Predicción de Riesgo Cardiovascular en una Población de Atención Primaria Mediante el uso de Machine learning

Predicting Cardiovascular Risk in a Primary Care Population Using Machine Learning

Fredy Troncoso-Espinosa[1], Juan San Martín-Durán[2]

Fecha de recepción: 30 de mayo, 2024
Fecha de aprobación: 3 de setiembre, 2024

Troncoso-Espinosa, F; Martín-Durán, J.S. Predicción de Riesgo Cardiovascular en una Población de Atención Primaria Mediante el Uso de Machine Learning. *Tecnología en Marcha*. Vol. 38, Nº 1. Enero-Marzo, 2025. Pág. 112-124.

https://doi.org/10.18845/tm.v38i2.7167

1 Departamento de Ingeniería Industrial. Facultad de Ingeniería. Universidad del Bío Bío. Concepción, Chile.
ftroncos@ubiobio.cl
https://orcid.org/0000-0002-9972-3123

2 Departamento de Ingeniería Industrial. Facultad de Ingeniería. Universidad del Bío Bío. Concepción, Chile.
juanmiguel.sd18@gmail.com
https://orcid.org/0009-0001-0529-4234

Palabras clave

Medicina preventiva; enfermedades cardiovasculares; machine learning.

Resumen

Las enfermedades cardiovasculares (CVD) representan un desafío global para la salud, siendo la principal causa de mortalidad a nivel mundial en 2023. En esta investigación, se construyen modelos predictivos para estimar el riesgo de un individuo de desarrollar una CVD. La población bajo estudio incluye a los usuarios del Centro de Salud Familiar Portezuelo (CESFAM) a través del programa crónico y preventivo 2023. Se emplearon cuatro modelos predictivos: Árbol de Decisión, Red Neuronal, Máquina de Soporte Vectorial (SVM) y Naive Bayes. El algoritmo de SVM destacó por su rendimiento, superando el 85% en las métricas evaluadas. Se identificaron atributos de alta importancia, categorizados como factores conductuales y metabólicos modificables, y se determinó un valor de umbral óptimo de 0.45 para distinguir entre pacientes propensos y no propensos a desarrollar una CVD. Estos hallazgos permiten trazar un plan preventivo para reducir la tasa de CVD en la población estudiada. En conclusión, el modelo predictivo demuestra ser una herramienta complementaria eficaz para la toma de decisiones clínicas.

Keywords

Preventive medicine; cardiovascular diseases; machine learning.

Abstract

Cardiovascular diseases (CVDs) represent a global health challenge, being the leading cause of mortality worldwide in 2023. This study constructs predictive models to estimate an individual's risk of developing CVD. The study population comprises users of the Centro de Salud Familiar Portezuelo (CESFAM) through the 2023 chronic and preventive program. Four predictive models were employed: Decision Tree, Neural Network, Support Vector Machine (SVM), and Naive Bayes. The SVM algorithm demonstrated superior performance, achieving over 85% in the evaluated metrics. High-importance attributes were identified, categorized as modifiable behavioral and metabolic factors, with an optimal threshold value of 0.45 to distinguish between patients likely and unlikely to develop CVD. These findings enable the development of a preventive plan to reduce the CVD rate in the study population. In conclusion, the predictive model proves to be an effective complementary tool for clinical decision-making.

Introducción

Las enfermedades cardiovasculares (CVD, por sus siglas en inglés) representan uno de los principales desafíos para la salud en este siglo [1]. Los costos humanos y económicos, referentes a procedimientos quirúrgicos, monitoreo constante, hospitalizaciones, rehabilitación y saturación del sistema son altamente significativos [2]. Por tanto, la estrategia de combate prioritaria debe ser su prevención en lugar de su tratamiento. Bajo lo indicado por la Federación Mundial del Corazón, las CVD se posicionan como la principal causa de mortalidad a nivel mundial, con una estimación de 17.9 millones de vidas cada año [3]. En el país de estudio, Chile no escapa de esta realidad, siendo que el 23.13% de los fallecimientos del año 2022 fueron a causa de una CVD. El problema se agrava debido a la falta de detección temprana producto de una escasa conciencia y conocimiento en la población sobre los perjuicios de esta enfermedad [4], [5].

Uno de los enfoques actuales de prevención corresponde al uso de machine learning. Desde la evaluación clínica hasta el diagnóstico y tratamiento, este enfoque permite fortalecer la precisión diagnóstica; mejorando la interpretación de datos y demostrando su validez como herramienta complementaria a la estadística tradicional [6].

