrollo de este proyecto, se prosiguió con el procesamiento de datos, generando matrices para variables de entrada y salida. Este procedimiento se hizo a través del programa Anaconda, usando Python 3.11.

3.3.2 Procesamiento de base de datos

Validación de IDs de pacientes. Como se mencionó anteriormente, se seleccionaron los archivos con información de los primeros dos años se seguimiento que constituyen la línea de base del estudio. Dado que el set de datos incorpora una cohorte suplementaria en el segundo año, se tuvo que filtrar los sets de datos eliminando los 687 pacientes de la nueva cohorte. Lo anterior se logró haciendo un estudio de las IDs de los pacientes presentes en los archivos. Se identificaron 687 IDs extras en la base 'basebothfinal.csv' que no estaban en los otros dos archivos, así que estos 687 pacientes fueron eliminados, quedando un total de 5201 pacientes en dicho archivo. Una vez verificado el total de pacientes y que estos fueran iguales en los tres archivos, se siguió con el procesamiento de las clases.

Definición de variable objetivo. Para poder construir las clases del problema, primero se extrajeron en una matriz 11 variables clínicas que dan cuenta de que el paciente sufrió o no un evento cardiovascular. Estas variables son: "ANGBASE", "MIBASE", "CHFBASE", "STRKBASE", "TIABASE", "AFIB", "MAJMIN", "ANAR1A06", "ANAR1B06", "ANAR1C06", y "ANAR306". Las descripciones detalladas de cada variable se encuentran en el Anexo A. La clase de salida se definió como "ha tenido" y "no ha tenido un evento cardiovascular". Así, si el paciente sufrió al menos un evento de los seleccionados, la variable guardaría el valor 1, si no, el 0, dejando la variable objetivo como una variable categórica. Pese a que los valores de las clases son binarios, los unos y ceros representan un Sí y un No respectivamente.

Distribución de variable objetivo. Con la variable objetivo definida, se observó un desbalance de clases importante, encontrando 3653 pacientes que sufrieron un evento cardíaco versus 1548 que no han sufrido ningún evento cardíaco (Figura 13). Luego, para efectos de la clasificación de estos pacientes, este desbalance de clases se abordó utilizando *.sample()* de la librería pandas para eliminar de forma aleatoria 2105 pacientes que sufrieron un evento cardiovascular, con el objeto de balancear las clases, dejando un total de 3096 pacientes. Finalmente se guardó la matriz con el nombre de *'target.csv'*, para usarla más tarde para entrenar el modelo.

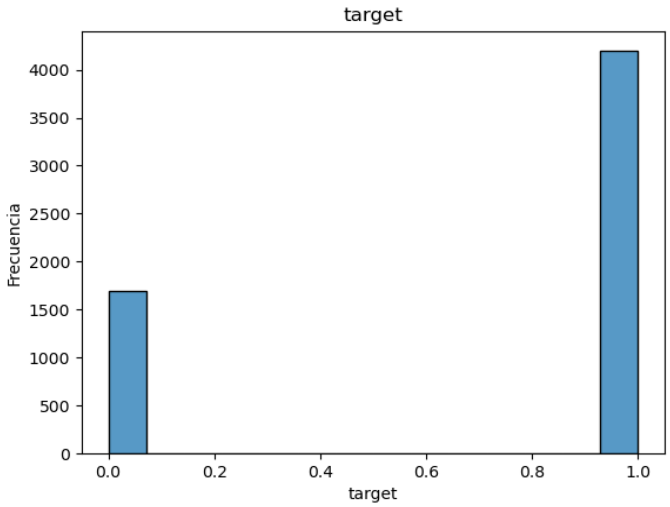


Figura 13: Distribución del target usando variables de eventos cardiovasculares, en donde 1 indica que sufrió un evento cardiovascular, y 0 que no.

Procesamientos variables de entrada. EL procesamiento de las variables de entrada constó principalmente del uso de métodos de imputación de datos, para lo cual se importó la matriz con las 86 variables de entrada. Los valores faltantes de la base de datos se imputaron utilizando la moda, dado que la mayoría de las características para clasificación son variables categóricas.

Selección de características. Para hacer la selección de características se implementó el método de filtro, método de envoltura y tres algoritmos para la selección de características: SelectKBest, featureWiz y SelectFromModel usando el modelo de Random Forest.

Para el método de filtro se detectaron variables duplicadas y variables que describían cosas similares, por ejemplo, hubo 5 variables que describían si la persona era fumadora, ya sea pasiva, fue fumador, etc. Por esto fue necesario hacer una correlación entre las variables de entrada con la clase. Se identificaron las variables categóricas y numéricas, resultando en 75 variables de tipo categóricas y 11 de tipo numéricas. Al ser la clase de tipo categórico, se usaron los métodos de correlación Chi-cuadrado y LDA para encontrar las más relevantes y eliminar las variables repetidas. Así, se pudo formar una nueva matriz con las variables necesarias para construir el modelo de predicción de riesgo cardiovascular.

Para el método de envoltura se utilizó el método de selección hacia delante, implementando el modelo de regresión logística de Scikit-Learn [45], tomando todas las variables de entradas. Primero sin reducción de características y luego con.

Finalmente, se implementaron los algoritmos de selección de características. Primero se implementó SelectKBest, en donde se seleccionaron las mejores 20 características. Luego, se implementó FeatureWiz, el cual seleccionó 38 mejores características. Posteriormente se utilizó el algoritmo de selección de SelectFromModel para seleccionar las mejores variables del modelo Random Forest, en donde se seleccionaron las mejores 20 características. Para los tres algoritmos se utilizó el data set con las 86 variables o características más la clase.

3.3.3 Desarrollo del modelo

Tras el procesamiento de los datos, se creó el modelo. Se utilizaron cinco modelos principales durante el desarrollo de este proyecto: Árbol de Decisión, Random Forest, Naive Bayes y Support Vector Machine (SVM) para método de filtro y LogisticRegression para el método envolvente. Todos ellos de la librería scikit-learn. Además de estos modelos tradicionales, se implementaron redes neuronales para entrenar el modelo.

Los cuatro modelos utilizados para el método de filtro se implementaron se manera similar: se importaron los data sets, del método de filtro, FeatureWiz, SelectKBest, SelectFromModel y el dataset con las 86 variables seleccionadas originalmente. Luego, cada conjunto de datos se dividió en conjuntos de entrenamiento y prueba, dejando el 30% del total como datos de prueba y fijando '`random_state = 0`' para otorgar aleatoriedad. Además, se importó GridSearchCV para seleccionar los mejores parámetros, y así encontrar el mejor puntaje de precisión. Finalmente se evaluaron los modelos con las métricas de precisión ('*accuracy'*) y F1 score, importando la librería *'classification_report'*, también de scikit-learn para tener una visión completa del rendimiento de cada modelo.

Por otro lado, para la implementación del método de envoltura se desarrolló el modelo *'LogisticRegression'* de Scikit-Learn [45]. Para este método se tomaron las 86 variables

seleccionadas, separando el conjunto de datos en entrenamiento y prueba y escalando los datos. Para este modelo, al igual que para los anteriores, se implementó GridSearchCV para seleccionar los mejores parámetros. Primero se implementó sin reducción de características con PCA y luego con reducción de características, para evaluar el rendimiento del modelo.

Finalmente, se implementaron redes neuronales, usando el modelo Sequential de Keras, definiendo la arquitectura de la red neuronal con una capa de entrada de 4 neuronas, una capa oculta de 16 neuronas, cada una de estas capas con activación ReLU y regularización L2, y una capa de salida con una neurona y activación sigmoid. Además, se añadieron dos capas ocultas de Dropout con una tasa de 0,3 después de cada capa oculta con el fin de evitar el sobreajuste. Después, el modelo se compilo con el optimizador RMSProp (Root Mean Square Propagation) y la función de perdida Binary Cross-Entropy con las métricas error absoluto medio y error cuadrático medio y accuracy para la evaluación. Durante el entrenamiento, el conjunto de datos se dividió en entrenamiento y validación utilizando un batch size de 128 y entrenado durante 500 épocas. Por último, se graficaron las métricas de rendimiento y la función de perdida. Además, se obtuvo el puntaje F1 para una evaluación más completa usando Classification_report.

3.3.4 Definición de métricas de evaluación

Para medir el rendimiento de los modelos entrenados se definieron las métricas para evaluar el desempeño de cada uno de los modelos. Estas métricas son accuracy y F1 score. Ambas métricas sirven para entender y medir el funcionamiento del modelo. La métrica accuracy es definida con la siguiente ecuación:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

En donde,

- TP (True Positive) se refiere a las predicciones verdaderas positivas.
- TN (True Negative) se refiere a las predicciones verdaderas negativas.
- FP (False Positive) se refiere a las predicciones falsas positivas.
- FN (False Negative) se refiere a las predicciones falsas negativas.

Mientras que F1 score se define como la media armónica de la precisión y el recall. Esta métrica evalúa la capacidad predictiva de un modelo, midiendo su desempeño por clase. De esta forma se puede llegar a la ecuación de la siguiente manera:

$$F1\ score = \frac{2TP}{2TP+FN+FP}$$, [46]

3.4. Discusión

El desarrollo de un modelo de aprendizaje parece sencillo a simple vista sin embargo existen muchas dificultades. La mayor dificultad está en el procesamiento de los datos, pues el proceso de estudio puede llegar a ser lento si la base de datos es compleja, como lo fue en este caso. Se lidia con más de 5000 pacientes en donde hay que evaluar qué datos pueden servir para entrenar el modelo y cuáles generarían ruido.

Seleccionar las mejores características para entrenar un modelo es otra de las principales dificultades, pues se requieren hartos conocimientos para poder emplear los diferentes métodos y algoritmos. Pese a esto, no es una tarea imposible. Una vez logrado el procesamiento de datos, lo que sigue es encontrar el modelo adecuado y el evaluar cuál da los mejores resultados.

Capítulo 4. Resultados

4.1. Introducción

En este capítulo se mostrarán los resultados obtenidos al haber empleado la metodología descrita en el capítulo 3, en donde se visualizarán los resultados estadísticos tras haber entrenado varios modelos con diferentes métodos y algoritmos de aprendizaje automático. Además, se explicarán los principales resultados.

4.2. Resultados

4.2.1 Desarrollo del modelo

Preprocesamiento. Para el desarrollo el modelo se hizo una selección manual de las variables correspondientes a eventos de la vida diaria. Estas son descritas en el Anexo B. Luego, tras haber identificado las variables necesarias se procedió a hacer un análisis de estas. En la Figura 14 se muestra de manera más sencilla la frecuencia del porcentaje de valores faltantes de las variables seleccionadas, en otras palabras, la cantidad de valores faltantes, en donde aproximadamente 78 variables tienen a los más un 16% de valores faltantes, lo que es un buen indicador. Dado que hay un 8,9% en total de valores faltantes, se realizó una imputación de estos valores faltantes reemplazándolos con la moda.