Dentro de esta área, Campo et al. [7] propone un estudio que utiliza diferentes técnicas y enfoques bajo un mismo entorno de prueba, los resultados de su estudio concluyen que Naive Bayes logra un mejor desempeño predictivo. Otro estudio, Bharti et al. [8] aplica diferentes algoritmos a un mismo conjunto de datos con motivo de predecir enfermedades cardiacas: Regresión Logística (LR), KNeighbors (KNN), SVM, Random Forest (RF), Decision Tree (DT), y XGBoost, además incluye un modelo de red neuronal con aprendizaje profundo, la investigación concluye que, el enfoque de aprendizaje profundo con capas densas proporciona resultados más eficaces.

Gupta et al. [9] propone un marco de inteligencia computacional para la predicción de enfermedades cardíacas, que utiliza el análisis factorial de datos mixtos para extraer y derivar características del conjunto de datos de enfermedades cardíacas de Cleveland de UCI. El estudio presenta la implementación de diversos modelos de aprendizaje automático, incluyendo LR, SVM, kNN, DT y RF. Los resultados indican que el modelo RF destaca como el mejor algoritmo.

En definitiva, múltiples investigaciones presentan diferentes enfoques en el uso del aprendizaje automático para la detección oportuna de CVD. Entre los métodos más efectivos se encuentra: Árbol de Decisión (DT), Máquina de Soporte Vectorial (SVM), Redes Neuronales (NA), Naive Bayes, entre otros [10].

Diversos estudios resaltan la importancia del preprocesamiento de datos para mejorar el desempeño predictivo de los algoritmos. Bharti et al. [8] aborda este proceso desde diferentes enfoques, explorando tres perspectivas: el uso de datos sin procesar, la depuración de datos sin valores atípicos y, finalmente, una depuración de outliers y estandarización de datos, concluyendo que, un tratamiento correcto de datos considerando eliminación de outliers, imputación de valores acorde a la distribución de la variable y un exhaustivo análisis exploratorio de los datos conlleva a una mejora el desempeño predictivo de los algoritmos.

Los atributos con mayor relevancia en la predicción de CVD suelen estar clasificados en diversos factores de riesgos; tales como conductuales modificables: fumador, actividad física, consumo de alcohol, tipo de dieta; factores de riesgos metabólicos como: frecuencia cardiaca, fluctuaciones lipídicas, niveles de colesterol, presion arterial y glicemia; e incluso factores sociales como: la clase socioeconómica del individuo y el nivel educacional [11].

Múltiples investigaciones coinciden en que la eficacia en el desarrollo y aplicación de los modelos predictivos mejoraría si es que hubiese un mejor nexo en el trabajo entre equipo de salud y desarrollador del modelo. Sin embargo, la aceptación de estos métodos aún se ve obstaculizada por la falta de algoritmos adecuados, la falta de formación médica en algoritmos predictivos, la preocupación por la sobre mecanización y el temor a perder el" toque humano" [6], [12].

El presente estudio busca diseñar un modelo de predicción de riesgo de enfermedades cardiovasculares (CVD) que detecte de manera oportuna si un individuo sufrirá una enfermedad cardiovascular o no. El estudio está aplicado bajo el registro de controles de salud 2023 de pacientes del Centro de Salud Familiar (CESFAM) Portezuelo, registro que almacena medidas antropométricas, perfil bioquímico y estilo de vida de los pacientes. Con ello se busca determinar y clasificar los atributos de los pacientes asociados con una mayor probabilidad de sufrir una CVD.

Metodología

El presente estudio consistió en la estimación de la predicción de CVD para pacientes del programa crónico y prevención cardiovascular año 2023 del Centro de Salud Familiar, Portezuelo, Región de Ñuble, Chile. Dicho conjunto de datos alberga alrededor de 3200 registros de información de los pacientes referentes a sus medidas antropométricas, perfil bioquímico, estilo de vida e historial clínico referente a enfermedades relacionadas a CVD.

Con motivo de preprocesar el conjunto de datos se realizó el proceso de Descubrimiento de Conocimiento en Bases de Datos (KDD), método que consta de 5 partes iterativas y cíclicas [13].

En la primera fase de selección, se extrae el conjunto de datos total o una muestra que sea representativa del estudio, se recopilan las variables de interés desde diversas fuentes buscando obtener la mayor cantidad de atributos que sean de relevancia.

En la segunda fase de preprocesamiento de datos, se elimina todo tipo de ruido que pudiese empeorar el desempeño predictivo. La limpieza y preprocesamiento de datos se realiza a través de técnicas de estadísticas como: eliminación de datos faltantes, outliers, normalización de datos, entre otras.

En la tercera fase de transformación de datos, se aplican técnicas como la ingeniería de características y la codificación de variables categóricas para preparar los datos para el entrenamiento de los modelos de machine learning seleccionados. Estas transformaciones son esenciales para garantizar que los modelos puedan aprender de manera efectiva.

En la cuarta fase de minería de datos, se implementan los modelos de machine learning, estos algoritmos son seleccionados en conformidad con su aplicación relevante en el campo de la medicina preventiva.

Los modelos predictivos utilizados en el presente son:

- Red Neuronal: Modelo basado en la estructura y comportamiento de las neuronas del cerebro y la comunicación que emiten entre ellas a través de impulsos nerviosos, este sistema es expresado de manera teórica como capas y nodos interconectados que aprenden patrones complejos mediante repeticiones de entrenamiento.
- Máquinas de Soporte Vectorial (SVM): Modelo de aprendizaje supervisado que encuentra el mejor hiperplano para separar distintas clases en un espacio multidimensional, utiliza vectores de soporte a fin de maximizar el margen entre ellas.
- Naive Bayes: Modelo probabilístico basado en el teorema de Bayes, realiza el cálculo de probabilidades asumiendo independencia entre los atributos, de esta manera predice la clase de un objeto en particular.
- Árbol de Decisión: Modelo de aprendizaje supervisado que divide el conjunto de datos en nodos basándose en el peso predictivo de sus atributos, en base a ello crea un árbol de decisión. Cada nodo representa una decisión basada en una característica y los nodos, corresponden a las ramas de un nodo superior.

En la quinta fase de interpretación, se evalúa la calidad de los modelos implementados, luego de ello se extrae análisis de valor en base a los objetivos planteados. Dicho análisis es realizado en búsqueda de mejorar la toma decisiones o en su defecto, proveer de retroalimentación al estudio.

Para la evaluación de la calidad de los modelos se utilizan cuatro medidas de desempeño, cada métrica tendrá un mayor peso que otra dependiendo de los objetivos planteados, dichas medidas corresponden a: Accuracy (Exactitud), Precisión, Recall (Exhaustividad) y F1- Score. Las métricas de desempeño son obtenidas a partir de la matriz de confusión (Véase Figura 1).

		Valor Real	
		1	0
Valor Predicho	1	Verdaderos Positivos (VP)	Falsos Positivos (FP)
	0	Falsos Negativos (FN)	Verdaderos Negativos (VN)

Figura 1: Descripción de la matriz de confusión.

La matriz de confusión es una herramienta que resume el rendimiento de un modelo predictivo, clasificado en cuatro segmentos: Verdaderos Positivos (VP), que corresponde aquellas predicciones correctas de las clases positivas; Verdaderos Negativos (VN), que corresponde a aquellas predicciones correctas de la clase negativa; Falsos Positivos (FP), que corresponde a aquellas predicciones incorrectas de la clase positiva; y por ultimo los Falsos Negativos (FN) que corresponde a aquellas predicciones incorrectas de la clase negativa. La composición de esa matriz permite evaluar la precisión y el rendimiento general del modelo en términos de clasificación.

Las medidas de desempeño a utilizar se describen tal que:

$$Accuraccy = \frac{TP+TN}{Total\ de\ observaciones} \quad (1)$$

Accuracy: Es el porcentaje de veces que el modelo predice correctamente.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Precision: Es el porcentaje de veces que se predice correctamente en relación con el porcentaje de veces que se predijo positivamente.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

Recall: Es el porcentaje de veces que se predice de manera correcta en relación con el total de veces en el que el resultado real fue positivo.

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

F1-Score: Medida que combina la precisión y Recall calculando una media armónica de ambas.