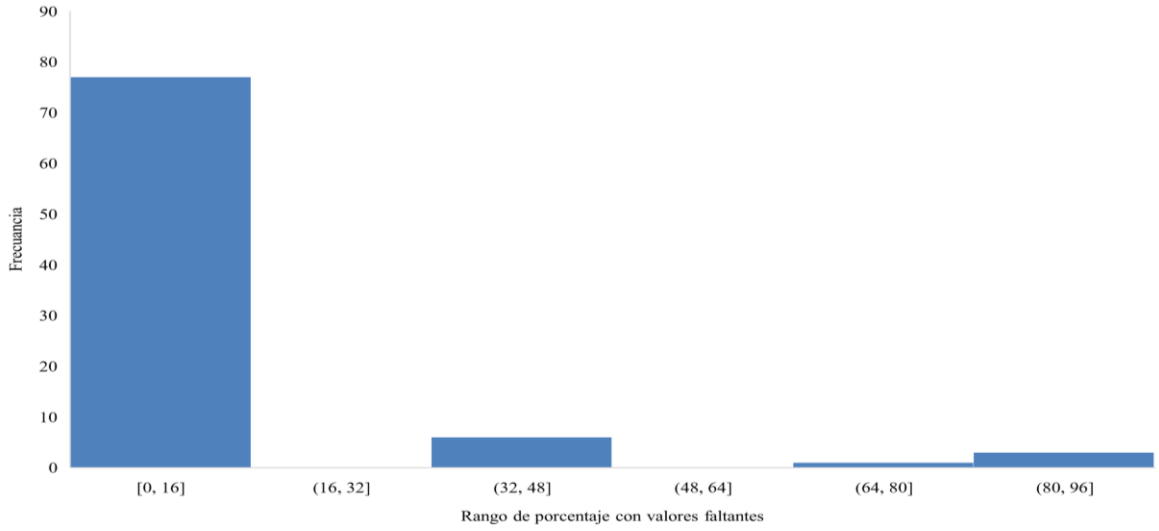


Figura 14: Histograma de la distribución de los rangos de porcentajes de valores faltantes.

Selección de Características: Previo a la clasificación se aplicaron 5 métodos de selección de características: método de filtro, método de envoltura, y 3 algoritmos de selección. Con cada método de obtuvo el mejor set de características, que luego fue llevado a calificación para observar con cuál se obtiene el mejor desempeño.

Método de filtros. Luego, se buscó seleccionar las mejores características para entrenar el modelo, para esto se utilizaron distintos métodos. Primero se utilizó el método de filtro, en donde las 75 variables categóricas fueron sometidas a la prueba chi cuadrado y las otras 11 variables numéricas fueron analizadas con el Análisis de Discriminante Lineal (LDA).

Primero se realizó la prueba de chi cuadrado para evaluar la correlación entre las variables categóricas con la variable objetivo, en donde se estudió el estadístico chi cuadrado y el valor p. El estadístico chi cuadrado indica qué tan relacionada está una variable con la variable objetivo, y el valor p mide la probabilidad de obtener un estadístico chi cuadrado alto con la teoría nula de que exista relación, por lo tanto, un valor p bajo es un buen indicador. Así, se buscó en el Anexo C las variables categóricas que tuvieran un valor p menor a 0,05, seleccionando las 24 variables que se muestran el a Tabla 2.

Tabla 2: Las 24 mejores características seleccionadas con la prueba chi cuadrado.

Variables	*Chi2*	*P - val*	*Variables*	*Chi2*	*P - val*
CHOR04	41,0343	1,50E-10	**EXINTEN2**	17,3554	0,0006
BURGER25	11,3562	0,0448	**FINANC05**	14,5640	0,0001
PIZZA25	20,5815	0,0004	**FNNCIN05**	9,8335	0,0017
MEXICA25	22,3043	0,0002	**GRDN04**	8,9550	0,0028
DIETW25	5,0199	0,0251	**INCOME01**	18,1166	0,0115
SALT25	14,0743	0,0009	**FLUSH06**	6,4350	0,0112
WINE25	13,9813	0,0002	**FALL22**	8,1775	0,0042
LIQUOR25	12,0274	0,0005	**CUROC01**	21,0431	0,0008
EDUC1	11,7615	0,0382	**SMKAMT**	11,3273	0,0101
EDUC2	11,7615	0,0382	**VISPROB**	8,5621	0,0034
DRIVE	8,5898	0,0034	**GEND01**	23,7335	1,11E-06
EXINTEN1	17,2158	0,0006	**AGE2**	76,9152	1,59E-11

Luego se estudiaron las variables numéricas con el análisis de discriminante lineal. En la Figura 15 se muestra un gráfico de barras con la importancia de cada variable numérica según el

modelo LDA, con sus respetivos coeficientes de discriminantes. Así se pudo notar que las principales características predictoras son tres: ADL1, MEALN25 y SNACKN25. Las otras variables al tener sus coeficientes cercanos a cero no aportan mucha información predictora por lo que se decidió dejarlos afuera del entrenamiento. De esa forma, con el método de filtros se seleccionaron 27 variables para entrenar los modelos.

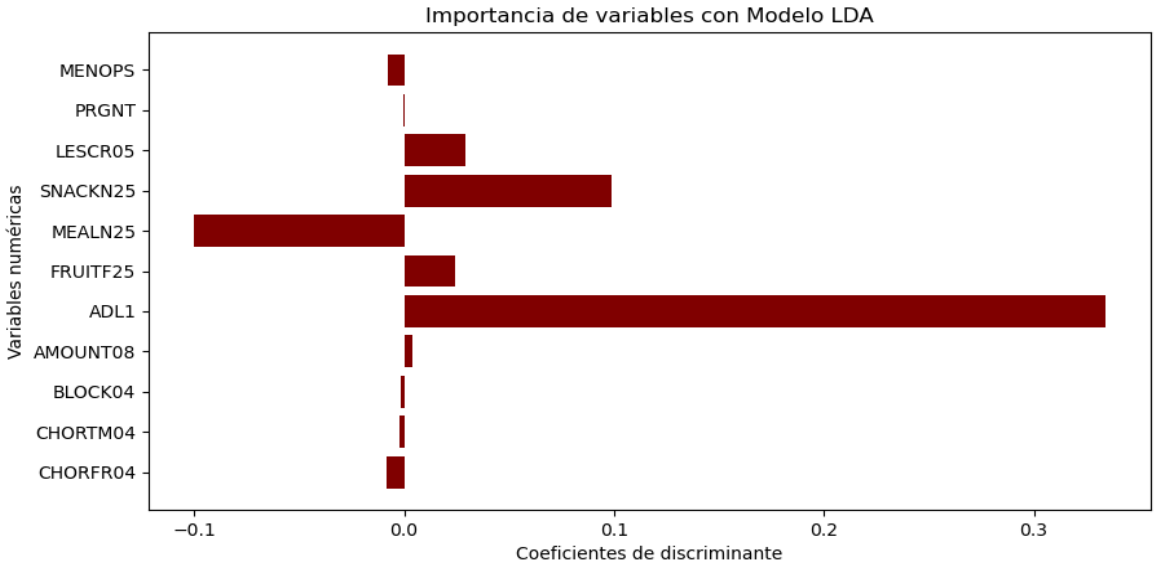


Figura 15: Gráfico de barra de la importancia de las variables numéricas usando modelo LDA.

Método de envoltura. Para el método de envoltura, se usó el modelo de regresión logística pues, como se estudió en el marco teórico, los modelos de regresión suelen tener buenos resultados a la hora de entrenar modelos de predicción. En este caso, se utilizó el conjunto de datos completo, obteniendo la curva AUC (Figura 16). La línea azul delimita el área bajo la curva, mientras que la línea roja delimita el umbral de rendimiento, es decir, sobre de la línea roja se dice que el modelo logra discriminar entre verdaderos positivos y falsos positivos. Así se ve que el rendimiento del modelo tiene un área bajo la curva de 0,57, lo que indica que la capacidad del modelo para discriminar entre falsos positivos y verdaderos positivos bastante baja. Lo esperado era que se acercara más al valor 1, sin embargo, la curva se eleva ligeramente hacia la esquina superior izquierda, indicando que no se entrenó tan bien.

Luego, para mejorar el modelo, se implementó la reducción de características usando un análisis de componentes principales (PCA), en donde para obtener estas componentes principales se graficaron

estas componentes versus la varianza explicada acumulada, estableciendo un umbral de 95%. Así. se obtuvieron 45 componentes principales (Anexo D). Con este método mejoró ligeramente el modelo, obteniendo un AUC de 0.60 (Figura 17).

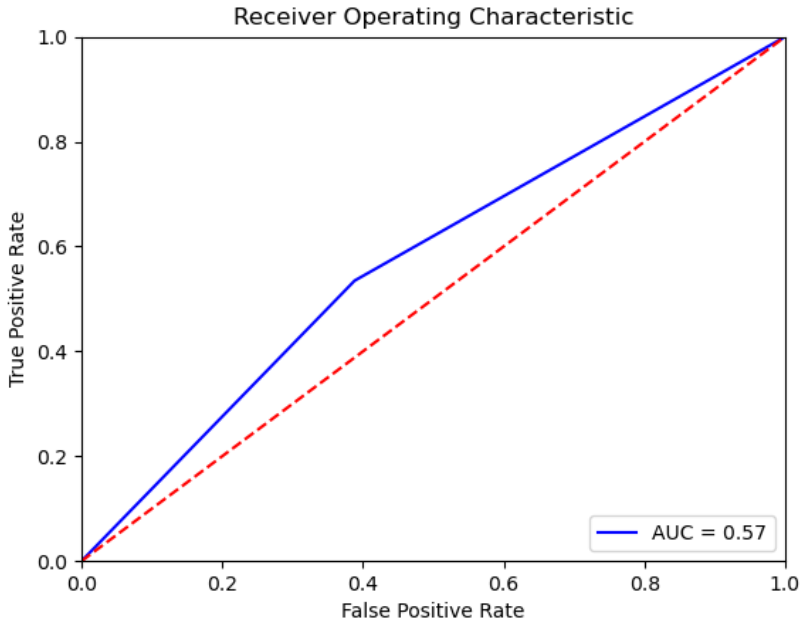


Figura 16: Gráfico del área bajo la curva del rendimiento del modelo LogisticRegression.

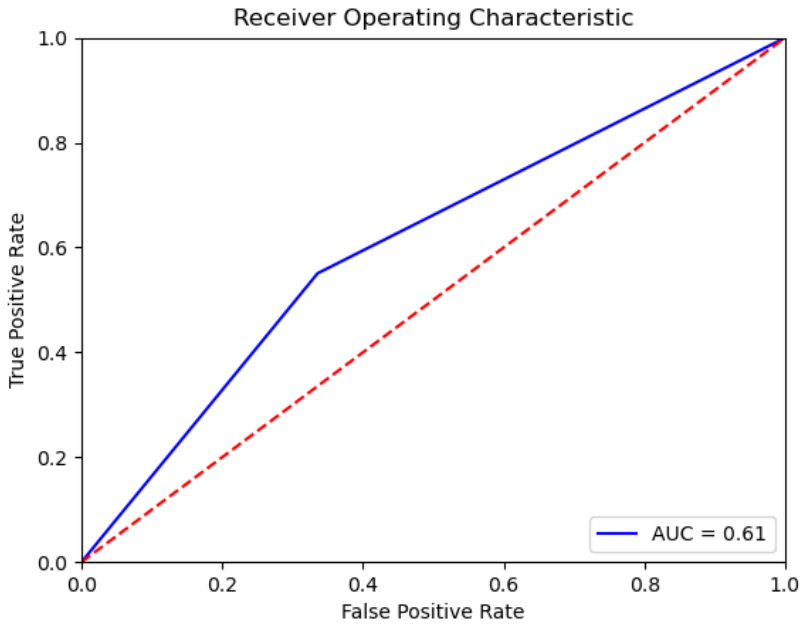


Figura 17: Gráfico del área bajo la curva del rendimiento del modelo LogisticRegression con PCA.

Algoritmos de selección. Se prosiguió a seleccionar las mejores características con algoritmos de selección. El primer algoritmo utilizado fue FeatureWiz que utiliza SULOV y XGBoost. Primeramente, con SULOV, el cual removió 6 variables altamente correlacionadas del conjunto de datos, dejando 80 variables en el conjunto de datos. Luego continuó el modelo recursivo XGBoost. Este modelo dividió el conjunto de datos en subconjuntos para entrenar reiteradamente los subconjuntos. De esta manera seleccionó las mejores 16 características por iteración, y así, el modelo seleccionó las mejores 38 características del conjunto de datos los que fueron utilizados posteriormente en la tarea de clasificación (Anexo E).

Luego se implementó el algoritmo SelectKBest, indicando que se quieren seleccionar las mejores 20 características. El modelo implementó la función *f_classif* para obtener las puntuaciones de las características, además se obtuvieron los valores p, los que, al igual que en la correlación de chi cuadrado, indican la probabilidad de ocurrencia de los mejores puntajes, siguiendo la teoría nula de que no se obtendrán buenos puntajes. Así, se buscó en el Anexo F los valores p menores a 0,05. Una vez identificado las características que un valor p que cumpla con el umbral, se seleccionaron las 20 variables cuyos valores p eran más cercanos a 0 (Tabla 3).

Tabla 3: Las 20 mejores características seleccionadas con SelectKBest.

Variables	*Importancia*	*P - val*	*Variables*	*Importancia*	*P - val*
INCOME01	15,9336	6,71E-05	**SALT25**	12,8670	0,0003
CHOR04	42,0527	1,03E-10	**WINE25**	14,3081	0,0002
CHORTM04	26,8848	2,30E-07	**LIQUOR25**	12,3256	0,0005
GRDN04	9,2168	0,0024	**DRIVE**	9,1234	0,0025
BLOCK04	10,5717	0,0012	**ADL2**	20,9263	4,96E-06
FINANC05	15,1435	0,0001	**EXINTEN2**	12,5283	0,0004
FNNCIN05	10,6048	0,0011	**EDUC2**	10,5372	0,0012
ADL1	20,9263	4,96E-06	**VISPROB**	9,0391	0,0027
EXINTEN1	12,7743	0,0004	**GEND01**	24,2650	8,84E-07
EDUC1	10,5372	0,0012	**AGE2**	64,6453	1,26E-15

Por último, se implementó el algoritmo SelecktFromModel. Este algoritmo de selección requiere de un modelo de aprendizaje automático. Este algoritmo utiliza la importancia de las características según el modelo que se utilice para seleccionar las mejores características. En este caso se utilizó el modelo Random Forest por su buen rendimiento como clasificador. Las características con mayor importancia son las características elegidas por el algoritmo, pues, la importancia de las

características indica la capacidad predictiva que poseen. De esta forma, del Anexo G se obtienen las mejores 20 características seleccionadas por el algoritmo, las que se muestran en la Tabla 4.

Tabla 4: Las 20 mejores características seleccionadas con SelectFromModel

Variables	*importancia*	*Variables*	*importancia*
OCCUP01	0,0210	**PIZZA25**	0,0195
CUROC01	0,0208	**CHINES25**	0,0192
INCOME01	0,0305	**FISH25**	0,0219
CHORFR04	0,0193	**OTFOOD25**	0,0279
CHORTM04	0,0318	**FRUITF25**	0,0230
BLOCK04	0,0474	**LESCR05**	0,0223
AMOUNT08	0,0239	**EDUC2**	0,0215
EDUC1	0,0215	**PRGNT**	0,0208
CHICKN25	0,0220	**AGE2**	0,0440
BURGER25	0,0230	**MENOPS**	0,0304

Redes Neuronales. Por último, se desarrolló una red neuronal con una capa densa de entrada, con 4 neuronas, una función de activación ReLU y un regularizador l2 = 0,001; una capa densa oculta con 16 neuronas, activación ReLU y regularizador l2; y una capa densa de salida con una neurona y activación sigmoid, cuyos resultados serán mostrados más adelante.

4.2.2 Rendimiento de los modelos

A continuación, se presentará en rendimiento de cada modelo entrenado usando todos los métodos antes descritos, cuyo detalle se especifica en el Anexo H. En primer lugar, en la Figura 18.a se muestran los resultados obtenidos entrenando los modelos con el método de filtro, en donde se ve de manera general que los puntajes de exactitud varían ligeramente entre modelos, siendo Random Forest el que obtuvo un mejor resultado con un 60% de exactitud. Por otro lado, comparando los puntajes accuracy con f1, se muestra que no hay gran variación entre ambos puntajes de cada modelo, indicando que los modelos están rindiendo moderadamente bien, sin sesgos.

El rendimiento de los modelos entrenados con las características seleccionadas por el algoritmo SelectKBest (Figura 18.b) se mantuvo similar al entrenamiento con el método de filtros, en donde la disminución más notoria de rendimiento fue con el modelo Naive Bayes, cuyo puntaje f1

descendió a 0,31. Aquí los mejores modelos entrenados fueron SVM y RandomForest con una exactitud del 59%. Luego, en la Figura 18.c se muestra el rendimiento de los modelos entrenados con las características seleccionadas con el algoritmo FeatureWiz. Con este algoritmo los resultados con similares obtenidos con el método de filtros, en donde nuevamente el modelo con mejor rendimiento fue Random Forest. Por otro lado, los modelos entrenados con el algoritmo SelectFromModel (Figura 18.d) fueron los que peor resultados tuvieron pues, pese a que se observa que random Forest tuvo un exactitud de un 60%, su puntaje f1 es considerablemente más bajo, lo que puede indicar que posiblemente esté teniendo problemas identificando los verdaderos positivos.

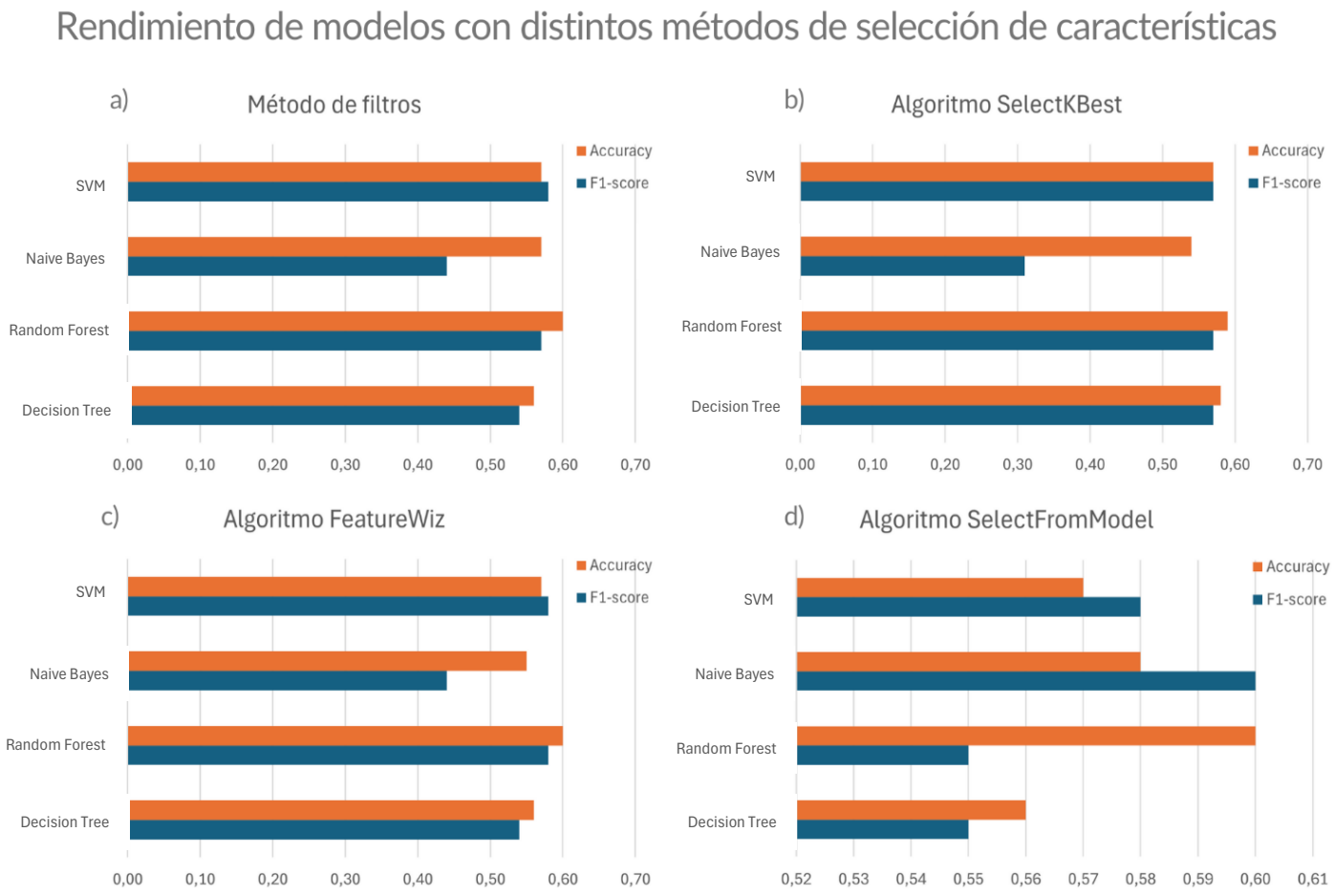


Figura 18: Rendimiento de los modelos Decision Tree, Random Forest, Naive Bayes y SVM entrenados con a) método de filtros, los algoritmos de selección de características b) SelectKBest, c) FeatureWiz y d) SelectFromModel.

Posteriormente, para el método de envoltura se utilizó el modelo regresión logística, el cual fue entrenado primero sin reducción de características y luego con reducción de características. En la Figura 19 se observa que claramente el modelo entrenado con reducción de características mejoró sus predicciones a un 61%, sin tener ninguna clase sesgada.

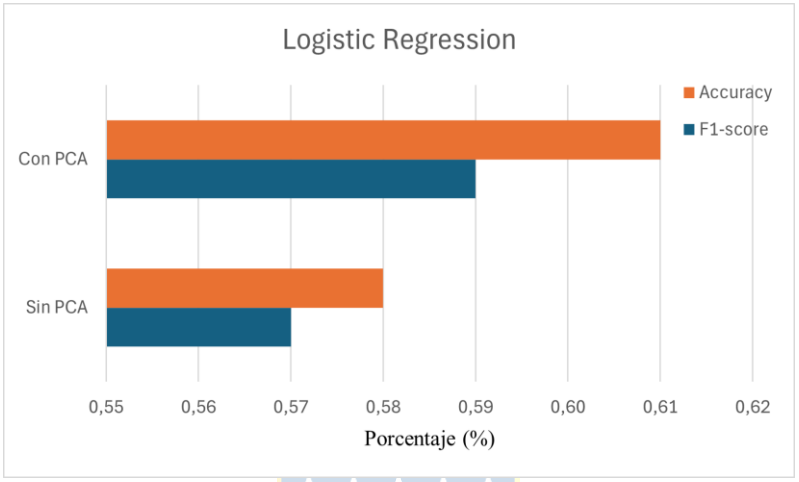


Figura 19: Rendimiento de los modelos entrenados con LogisticRegression

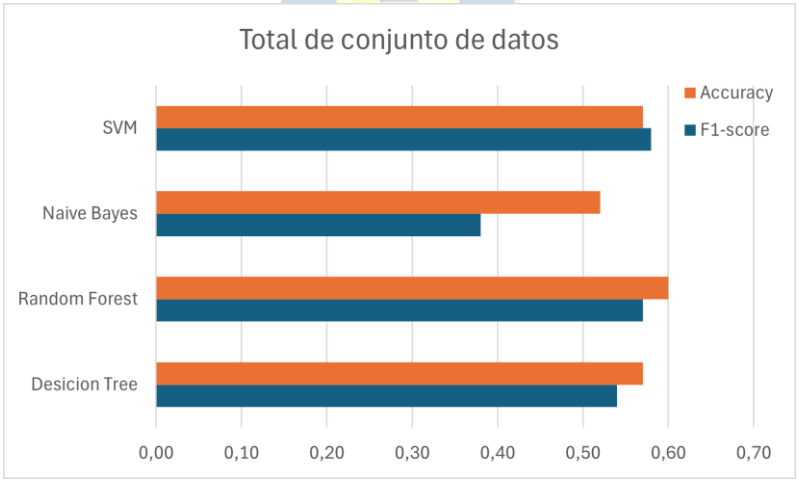


Figura 20: Rendimiento de los modelos entrenados con el total del conjunto de datos.