Materiales y Métodos

Diseño del Estudio

El estudio corresponde a un tipo observacional retrospectivo cuantitativo, en el que, mediante información previa permite anticipar una CVD en los pacientes en estudio. La población de estudio comprende el control anual de pacientes de atención primaria año 2023 en Cesfam de Portezuelo, dichos pacientes son asistidos en los programas de prevención y tratamiento crónico cardiovascular. En relación con la recopilación de datos, dicha información comprende desde datos de controles médicos protocolares, resultados de pruebas diagnósticas y encuestas de prevención relacionadas al estilo de vida conductual de los pacientes, además de antecedentes familiares relacionados al CVD. La variable predictora, está calculada en base al índice de Framingham, y con ello, se estima el riesgo a desarrollar una CVD [14].

Procedimientos

Se detallan los procedimientos aplicados en cada una de las fases del enfoque KDD para la presente investigación.

a. Selección de Variables: El conjunto de datos es la unificación del conjunto de base de datos de los programas crónico cardiovascular y de prevención del control anual de pacientes año 2023. En el Cuadro 1, se puede visualizar tanto el atributo como una breve descripción del mismo.

Cuadro 1. Variables consideradas para estimar el riesgo a sufrir una CVD

Variables	Descripción
Edad	Edad en años del paciente.
Sexo	Condición biológica del paciente (hombre, mujer, no informa)
Fecha control médico	Fecha del control médico del paciente.
Peso	Medida en kilogramos del paciente.
Talla	Estatura de la paciente dada en centímetros.
IMC	Índice que evalúa la relación peso-altura en la salud.
Estado Nutricional	Categoría Nutricional del Paciente (Bajo peso, Peso normal, Sobrepeso, Obesidad.)
Presión Arterial	Fuerza de la sangre contra las paredes arteriales, categorizada en presión diastólica y sistólica.
Glicemia	Nivel de glucosa en la sangre.
Creatinina	Producto de desecho muscular filtrado por los riñones.
Colesterol (Total, HDL y LDL)	Niveles de lipoproteínas en la sangre.
Triglicéridos	Grasas en sangre, almacenadas en tejido adiposo o utilizadas como fuente de energía.
Sodio	Elemento que regula balance hídrico y presión sanguínea.
Potasio	Elemento que controla actividad muscular y ritmo cardiaco.
Cloro	Elemento que mantiene equilibrio de fluidos y PH.
Realiza actividad física	Si realiza actividad física moderada de manera frecuente (2 veces por semana).
Adherencia a Dieta	Si el paciente cuida su alimentación mediante nutricionista
Fumador	Si consume cigarrillos de manera frecuente.

Variables	Descripción
Factores de Riesgo CVD	Factores de riesgo metabólicos para sufrir una CVD (Hipertenso, Diabetes Mellitus Tipo II, Dislipidémico)
Antecedentes cardiovasculares	IAM, AVE, Hipertensión Arterial Refractaria y Aterosteclosis
Riesgo Cardiovascular	Riesgo de sufrir una enfermedad cardiovascular calculado en base al Indice de Framingham.

b. **Preprocesamiento de Datos:** Se llevaron a cabo un conjunto de técnicas de limpieza de datos a fin de eliminar todo tipo de errores ocasionados por no asistencia de pacientes a sus controles, errores del equipo médico al ingresar los datos, entre otros. Además, se eliminaron aquellos atributos que tuviesen una cantidad elevada de registros nulos (mayor al 40%). A su vez, se eliminaron los datos duplicados y se corrigieron aquellos errores de formato en los datos. Con motivo de eliminar los sesgos de los atributos y robustecer el modelo se aplicó una eliminación de outliers mediante el método de caja de bigotes. Finalmente, aquellos atributos con valores nulos que estuviesen en el rango esperado se les aplicó una imputación de valores mediante método K-Nearest Neighbor (KNN), de esta forma se evitó la perdida de información valiosa.

c. **Transformación de Datos:** En esta etapa se identificaron y se extrajeron características de interés con las que se pudiese obtener un mejor desempeño predictivo, en primer lugar, se aplicó una matriz de correlación para evaluar para evaluar la relación entre variables numéricas y tabla de chi-cuadrado para evaluar la significancia entre variables categóricas respecto a la variable predictora. Se crearon nuevas variables a partir del conjunto original, tales como: R.SP (Ratio entre sodio y potasio), medida que tiene cierta relación con la probabilidad de sufrir CVD; índice de aterogenicidad (ratio entre colesterol LDL y HDL); cálculo del IfGE. (Índice de filtración glomerular estimado) y variables de acumulación de número de factores de riesgo de CVD y de enfermedades anteriores CVD en el paciente. Posterior a ello, se realizó la estandarización de los datos mediante método MinMaxScaler, de esta forma se asegura la homogeneidad de los datos y una disminución en la probabilidad en el sobreajuste del modelo. Las variables categóricas fueron convertidas a variables numéricas mediante el método de creación de variables dummy. Este proceso fortaleció la calidad de los datos.