Ahora bien, pese a que en general los modelos tuvieron un rendimiento moderado, se intentó mejorarlo implementando validación cruzada. Aquí se evaluó el nivel de exactitud de cada modelo, la cantidad de aciertos que tuvieron. Sin embargo, como se muestra en la Figura 21, pese a que hubo modelos cuyo rendimiento mejoró con la implementación de validación cruzada, no fue suficiente

para superar el rendimiento de Random Forest sin la implementación de validación cruzada en los cinco métodos mostrados en la figura 21.

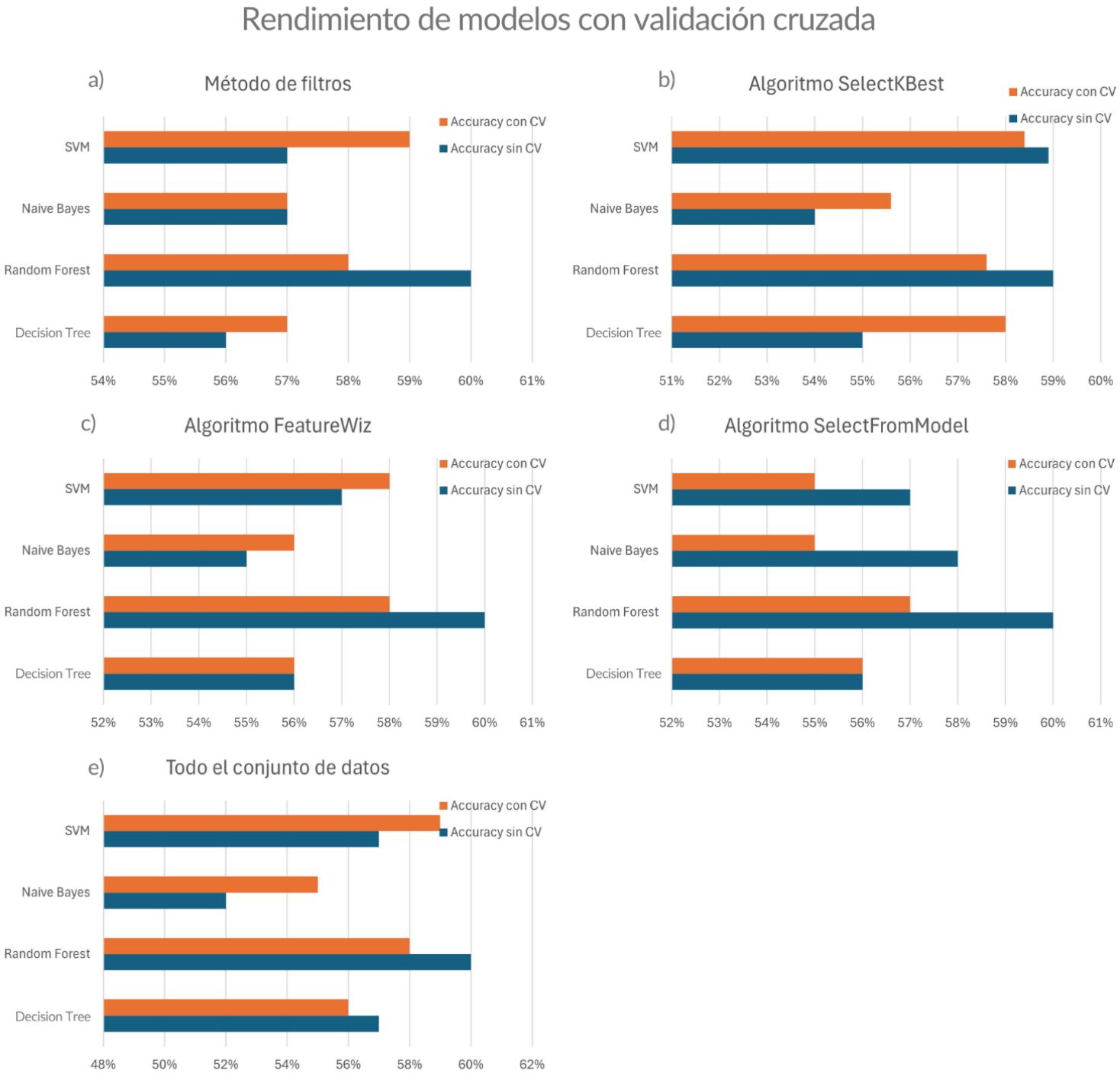


Figura 21: Efecto de la validación cruzada en el rendimiento de los modelos Decision Tree, Random Forest, Naive Bayes y SVM entrenados con a) método de filtros, los algoritmos de selección de características b) SelectKBest, c) FeatureWiz, d) SelectFromModel y e) con el total del conjunto de datos.

En última instancia se buscó obtener mejores resultados predictivos usando redes neuronales. En la Figura 22 se muestra la precisión del modelo de red neuronal en función en las épocas de entrenamiento. La precisión del conjunto de entrenamiento se ve que va en aumento durante las 100 épocas de entrenamiento, alcanzando un máximo en 0,58 aproximadamente. Al mismo tiempo, la precisión del conjunto de validación tiene la misma tendencia creciente y alcanza su máximo en 0,56. Pese a que el hecho de que la curva de entrenamiento sea creciente es un buen indicador, este gráfico sugiere un leve sobreajuste, pues la curva de validación tiene más fluctuaciones que la curva de entrenamiento, en donde lo ideal sería que ambas curvas lograran estabilizarse en un punto y seguir en paralelo.

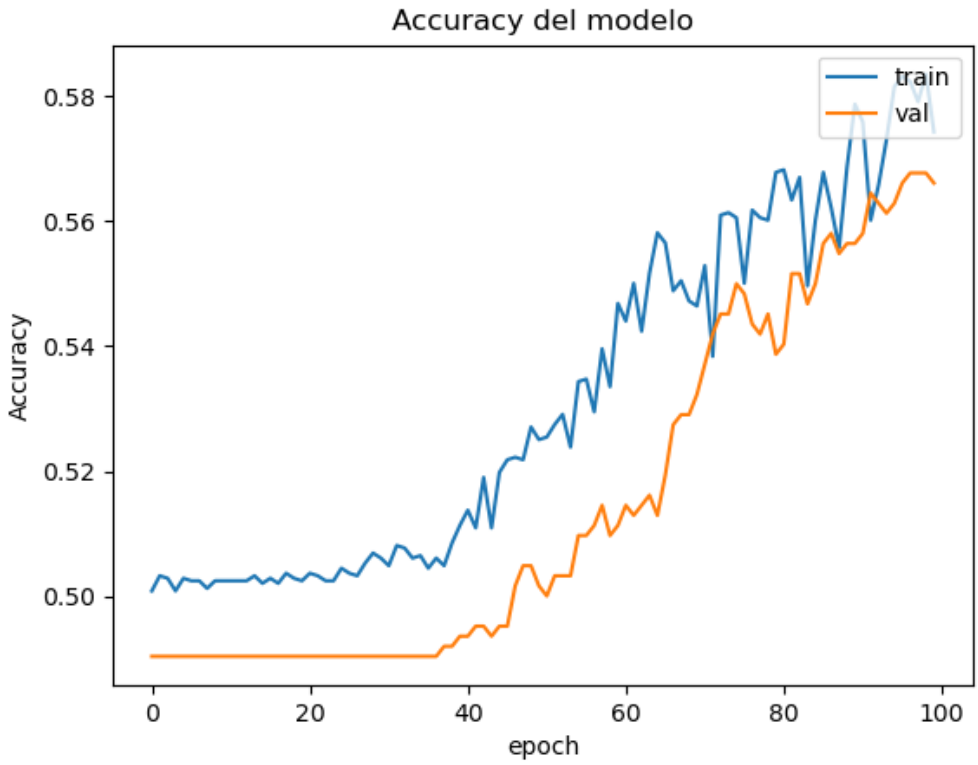


Figura 22: Curca de accuracy del modelo de red neuronal.

En la Figura 23 se muestra la evaluación de la función de pérdida del modelo en 100 épocas de entrenamiento. La curva del conjunto de entrenamiento disminuye constantemente a lo largo de las épocas, llegando cerca de 0,685. De igual forma, tanto la curva de validación, al igual que la curva de entrenamiento, decrece, sin embargo, lo hace de forma menos pronunciada y con menos fluctuaciones,

alcanzando un puntaje de pérdida de 0,688. La disminución inicial de ambas curvas es un buen indicador de que el modelo está mejorando su rendimiento y podría decirse que el modelo está aprendiendo de sus errores.

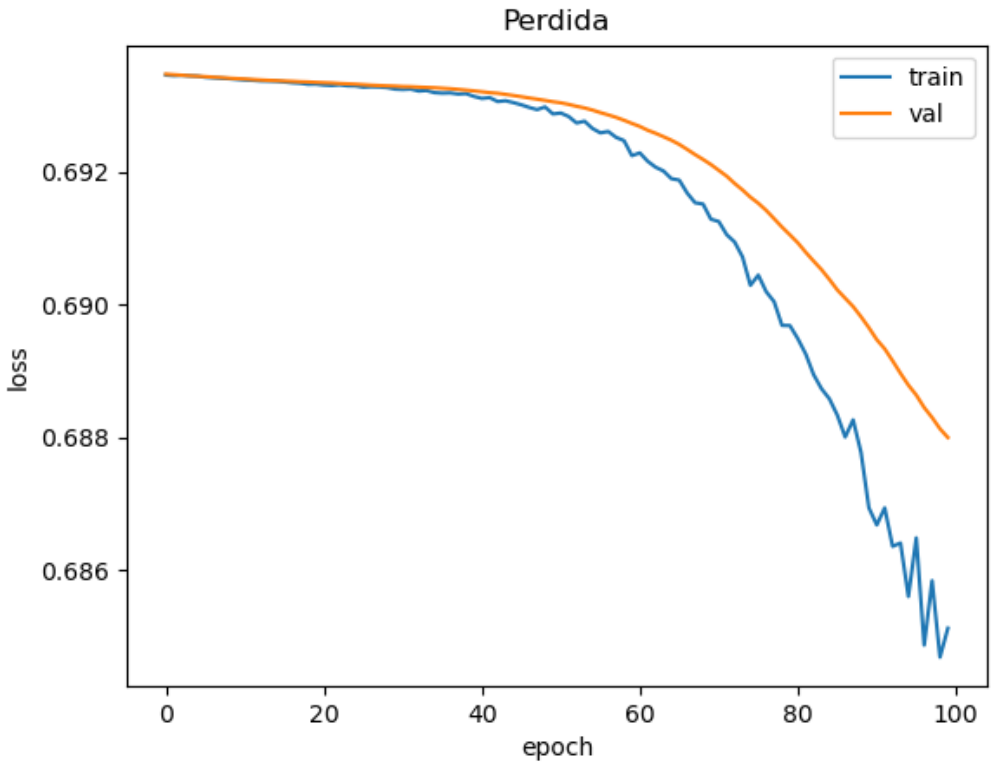


Figura 23: Curva de la función pérdida del modelo de red neuronal.