d. **Minería de Datos:** Se llevó a cabo la implementación de 4 modelos predictivos: Árbol de Decisión, Red Neuronal, Support Vector Machine y Naive Bayes. Los algoritmos implementados se llevaron a cabo a través del lenguaje de programación Python. El proceso de entrenamiento, validación y prueba se puede observar mediante figura 2, la subdivisión de la metodología consistió en una división trimestral: los primeros dos trimestres (enero a junio) de 2023 fueron para el conjunto de entrenamiento de los modelos predictivos, el posterior trimestre de 2023, fue destinado para la validación y ajuste de parámetros, en el que, se utilizaron k=5 pliegues para realizar un proceso de validación cruzada y un ajuste de parámetros mediante el método Gridsearch. Una vez teniendo los mejores parámetros para el modelo, se utiliza el último trimestre del año para realizar las predicciones y estimar a los pacientes propensos a sufrir una CVD.

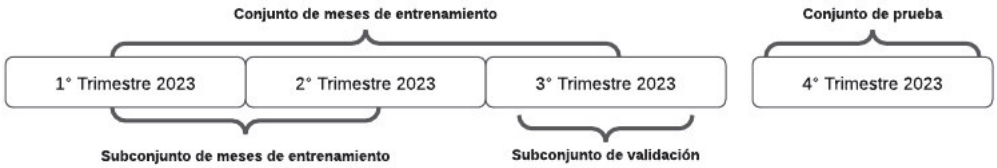

Figura 2: Conjunto de entrenamiento y validación.

e. Interpretación de Resultados: Se determinó el mejor modelo mediante de las métricas de desempeño Recall y F1-Score, se consideran estas métricas como prioridad dada la naturaleza del estudio, en la que cometer un error Tipo II es tanto económica y humanitariamente más costoso que cometer un error Tipo I. Posterior a ello, se determinó el umbral óptimo del mejor modelo, lo cual permitió minimizar la combinación de costos asociados a falsos positivos y negativos, revelando así un equilibrio óptimo entre sensibilidad y especificidad para la toma de decisión, donde se prioriza la identificación precisa de pacientes propensos a padecer una CVD.

Resultados

Decisión del mejor Modelo

El cuadro 2 visualiza las métricas de desempeño de los 4 modelos: RN, NB, SVM y TD. SVM es el modelo que mejor rendimiento presenta acorde a los objetivos planteados teniendo los valores de Recall y F1-Score más altos, lo que indica que posee un mejor equilibrio entre la capacidad de prever correctamente aquellos pacientes propensos a sufrir una CVD (precisión) y la capacidad de identificar correctamente todos los casos positivos (recall). Además, se observa que los modelos predictivos poseen un porcentaje de predicción aceptable en virtud de estimar correctamente si un individuo sufrirá una CVD o no, siendo Naive Bayes el único modelo que presenta un porcentaje de exactitud y F1-Score considerablemente menor al resto, siendo inferior al 70%.

Cuadro 2. Métricas de desempeño

Modelo	Exactitud	Precisión	Recall	F1-Score
Red Neuronal	0,86	0,86	0,86	0,86
Naive Bayes	0,67	0,76	0,67	0,67
Support Vector Machine	0,85	0,90	0,86	0,87
Árbol de Decisión	0,79	0,79	0,79	0,79

Por medio de figura 3 se visualizan las Curvas ROC de los 4 Modelos. Si bien SVM es el que posee un mayor valor, los demás modelos predictivos presentan valores superiores a 0.8, por lo que se consideran aceptables y significativos para el estudio, lo que destaca la capacidad de los algoritmos para discriminar entre las clases en estudio (pacientes que sufrirán un infarto respecto a los que no).

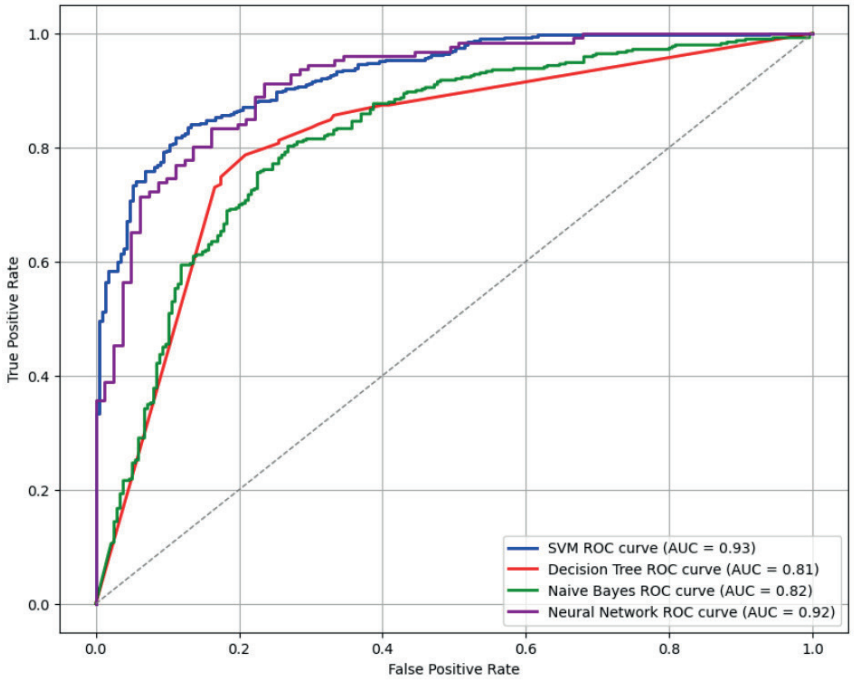

Figura 3: Curva ROC de modelos predictivos

Dada las características del estudio en el que se prioriza el Recall y F1-Score como medidas de desempeño decisivas es que se determina que SVM es el modelo es el más apto para el estudio, es por ello que el análisis posterior está dado en base a la extracción de análisis sobre el mismo.

Importancia de los atributos

La implementación del modelo SVM permite extraer los pesos de las variables (Véase Cuadro 3, de esta forma es posible entender la importancia de cada atributo en la predicción de cada clase (sufrirá una CVD o no). Estos atributos con mayor relevancia dentro del estudio pueden ser clasificados como variables metabólicas modificables, conductuales y antecedentes de enfermedades previas de un paciente.

Los atributos que aumentan la probabilidad de sufrir una enfermedad cardiovascular en el paciente pueden clasificarse en tres categorías principales. En primer lugar, las variables metabólicas modificables incluyen el aumento en los niveles de glicemia, colesterol total, creatinina y el índice de filtración glomerular estimado. En segundo lugar, las variables conductuales modificables comprenden la situación de obesidad, la falta de actividad física y el hábito de fumar. Por último, los antecedentes de enfermedades previas que contribuyen a esta probabilidad son sufrir de diabetes mellitus tipo II, ser hipertenso y tener antecedentes previos de una enfermedad cardiovascular, como un infarto agudo de miocardio o un accidente cerebrovascular.

Los atributos que poseen un mayor peso en la disminución de la probabilidad de padecer una enfermedad cardiovascular incluyen realizar actividad física de manera regular, mantener un peso dentro de los rangos considerados como normales, no padecer de síndrome metabólico y mantener un perfil bioquímico estable, es decir, no exceder los niveles normales de triglicéridos, hemoglobina, presión arterial, entre otros.

Cuadro 3. Características de pacientes con y sin riesgo de sufrir una CVD

Pacientes con riesgo de CVD		Pacientes sin riesgo de CVD	
Atributo	Peso	Atributo	Peso
N_Fact_Riesgo	2.25	Actividad Física = SI	-1.26
N_Hist_CVD	1.42	Estado Nutricional = Peso Normal	-1.05
Estado nutricional = Obesidad	1.41	Síndrome metabólico = NO	-0.81
Actividad física = NO	1.25	Í. Aterogenecidad= Bajo Riesgo	-0.54
Glicemia	1.03	Ratio S/P	-0.40
Síndrome Metabólico = SI	0.82	Fumador = NO	-0.35
Creatinina	0.56	Presión sistólica	-0.17
Í. Aterogenecidad= Riesgo Intermedio	0.54	Colesterol HDL	-0.16
Fumador = SI	0.35	Hemoglucotest	-0.16
Colesterol Total	0.35	Triglicéridos	-0.13
Edad	0.26	Sexo = Mujer	-0.04
IfGe	0.26		
Adherencia Dieta = NO	0.11		
Sexo = Hombre	0.04		