4.3. Discusión

Durante la selección de características, se observó que cada método utiliza el valor p o el nivel de importancia para la selección de cada característica. No obstante, cada algoritmo selecciona diferentes variables según el método implementado, lo que sugiere que cada algoritmo evalúa el nivel de importancia de las variables de manera distinta. Hubo variables que, pese a tener distintos niveles de importancia, las seleccionaron consistentemente los cuatro métodos implementados. Estas variables fueron BLOCK04, CHORTM04, INCOME01, AGE2 y EDUC1, indicando que son fuertes predictores para este proyecto.

En cuanto a los resultados, considerando los métodos de selección de características, el modelo con mejor desempeño fue Random Forest, obteniendo un accuracy de 60% en 3 de los 4 métodos implementados. Dentro de este contexto, el método de filtro fue el que obtuvo mejores resultados, considerando una relación entre el exactitud con el puntaje f1. No obstante, al evaluar todos los métodos y modelos entrenados, el que tuvo mejor desempeño fue Logistic Regression con reducción de características, con un exactitud de 61% y un puntaje f1 de 0,58. Además, pese a los esfuerzos por mejorar el rendimiento de los modelos utilizando validación cruzada, se identificó una disminución del rendimiento al implementarla.

En cuanto a la red neuronal entrenada la optimización de la arquitectura de esta resultó ser un desafío, pues pese a los ajustes que se implementaron, no se lograron mejores resultados. Se pudo notar que el modelo aprende bien del conjunto de entrenamiento. Por el gráfico de perdida se puede observar que el modelo va aprendiendo de sus errores aprendiendo a generalizar, sin embargo, por el gráfico de exactitud, se pudo notar un leve sobreajuste, aunque no significativo, lo que sugiere que el modelo generaliza razonablemente bien.

Capítulo 5. Conclusiones

5.1. Conclusiones

Durante el estudio del marco teórico fue posible entender la importancia de cuidar uno de los órganos más importantes para la vida. Sin embargo, no basta con mantener una vida de hábitos saludables, ya que pueden existir otros factores que contribuyan al fallo de este órgano crucial. Por ello, se decidió estudiar la mejor forma de prevenir o advertir sobre la posibilidad de contraer alguna enfermedad cardio vascular (ECV). Se consideró una buena estrategia estudiar variables que describieran eventos de la vida diaria para desarrollar un modelo de aprendizaje automático para predecir ECV.

Al haber empezado a estudiar la base de datos CHS, se llegó a la conclusión que la mejor forma de abordar esta problemática era usando modelos de aprendizaje automáticos, además de realizar una selección de características precisa. Para ello, se implementaron distintos métodos de selección de características y se entrenaron varios modelos para evaluar su desempeño y poder escoger el modelo que ofreciera los mejores resultados.

Al comenzar la exploración de datos de la base CHS, se observó que la complejidad de la base de datos era mayor de lo anticipado, debido al volumen considerable de datos y a la necesidad de aplicar conocimientos avanzados en ingeniería de datos. El procesamiento de los datos incluyó la eliminación de IDs de pacientes presentes en uno de los archivos de las bases de datos y ausentes en otros, y el equilibrio de la distribución de la variable objetivo. Con este proceso se logró obtener una base de datos mucho más uniforme y confiable, mejorando la capacidad predictiva de los modelos entrenados, evitando sesgos hacia la clase mayoritaria, en este caso hubiera sido la clase que describe que el paciente “ha tenido” un evento cardiovascular, aumentando la probabilidad de obtener falsos positivos. Para la selección de características, la combinación de distintos métodos, como lo fueron método de filtros, envoltura y algoritmos de selección, proporcionó un análisis mucho más profundo de la capacidad predictiva de las variables. Se identificaron tanto variables altamente predictoras como variables que aportaban poco al entrenamiento del modelo. Así que la implementación de distintos métodos de selección contribuyó a un estudio más profundo del comportamiento de los modelos implementados.

Para la implementación de los modelos finalmente se pudo evaluar qué tan viable era predecir el riesgo cardiovascular usando variables que describieran el diario vivir de pacientes. Primero se evaluaron los modelos de clasificación utilizando los métodos de selección de características, el cual tuvo resultados moderados, en donde el modelo con mejores resultados fue Random Forest empleando el método de filtros, con un exactitud de 60% y un puntaje f1 de 0,57, lo que sugiere que el modelo logra hacer predicciones correctas, pero no es consistente con esas predicciones, por lo que el modelo se presta a mejoras.

Luego se implementó el modelo de regresión logística con y sin reducción de características con PCA, el cual de forma interna hace su selección de características. Con este modelo los resultados mejoraron levemente utilizando PCA, obteniendo un exactitud de 61% y un puntaje f1 de 0,59. Estos resultados sugieren un ligero aumento en la capacidad de identificar casos positivos. Pese a que este modelo obtuvo mejor rendimiento, sigue sin ser lo ideal, dando a lugar buscar mejoras.

Para comparar estos resultados, se entrenaron los modelos con el conjunto de datos, sin seleccionar características, y sus resultados no fueron mejores que el método de filtro. Es por esta razón que se buscó mejorar el rendimiento de estos modelos implementando validación cruzada, sin embargo, pese a que la implementación de esta mejoró el rendimiento de algunos modelos, no logró superar el 60% de exactitud de Random Forest implementado con el método de filtros.

Finalmente, se entrenó una red neuronal para aumentar la precisión en la predicción del riesgo cardiovascular. Esta demostró tener resultados aceptables, en donde se observó un leve sobreajuste en la curva de accuracy por la diferencia entre las fluctuaciones entre las curvas de entrenamiento y de validación. Por otro lado, la curva de pérdida demostró que el modelo, pese a tener un leve sobreajuste, tiende a mejorar, aprendiendo de sus errores. Aunque el modelo demostró que puede mejorar, no se obtuvieron mejores resultados que los modelos descritos.

Teniendo en cuanta el análisis expuesto, se puede afirmar que los resultados obtenidos no fueron los esperados. Aunque se entrenaron varios modelos con distintos métodos de selección de características y utilizando el conjunto de datos, la exactitud de los modelos no superó el 61 %, sin alcanzar el objetivo principal de este proyecto. Esto podría deberse a varias razones. Una de ellas es

que las 86 variables seleccionadas directamente de la base de datos CHS no tenían suficiente capacidad predictiva para predecir el riesgo cardiovascular. Otra posible razón es que los modelos entrenados no lograron captar la complejidad de las interacciones entre las variables seleccionadas. La combinación de estos factores pudo haber afectado a los malos resultados obtenidos.

5.2. Trabajo Futuro

Pese a que ya se han estudiado distintos métodos de selección de características y se entrenaron distintos modelos de aprendizaje automático, es posible volver a aplicar una limpieza de datos para mejorar resultados, implementando además, métodos integrales como la regresión LASSO (método estadístico que ajusta un modelo lineal al reducir la importancia de variables menos relevantes). Además, sería beneficioso aplicar otras técnicas de ingeniería de datos que no se aplicaron en este proyecto, como la creación de nuevas características a partir de otras ya existentes, ya sea combinando características seleccionadas en este proyecto o combinando características que describan eventos de la vida diaria con eventos cardiovasculares. También se puede considerar el uso de otros métodos de correlación como la correlación de Spearman, usada cuando las variables no necesariamente tienen una relación lineal. Además, se puede incluir otros archivos dentro de la base de datos CHS, en donde se puedan encontrar variables que describan otros eventos de la vida diaria no encontrados en los archivos utilizados. Otra opción sería implementar otros algoritmos para predecir el riesgo cardiovascular, como TabNet, que utiliza redes neuronales o AdaBoost, que entrena el modelo usando un conjunto de modelos débiles. Ambos podrían ser usados para resolver la problemática de este proyecto.

Glosario

ECV	: Enfermedad Cardiovascular
PSCV	: Programa de Salud Cardiovascular
CHS	: Cardiovascular Health Study
OMS	: Organización Mundial de la Salud
ACV	: Accidente Cerebrovascular
ANOVA	: Analysis Of Variance
LDA	: Linear Discriminant Analysis
SULOV	: Searching for Uncorrelated List Of Variables
MIS	: Mutual Information Score
PCA	: Principal Component Analysis
ML	: Machine Learning (Aprendizaje Automático)
SVM	: Support Vector Machine
ReLU	: Rectified Linear Unit
CC	: Cardiopatías Coronarias
EEUU	: Estados Unidos
HCFA	: Health Care Financing Administration
ECG	: Electrocardiograma
CVF	: Capacidad Vital Forzada
FEV1	: Forced Expiratory Volume in one second

Referencias

[1] La Vanguardia, «¿Cuáles son las enfermedades cardiovasculares más frecuentes?,» La Vanguardia, 01 10 2023. [En línea]. Available: https://www.lavanguardia.com/vida/salud/20231001/9263994/cuales-son-enfermedades-cardiovasculares-mas-frecuentes.html. [Último acceso: 01 10 2023].

[2] MINSAL, «Indicadores Básicos de Salud 2020,» Depto. Estadísticas e Informción de Salud, 30 03 2021. [En línea]. Available: https://estadisticas.ssosorno.cl/publicaciones/archivos/IBS-Indicadores%20Basicos%20de%20Salud%20SSO%202020.pdf. [Último acceso: 01 10 2023].

[3] MINSAL, «INDICADORES BÁSICOS DE SALUD 2016,» Depto. Estadísticas e Información de Salud, 2017. [En línea]. Available: https://repositoriodeis.minsal.cl/Publicaciones/2018/10/IBS%202016.pdf. [Último acceso: 01 10 2023].

[4] P. Galindez, M. A. Palacios Aguilar, A. J. Flórez Castro y P. Keynis, «Factores de riesgo para enfermedades cardiovasculares identificados en los colaboradores de la alcaldía del municipio de Almaguer departamento del Cauca,» Repositorio Digital ECCI, 2023. [En línea]. Available: https://repositorio.ecci.edu.co/handle/001/3436. [Último acceso: 01 10 2023].

[5] J. P. Navarrete, J. Pinto, R. L. Figueroa, M. E. Lagos, Z. Qing y C. Taramasco, «Supervised Learning Algorithm for Predicting Mortality Risk in Older Adults Using Cardiovascular Health Study Dataset,» MDPI, 14 11 2022. [En línea]. Available: https://www.mdpi.com/2076-3417/12/22/11536. [Último acceso: 2024].

[6] L. P. Fried, N. O. Borhani, P. Enright, C. D. Furberg, J. M. Gardin, R. A. Kronmal, L. H. Kuller, T. A. Manolio, M. B. Mittelmark, , A. Newman, , D. H. O'Leary, B. Psaty, P. Rautaharju, R. P. Tracy y P. G. Weiler, «The cardiovascular health study: Design and rationale,» *Annals of Epidemiology,* vol. 1, nº 3, pp. 263-276, 1991.

[7] OMS, «Constitución,» Organización Mundial de la Salud, 07 04 1948. [En línea]. Available: https://www.who.int/es/about/governance/constitution. [Último acceso: 01 10 2023].

[8] S. Herrero Jaén, «Formalización del concepto de salud a través de la lógica: impacto del lenguaje formal en las ciencias de la salud,» *Ene,* vol. 10, nº 2, 2016.

[9] Bupa, «Sistema cardiovascular,» Bupa, 2020. [En línea]. Available: https://www.bupasalud.com/salud/sistema-cardiovascular. [Último acceso: 03 10 2023].

[10] OMS, «Enfermedades cardiovasculares,» organización Mundial de la Salud, 17 05 0217. [En línea]. Available: https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). [Último acceso: 03 10 2023].

[11] Clínica Alemana|, «Factores de Riesgo Cardiovasculares,» Clínica Alemana, 2023. [En línea]. Available: https://www.clinicaalemana.cl/centro-de-extension/material-educativo/factores-de-riesgo-cardiovascular. [Último acceso: 03 10 2023].

[12] J. M. Mostaza, X. Pintó, P. Armario, L. Masana, J. F. Ascaso y P. Valdivielso, «Estándares SEA 2019 para el control global del riesgo cardiovascularStandards for global cardiovascular risk management arteriosclerosis,» *Clínica e Investigación en Arteriosclerosis,* vol. 31, nº 1, pp. 1-43, 2019.

[13] C. Morales Boada, «Introducción al Feature Selection (o cómo escoger las variables adecuadas),» LinkedIn, 03 07 2018. [En línea]. Available:

https://www.linkedin.com/pulse/introducci%C3%B3n-al-feature-selection-o-c%C3%B3mo-escoger-las-morales-boada/. [Último acceso: 19 04 2024].

[14] T. Ahsan, «Beginner's guide for feature selection,» medium, 08 05 2021. [En línea]. Available: https://towardsdatascience.com/beginners-guide-for-feature-selection-by-a-beginner-cd2158c5c36a. [Último acceso: 19 03 2024].

[15] R. E. Lopez Briega, «Análisis de datos categóricos con Python,» Raul E. Lopez Briega, 29 02 2016. [En línea]. Available: https://relopezbriega.github.io/blog/2016/02/29/analisis-de-datos-categoricos-con-python/. [Último acceso: 06 05 2024].

[16] L. Gonzalez, «Métodos de Selección de Características,» Aprende IA, 04 01 2019. [En línea]. Available: https://aprendeia.com/metodos-de-seleccion-de-caracteristicas-machine-learning/. [Último acceso: 01 07 2024].

[17] J. Amat Rodrigo, «Análisis de varianza (ANOVA) con Python,» Ciencia de Datos, 12 2021. [En línea]. Available: https://cienciadedatos.net/documentos/pystats09-analisis-de-varianza-anova-python. [Último acceso: 19 04 2024].

[18] E. Kavlakoglu, «Implementing linear discriminant analysis (LDA) in Python,» IBM, 18 03 2024. [En línea]. Available: https://developer.ibm.com/tutorials/awb-implementing-linear-discriminant-analysis-python/. [Último acceso: 05 05 2024].

[19] M. Narvaez, «Prueba de chi-cuadrado: ¿Qué es y cómo se realiza?,» QuestionPro, 2024. [En línea]. Available: https://www.questionpro.com/blog/es/prueba-de-chi-cuadrado-de-pearson/. [Último acceso: 05 05 2024].

[20] Geeks for Geeks, «Feature Selection Techniques in Machine Learning,» Geeksforgeeks, 19 03 2024. [En línea]. Available: https://www.geeksforgeeks.org/feature-selection-techniques-in-machine-learning/. [Último acceso: 01 07 2024].

[21] IBM, «Nodo PCA/Factorial,» International Business Machines, 23 02 2023. [En línea]. Available: https://www.ibm.com/docs/es/spss-modeler/18.4.0?topic=models-pcafactor-node. [Último acceso: 05 2024].

[22] R. Seshadri, «Use advanced feature engineering strategies and select the best features from your data set fast with a single line of code,» GitHub, 2020. [En línea]. Available: https://pypi.org/project/featurewiz/0.0.7/. [Último acceso: 20 06 2024].

[23] Scikit-Learn, «Feature selection,» Scikit-Learn, 2024. [En línea]. Available: https://scikit-learn.org/stable/modules/feature_selection.html. [Último acceso: 20 06 2024].

[24] Microsoft, «Modelo de validación cruzada,» Microsoft Learn, 01 06 2023. [En línea]. Available: https://learn.microsoft.com/es-es/azure/machine-learning/component-reference/cross-validate-model?view=azureml-api-2. [Último acceso: 22 06 2024].

[25] Geeks for Geeks, «Feature selection using SelectFromModel and LassoCV in Scikit Learn,» Geeksforgeeks, 06 03 2024. [En línea]. Available: https://www.geeksforgeeks.org/feature-selection-using-selectfrommodel-and-lassocv-in-scikit-learn/. [Último acceso: 05 2024].

[26] Q. Radich, A. Jenks y E. Cowley, «¿Qués es un modelo de aprendizaje automático?,» Microsoft, 18 04 2024. [En línea]. Available: https://learn.microsoft.com/es-es/windows/ai/windows-ml/what-is-a-machine-learning-model. [Último acceso: 03 10 2023].

[27] J. Luna Gonzalez, «Tipos de aprendizaje automático,» Medium, 08 02 2018. [En línea]. Available: https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2. [Último acceso: 04 10 2023].

[28] IBM, «¿Qué es un árbol de decisión?,» International Business Machines, 2023. [En línea]. Available: https://www.ibm.com/es-es/topics/decision-trees. [Último acceso: 14 11 2023].

[29] Scikit-Learn, «Decision Trees,» Scikit-Learn, 2024. [En línea]. Available: https://scikit-learn.org/stable/modules/tree.html. [Último acceso: 05 05 2024].

[30] IBM, «¿Qué es el random forest?,» International Business Machines, 2023. [En línea]. Available: https://www.ibm.com/mx-es/topics/random-forest. [Último acceso: 14 11 2023].

[31] M. Toprak, «Bayes' Theorem,» Medium, 29 08 2020. [En línea]. Available: https://medium.com/@toprak.mhmt/bayes-theorem-89daf9f11769#id_token=eyJhbGciOiJSUzI1NiIsImtpZCI6ImY4MzNlOGE3ZmUzZmU0Yjg3ODk0ODIxOWExNjg0YWZhMzczY2E4NmYiLCJ0eXAiOiJKV1QifQ.eyJpc3MiOiJodHRwczovL2FjY291bnRzLmdvb2dsZS5jb20iLCJhenAiOiIyMTYyOTYwMzU4MzQtazFrNnFl. [Último acceso: 13 11 2023].

[32] Scikit-Learn, «Naive Bayes,» Scikit-Learn, 2007. [En línea]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html. [Último acceso: 14 11 2023].

[33] S. Taylor, «Bayes' Theorem,» CFI, 2023. [En línea]. Available: https://corporatefinanceinstitute.com/resources/data-science/bayes-theorem/. [Último acceso: 13 11 2023].

[34] L. Gonzalez, «Naive Bayes – Teoría,» Aprende IA, 20 09 2019. [En línea]. Available: Este modelo tiene sus ventajas y desventajas. Dentro de las ventajas se destaca su rapidez y lo fácil de predecir las clases, incluso para los multiclases. Además, este modelo funciona mejor que otros modelos de clasificación, como la regresión logística,. [Último acceso: 14 11 2023].

[35] IBM, «Funcionamiento de SVM,» International Business Machines, 17 08 2021. [En línea]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=models-how-svm-works. [Último acceso: 01 07 2024].

[36] Scikit-Learn, «Support Vector Machines,» Scikit-Learn, 2024. [En línea]. Available: https://scikit-learn.org/stable/modules/svm.html. [Último acceso: 01 07 2024].

[37] Scikit-Learn, «Plot classification boundaries with different SVM Kernels,» Scikit-Learn, 2024. [En línea]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py. [Último acceso: 01 07 2024].

[38] M. Stewart, «Intermediate Topics in Neural Networks,» Medium, 19 06 2019. [En línea]. Available: https://towardsdatascience.com/comprehensive-introduction-to-neural-network-architecture-c08c6d8e5d98. [Último acceso: 28 06 2024].

[39] NIH, «Diccionario de cáncer del NCI,» Instituto Nacional del Cáncer, 2023. [En línea]. Available: https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer/def/estudio-retrospectivo. [Último acceso: 27 11 2023].

[40] C. Ortega, «Análisis de regresión: Qué es, tipos y cómo realizarlo,» Question Pro, 2024. [En línea]. Available: https://www.questionpro.com/blog/es/analisis-de-regresion/. [Último acceso: 05 07 2024].

[41] C. Li, X. Liu , P. Shen, Y. Sun, T. Zhou, W. Chen, Q. Chen, H. Lin, X. Tang y P. Gao, «Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records,» *European Heart Journal - Digital Health*, vol. 5, nº 1, pp. 30-40, 2024.

[42] A. V. Gusev, R. E. Novitsky, D. V. Gavrilov, I. N. Korsakov, L. M. Serova y T. Y. Kuznetsova, «Prospects for the using of machine learning methods for predicting cardiovascular disease,» Manager Of Health Care, 04 10 2019. [En línea]. Available: https://www.idmz.ru/journals/information-technologies-for-the-physician/2019/3/prospects-for-the-use-of-machine-learning-methods-for-predicting-cardiovascular-disease. [Último acceso: 29 11 2023].

[43] H. Weiting, W. Tan, L. C. C. Woon, B. Lohendran, E. M. Ong, K. Y. Khung y S. K. Ng, «Application of ensemble machine learning algorithms on lifestyle factors and wearables for cardiovascular risk prediction,» *Scientific Reports*, vol. 12, nº 1033, 2022.

[44] NHLBI, «NHLBI Data Repository,» CHS NHLBI, 2018. [En línea]. Available: https://chs-nhlbi.org/CHS_PublicData. [Último acceso: Septiembre 2023].

[45] Scikit-Learn, «LogisticRegression,» Scikit-Learn, 2024. [En línea]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. [Último acceso: 05 2024].

[46] R. Kundu, «F1 Score in Machine Learning: Intro & Calculation,» V7 Labs, 16 12 2022. [En línea]. Available: https://www.v7labs.com/blog/f1-score-guide. [Último acceso: 06 2024].

Anexo A. Variables que define la variable objetivo

Variable	*Descripción*
ANGBASE	Estado de Angina en el Baseline
MIBASE	Estado de Infarmo al Miocardio en el Baseline
CHFBASE	Estado de insuficiencia cardíaca en el Baseline
STRKBASE	Estado de infartos en el Baseline
TIABASE	Estado de accidentes isquémicos transitorios en el Baseline
AFIB	paciente tuvo fibrilación atrial
MAJMIN	ECG anormal
ANAR1A06	Arritmia, clase 1A
ANAR1B06	Arritmia, clase 1B
ANAR1C06	Arritmia, clase 1C
ANAR306	Arritmia, clase 3

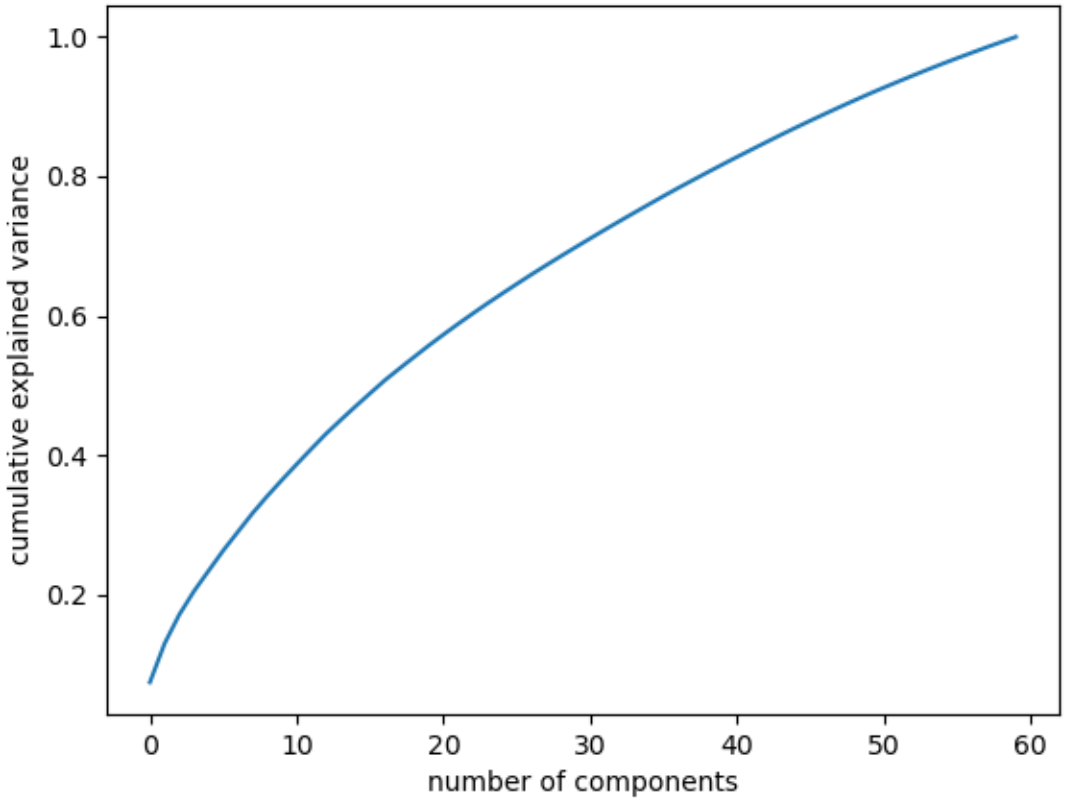
Anexo B. Variables que describen eventos de la vida diría disponibles en la base de datos CHS