Determinación umbral óptimo

Para un análisis más profundo para la implementación del modelo, se analizan los costos de error de clasificación. Los costes de clasificación son calculados de la siguiente manera:

$$Costo = \text{TFN} * C(0,1) + TFP * C(1,0) \quad (5)$$

Donde TFN representa la tasa de falsos negativos; TFP la tasa de falsos positivos; C(0,1) como el costo de cometer un Error de Tipo I y a C(1,0) como el costo de cometer un Error de Tipo II. Dado que TFN=1-TFP, la curva de costos puede ser representada en términos de las dimensiones de la Curva ROC por medio de la siguiente ecuación.

$$TVP = \frac{C(1,0)}{C(0,1)} * TFP + 1 \quad (6)$$

El umbral óptimo se selecciona minimizando la combinación de los costos asociados a los falsos positivos y falsos negativos, esto se logra identificando el punto de intersección entre la curva de costos y la Curva ROC. Este enfoque permite encontrar un equilibrio entre la sensibilidad y la especificidad del modelo, optimizando la toma de decisiones según los costos asociados a los errores de clasificación. Dado el contexto de la investigación los dos tipos de costos son estimados mediante consultoría al equipo médico y finanzas del Cesfam Portezuelo de tal manera que:

- Error Tipo I : Indicar que un paciente sufrirá un infarto cuando no lo sufrirá, por ende, la evaluación económica está dada por el cálculo de asistir a un paciente mediante un programa crónico cardiovascular, teniendo un costo de 575.063 pesos chilenos, en el que se consideran: constantes monitores trimestrales, pago de remuneraciones al

personal (administrativo, kinesiólogo, nutricionista, medico, enfermero, etc), mantención de infraestructuras, servicios básicos y todo de tipo de coste relacionado a un centro de atención cardiovascular.

- Error Tipo II: Indicar que un paciente no sufrirá un infarto cuando lo sufrirá, por ende, la evaluación económica está relacionado a los costos de asistir al paciente tras sufrir una enfermedad coronaria y posteriormente tratarlo en el programa crónico, teniendo un costo total de 1.092.773 pesos chilenos, en el cual se consideran: traslado de paciente, costo día-cama hospitalización y asistencia en un programa CVD.

Tras obtener ambos costes de clasificación se obtiene el umbral óptimo con el que se distinguirá entre clases positivas y negativas. El umbral que minimiza el costo de error de clasificación resulta ser de 0.45. Esto quiere decir que para clasificar un paciente que sufrirá una enfermedad cardiovascular su probabilidad debe ser mayor a este valor.

Esquema de implementación del modelo predictivo

Se muestra el proceso para llevar a cabo la implementación de la investigación en el contexto de los programas preventivos de CVD. Por medio de la Figura 4 se visualizan las diferentes fases de la estrategia, desde la extracción de datos de los pacientes a predecir desde la base de datos, aplicación del modelo de SVM y la clasificación de pacientes entre propensos a sufrir o no una CVD, luego de clasificar a estos pacientes se busca que el equipo médico trabaje con aquellos pacientes con un mayor riesgo acorde a los resultados del modelo.

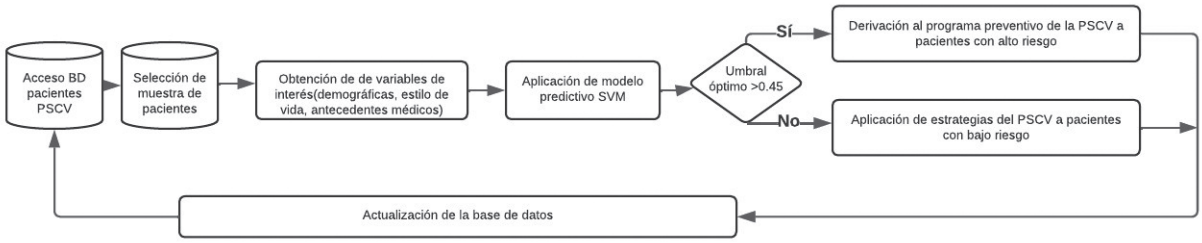

Figura 4: Esquema de implementación modelo predictivo

Discusiones

La detección de enfermedades cardiovasculares mediante el uso de machine learning ha sido un problema ampliamente estudiado por la literatura durante la última década, permitiendo una mejora en la precisión diagnóstica [15]. Los modelos implementados durante la investigación entregan resultados aceptables en función de detectar oportunamente una CVD, con ello es posible afirmar que el uso de algoritmos de aprendizaje son una herramienta eficaz que permita servir como punto inicial para con ella trazar planes de acción y medidas preventivas en aquellos pacientes clasificados como pacientes propensos a sufrir una CVD. Entre los modelos, destaca el desempeño del modelo SVM, lo cual es coincidente con múltiples investigaciones aplicadas a este mismo sector [7], [12], [16].