1. MARIT01
2. OCCUP01
3. CUROC01
4. INCOME01
5. SMOKE101
6. SMOKE201
7. TALK03
8. WALK04
9. CHOR04
10. CHORFR04
11. CHORTM04
12. MOW04
13. GRDN04
14. HIKE04
15. BIKE04
16. BLOCK04
17. DIE05
18. WORSE05
19. FINANC05
20. FNNCIN05
21. ROB05
22. BORN05
23. SMOKE08
24. AMOUNT08
25. ANYONE08
26. SNORE08
27. CAFFHR11
28. ADL1
29. SMOKE1
30. EXINTEN1
31. SLPILL06
32. FLUSH06
33. EDUC1
34. RETIRE31
35. GRNDCH31
36. CARHRD31
37. CHGFIN31
38. POSFIN31
39. ACCILL31
40. ROBBED31
41. RELWRS31
42. CLSDIE31
43. CHICKN25
44. BURGER25
45. PIZZA25
46. CHINES25
47. MEXICA25
48. FISH25
49. OTFOOD25
50. DIETW25
51. SKIN251
52. FAT252
53. SALT25
54. FATCK1251
55. FATAD1251
56. FRUITF25
57. BRKAST25
58. MEALN25
59. SNACKN25
60. BEER25
61. WINE25
62. LIQUOR25
63. PATTRN25
64. LESCR05
65. ANYONE
66. DRIVE
67. RECOGN
68. CONVER
69. EVERSM
70. SMKAMT
71. ADL2
72. FALL22
73. DIET25
74. SKIN252
75. FAT251
76. FATAD1252
77. FATCK1252
78. EXINTEN2
79. SMOKE2
80. EDUC2
81. PRGNT
82. HEARPROB
83. VISPROB
84. GEND01
85. AGE2
86. MENOPS

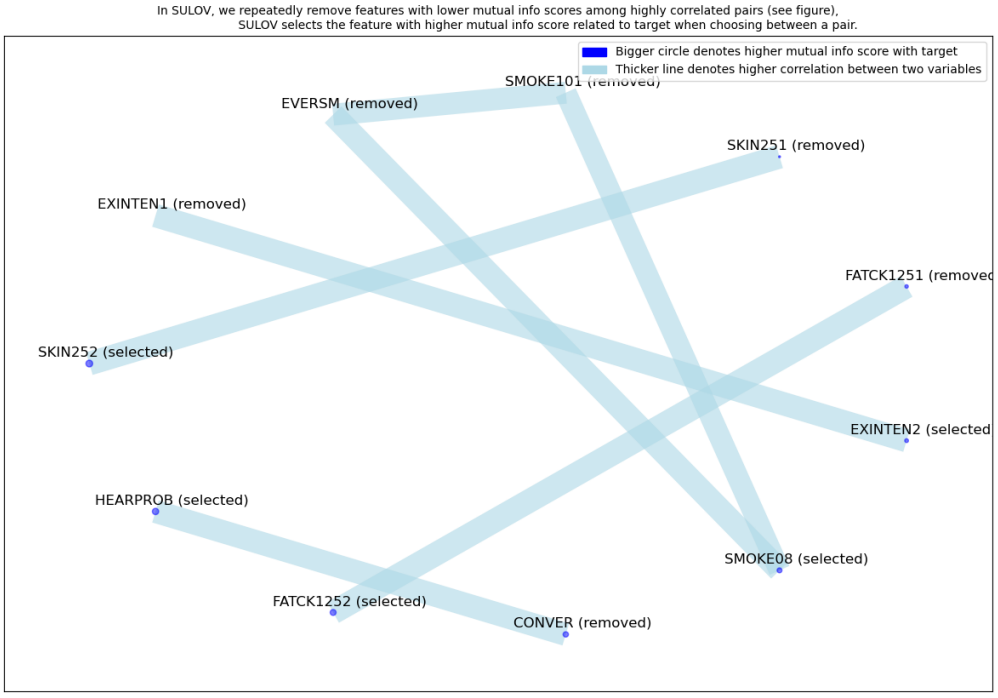
Anexo C. Tabla de correlación de variables categóricas con Chi2

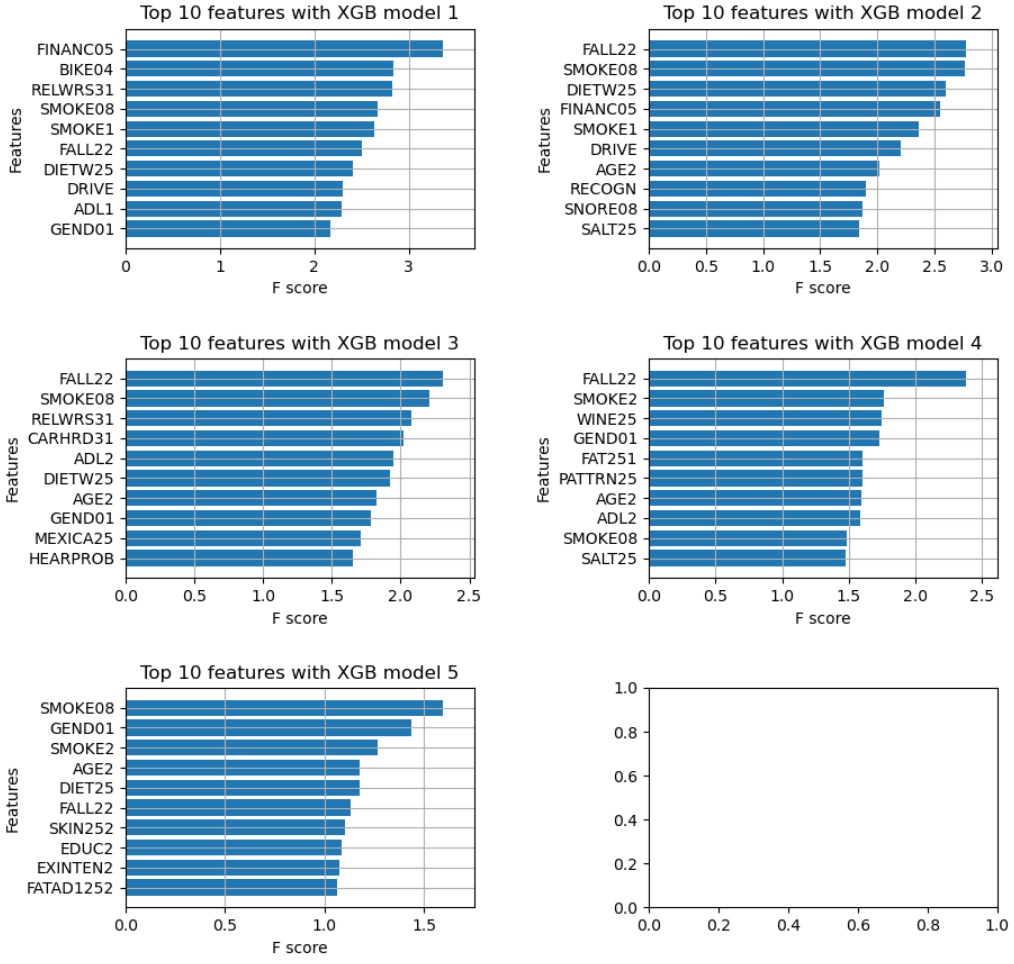
Variables	*Estadístico Chi2*	*P - Val*	*Variables*	*Estadístico Chi2*	*P - Val*
ROB05	1,398172198	0,237029844	**MOW04**	0,092145767	0,761466777
ROBBED31	1,254005401	0,262788844	**GRDN04**	8,95501666	0,002767087
CHOR04	41,03427057	1,49583E-10	**BORN05**	0,542841105	0,46125766
DIE05	0,397664117	0,528298077	**GRNDCH31**	3,688920133	0,054775068
CLSDIE31	0	1	**MARIT01**	5,817062774	0,12085777
CHICKN25	3,174448142	0,673111445	**INCOME01**	18,11656761	0,011454997
BURGER25	11,35623698	0,044756859	**TALK03**	4,656159653	0,19877703
PIZZA25	20,58152021	0,000383268	**HIKE04**	0,411465754	0,521226483
CHINES25	8,090432361	0,08832115	**BIKE04**	1,238008789	0,265855154
MEXICA25	22,30427667	0,000174317	**SNORE08**	0,307041923	0,579501078
FISH25	6,225391763	0,284901056	**CAFFHR11**	1,06367096	0,302379415
OTFOOD25	7,885041131	0,246647083	**SLPILL06**	2,440181149	0,118262568
DIETW25	5,019904548	0,02505755	**FLUSH06**	6,434965928	0,01118953
SKIN251	2,16352902	0,338996834	**RETIRE31**	0	1
FAT252	4,622695307	0,099127572	**CARHRD31**	1,57145395	0,20999599
SALT25	14,07426952	0,00087864	**ACCILL31**	0,00764377	0,930330793
FATCK1251	5,708527489	0,456619881	**FALL22**	8,177451306	0,004241431
FATAD1251	4,554015432	0,472689049	**OCCUP01**	8,269405749	0,082193349
BRKAST25	7,69194435	0,103537202	**CUROC01**	21,04312536	0,000795002
BEER25	2,159447375	0,141695643	**CONVER**	0,775192604	0,378615457
WINE25	13,98130751	0,000184637	**HEARPROB**	2,113778613	0,145978476
LIQUOR25	12,02735543	0,000524254	**SMOKE101**	2,995657186	0,083488032
PATTRN25	0,638910385	0,424105638	**SMOKE201**	1,157127949	0,282061976
DIET25	0,139295236	0,70898265	**SMOKE08**	1,334708049	0,247969375
SKIN252	1,967067221	0,373987236	**ANYONE08**	2,119742539	0,145410986
FAT251	4,622695307	0,099127572	**SMOKE1**	2,984114597	0,224909473
FATAD1252	4,554015432	0,472689049	**ANYONE**	2,119742539	0,145410986
FATCK1252	5,87146425	0,437741846	**EVERSM**	1,019783839	0,312570289
EDUC1	11,76149412	0,038206118	**SMKAMT**	11,32729064	0,010081558
EDUC2	11,76149412	0,038206118	**SMOKE2**	2,984114597	0,224909473
WALK04	0,001324907	0,970963996	**DRIVE**	8,589782914	0,003380543
EXINTEN1	17,21578859	0,000638069	**RECOGN**	0,643424949	0,422473288
EXINTEN2	17,35539779	0,000597216	**VISPROB**	8,562068246	0,003432393
FINANC05	14,56402208	0,000135477	**WORSE05**	0,020708234	0,885576627
FNNCIN05	9,833495992	0,001713623	**RELWRS31**	2,219070779	0,136315082
CHGFIN31	0,152233945	0,696409597	**GEND01**	23,7335402	1,10638E-06
POSFIN31	1,443198951	0,229622381	**AGE2**	76,91521627	1,59395E-11

Anexo D. Curva de componentes principales versus varianza explicada acumulada.



Anexo E. Selección de características con FeatureWiz





Anexo F. Tabla de puntuaciones por el algoritmo SelectKBest

Variables	*P - val*	*Puntaje*	*Variables*	*P - val*	*Puntaje*
MARIT01	0,262033266	1,258441954	**BURGER25**	0,035551319	4,422369462
OCCUP01	0,413207515	0,669733626	**PIZZA25**	0,030419923	4,68971099
CUROC01	0,228573549	1,450289957	**CHINES25**	0,062324923	3,476924192
INCOME01	6,71268E-05	15,93361476	**MEXICA25**	0,053207117	3,740264786
SMOKE101	0,077297432	3,122908692	**FISH25**	0,293159067	1,105432604
SMOKE201	0,257645153	1,281838821	**OTFOOD25**	0,366320777	0,816346823
TALK03	0,326191029	0,964266584	**DIETW25**	0,021721034	5,273333694
WALK04	0,941989801	0,005296213	**SKIN251**	0,570450193	0,321998606
CHOR04	1,03049E-10	42,05269077	**FAT252**	0,033919304	4,502767245
CHORFR04	0,036101514	4,396133082	**SALT25**	0,000339601	12,86702774
CHORTM04	2,29837E-07	26,88482068	**FATCK1251**	0,415618826	0,662848334
MOW04	0,728753959	0,120280583	**FATAD1251**	0,257776325	1,281132297
GRDN04	0,002417963	9,216843808	**FRUITF25**	0,552448974	0,35302534
HIKE04	0,441612487	0,592241269	**BRKAST25**	0,231005551	1,435223871
BIKE04	0,22111144	1,497747622	**MEALN25**	0,099621465	2,71322017
BLOCK04	0,001160625	10,57172349	**SNACKN25**	0,189545794	1,721901584
DIE05	0,503002669	0,448700566	**BEER25**	0,131594198	2,274799854
WORSE05	0,829167829	0,046564127	**WINE25**	0,00015813	14,30812809
FINANC05	0,000101739	15,14354056	**LIQUOR25**	0,000453173	12,32558868
FNNCIN05	0,001140125	10,60478456	**PATTRN25**	0,398986284	0,711576123
ROB05	0,202999239	1,621350502	**LESCR05**	0,6669242	0,185256557
BORN05	0,424931791	0,636803839	**ANYONE**	0,130635767	2,286146648
SMOKE08	0,233632359	1,419164431	**DRIVE**	0,002544186	9,123438365
AMOUNT08	0,007707274	7,109563879	**RECOGN**	0,349522639	0,875454745
ANYONE08	0,130635767	2,286146648	**CONVER**	0,343199155	0,898720145
SNORE08	0,550831944	0,355906858	**EVERSM**	0,2957553	1,093606026
CAFFHR11	0,266852316	1,233301065	**SMKAMT**	0,21615849	1,53031878
ADL1	4,96061E-06	20,9262522	**ADL2**	4,96061E-06	20,9262522
SMOKE1	0,442985655	0,58868748	**FALL22**	0,002972885	8,838060234
EXINTEN1	0,000356788	12,77426999	**DIET25**	0,678485809	0,171867883
SLPILL06	0,102509998	2,66763263	**SKIN252**	0,587569423	0,29421764
FLUSH06	0,010073389	6,630035699	**FAT251**	0,033919304	4,502767245
EDUC1	0,001182408	10,53723621	**FATAD1252**	0,257776325	1,281132297
RETIRE31	0,868689114	0,027336502	**FATCK1252**	0,372706416	0,794851727
GRNDCH31	0,047199232	3,941416948	**EXINTEN2**	0,00040674	12,52829162
CARHRD31	0,168021516	1,901398722	**SMOKE2**	0,442985655	0,58868748
CHGFIN31	0,639766543	0,21909078	**EDUC2**	0,001182408	10,53723621
POSFIN31	0,176646671	1,826446281	**PRGNT**	0,703017137	0,1453791
ACCILL31	0,895703949	0,017187467	**HEARPROB**	0,128601635	2,310539348
ROBBED31	0,213535718	1,547928757	**VISPROB**	0,002663846	9,039137801
RELWRS31	0,110546342	2,547853457	**GEND01**	8,8357E-07	24,26503443
CLSDIE31	0,965406306	0,001881293	**AGE2**	1,26402E-15	64,64525242
CHICKN25	0,221131826	1,497615369	**MENOPS**	0,727002984	0,121903375

Anexo G. Importancia de características por algoritmo SelectFromModel

Variables	*importancia*	*Variables*	*importancia*
MARIT01	0,01362642	**BURGER25**	0,023022466
OCCUP01	0,021048453	**PIZZA25**	0,019477855
CUROC01	0,020800203	**CHINES25**	0,01916986
INCOME01	0,030511361	**MEXICA25**	0,013092841
SMOKE101	0,006384885	**FISH25**	0,021902131
SMOKE201	0,003634363	**OTFOOD25**	0,027932881
TALK03	0,010855819	**DIETW25**	0,006568702
WALK04	0,009417451	**SKIN251**	0,010056609
CHOR04	0,009800543	**FAT252**	0,006391128
CHORFR04	0,01925976	**SALT25**	0,015444778
CHORTM04	0,031776831	**FATCK1251**	0,01771853
MOW04	0,007504302	**FATAD1251**	0,016332264
GRDN04	0,009152035	**FRUITF25**	0,023040283
HIKE04	0,001634419	**BRKAST25**	0,009292471
BIKE04	0,002502798	**MEALN25**	0,011024308
BLOCK04	0,047366174	**SNACKN25**	0,017430051
DIE05	0,008388681	**BEER25**	0,007773574
WORSE05	0,003477462	**WINE25**	0,008028524
FINANC05	0,005466415	**LIQUOR25**	0,008544018
FNNCIN05	0,002131856	**PATTRN25**	0,006983461
ROB05	0,002749237	**LESCR05**	0,022266193
BORN05	0,005274717	**ANYONE**	0,004801529
SMOKE08	0,005541142	**DRIVE**	0,003214574
AMOUNT08	0,023887189	**RECOGN**	0,001388342
ANYONE08	0,004416332	**CONVER**	0,003827512
SNORE08	0,008187306	**EVERSM**	0,004944878
CAFFHR11	0,003513759	**SMKAMT**	0,011784095
ADL1	0,004676284	**ADL2**	0,004588536
SMOKE1	0,008778261	**FALL22**	0,002658771
EXINTEN1	0,014101732	**DIET25**	0,007749792
SLPILL06	0,004076626	**SKIN252**	0,010280337
FLUSH06	0,010066247	**FAT251**	0,006979938
EDUC1	0,021502206	**FATAD1252**	0,017392238
RETIRE31	0,001068969	**FATCK1252**	0,017831089
GRNDCH31	0,004570022	**EXINTEN2**	0,014014201
CARHRD31	0,001662423	**SMOKE2**	0,008129262
CHGFIN31	0,004089408	**EDUC2**	0,021526709
POSFIN31	0,00130729	**PRGNT**	0,020774525
ACCILL31	0,007633372	**HEARPROB**	0,004314687
ROBBED31	0,001978246	**VISPROB**	0,003346315
RELWRS31	0,002574297	**GEND01**	0,008364691
CLSDIE31	0,007801445	**AGE2**	0,043954901
CHICKN25	0,022017113	**MENOPS**	0,030426293

Anexo H. Métricas para evaluar el rendimiento de los modelos obtenidos.

	Modelo	*Precision*	*Recall*	*F1-score*	*Accuracy*
Método de filtro	Desicion Tree	0.56	0.52	0.54	0.56
	Random Forest	0.61	0.54	0.57	0.60
	Naive Bayes	0.64	0.33	0.44	0.57
	SVM	0.57	0.58	0.58	0,57
Método de envoltura	Sin PCA	0.60	0.54	0.57	0.58
	Con PCA	0.63	0.56	0.59	0.61
Algoritmo SelectKBest	Desicion Tree	0.56	0.48	0.52	0.55
	Random Forest	0.61	0.53	0.57	0.59
	Naive Bayes	0.63	0.20	0.31	0.54
	SVM	0.61	0.52	0.56	0.59
Algoritmo FeatureWiz	Desicion Tree	0.57	0.52	0.54	0.56
	Random Forest	0.61	0.55	0.58	0.60
	Naive Bayes	0.58	0.35	0.44	0.55
	SVM	0.57	0.60	0.58	0.57
Algoritmo SelectFromModel	Desicion Tree	0.55	0.56	0.55	0.56
	Random Forest	0.60	0.51	0.55	0.60
	Naive Bayes	0.59	0.60	0.60	0.58
	SVM	0.57	0.60	0.58	0.57
Dataset 86 variables	Desicion Tree	0.57	0.52	0.54	0.57
	Random Forest	0.61	0.53	0.57	0.60
	Naive Bayes	0.56	0.28	0.38	0.52
	SVM	0.57	0.60	0.58	0.57

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TITULO

Departamento : Departamento de Ingeniería Eléctrica
Carrera : Ingeniería Civil Biomédica
Nombre del memorista : **Rocío Hernández Torres**
Título de la memoria : **MODELO DE APRENDIZAJE AUTOMÁTICO USANDO VARIABLES DEL DIARIO VIVIR Y DEL DIARIO VIVIR**

Fecha de la presentación oral : 30 de Agosto, 2024

Profesor(es) Guía : Rosa Figueroa I.
Profesor(es) Revisor(es) : Álvaro Saldaña V. y Miguel Figueroa T.
Concepto :
Calificación :

Resumen

Las enfermedades cardiovasculares (ECV) son una de las principales causas de defunción en el mundo. Solo en Chile representan el 23% de los decesos. Por esta razón que existen diversos programas mundiales y nacionales para prevenir eventos cardiovasculares, tales como Programa de Salud Cardiovascular (PSCV), los promueven un estilo de vida saludable, como llevar una dieta balanceada o ejercitar frecuentemente. Pese a esto, muchas personas cuyos hábitos de vida son considerados saludables, también pueden llegar a sufrir una ECV, lo que ha motivado a muchos investigadores a estudiar las causas de estas enfermedades, buscando métodos para medir el riesgo cardiovascular.

Por esta razón que en este trabajo de memoria de título se busca estudiar la forma de predecir el riesgo cardiovascular usando variables del diario vivir presentes en la base de datos Cardiovascular Health Study (CHS). Para lograr el objetivo propuesto, primero se hizo un análisis exploratorio dentro de la base de datos y un procesamiento de las características seleccionadas, con el fin de identificar las variables más relevantes para entrenar el modelo usando distintos métodos. Posteriormente se entrenaron los modelos con los distintos métodos seleccionados. Para la selección de características, se utilizaron métodos de filtros, de envoltura, y algoritmos específicos como FeatureWiz, SelectKBest y SelectFromModel. Además, se implementaron redes neuronales para mejorar la capacidad de predicción.

Los resultados obtenidos no fueron excelentes, pero tampoco desalentadores. El modelo con el mejor rendimiento fue el modelo entrenado con el método de envoltura, alcanzando un 61% de accuracy y un puntaje f1 de 0,59. Estos resultados, aunque no óptimos, indican que las variables seleccionadas para este proyecto no fueron suficientemente relevantes para predecir el riesgo cardiovascular. Sin embargo, este resultado no descarta la posibilidad de seguir investigando si eventos de la vida diaria pueden afectar el correcto funcionamiento del músculo cardíaco, lo que sugiere la necesidad de aplicar más técnicas de ingeniería de datos.