Respecto al estudio de la contribución de las variables en el modelo predictivo, estas son concordantes con lo hallado en la literatura. En la cual se identifican factores que pueden ser categorizados en factores de riesgos conductuales modificables y factores de riesgo metabólicos y genéticos. El método de extracción de pesos en el modelo SVM entregó una estimación aproximada del grado de contribución de los atributos sobre la probabilidad de

padecer una CVD, de esta forma se afirma que la afección de una enfermedad cardiovascular puede reducir su probabilidad si es que el individuo altera o disminuye sus factores de riesgo metabólicos que contribuyen de manera importante tales como: presión arterial, glicemia, triglicéridos, entre otros mediante el cambio de factores de riesgo conductuales tales como: un aumento en la actividad física y el no consumo de tabaco, entre otros [11].

Se comprueba empíricamente la necesidad de realizar un trabajo estrecho con el equipo médico, existiendo una entrega de conocimiento bidireccional en la que el equipo médico pudiese proveer de información al desarrollador del modelo, generando así una mejora continua sobre el enfoque predictivo, y de manera simultánea, el investigador puede resaltar la importancia de capturar una mayor cantidad de atributos del paciente y la necesidad de evitar inconsistencias en el ingreso de datos del paciente, lo que proporcionará un conjunto de datos más robusto que mejore el desempeño predictivo de los modelos [12].

Conclusiones

La presente investigación aborda el estudio del problema de detección de enfermedades cardiovasculares mediante la implementación de modelos predictivos basados en el conjunto de datos del programa cardiovascular y preventivo 2023 del Centro de Salud Familiar Portezuelo. Dado los resultados obtenidos de los cuatro modelos predictivos: Naive Bayes, Red Neuronal, Support Vector Machine y Árbol de Decisión, en el que todos los modelos se obtiene un porcentaje de precisión superior al 75% es posible afirmar que son una herramienta eficaz para discriminar entre individuos que padecerán de una CVD y los que no.

El modelo de SVM es el modelo que mejor desempeño presenta acorde a los objetivos del estudio planteados, obteniendo en todas sus métricas de desempeño un valor superior al 85% (Precisión, Recall, Exactitud y F1-Score). Además, se obtiene un valor de curva ROC de 0.93 para este modelo. Lo que indica que el modelo es eficaz para predecir correctamente y de ofrecer un equilibrio entre la predicción de las clases.

Por medio del método de obtención de pesos del modelo SVM se identificaron factores de riesgos que aportan una mayor contribución para el modelo predictivo, tales como: glicemia, presión arterial, antecedentes familiares, nivel del colesterol y edad del individuo. La contribución de las demás variables de estudio, aunque fuese menor también se considera relevante para el estudio. Se identifica que existen factores de riesgo conductuales y metabólicos que son categorizados como modificables, por ende, estas variables pueden ser tratadas y alteradas mediante la aplicación de planes de acción preventivos a fin de disminuir la tasa de CVD en la población bajo estudio.

El estudio abordó la determinación de asumir costos de clasificación para los errores de Tipo I y Tipo II, de esta forma se calculó un valor umbral óptimo de 0.45, lo que significó que para que el modelo considere que un paciente sufra una CVD este debe tener una probabilidad superior a la de este umbral. De esta forma se optimiza la toma de decisiones clínicas.

A raíz del presente estudio surgen nuevas consideraciones que pueden ser tomadas como futuras investigaciones, se proponen algunas como: la incorporación de nuevos atributos para el conjunto de datos que considere factores sociales tales como: niveles de estudio y clase socioeconómico del individuo. Como último punto, se propone construir una aplicación web móvil para el equipo médico del Cesfam Portezuelo con el cual sea posible utilizar el modelo predictivo y entregar un resultado de forma automática al ingresar registros de nuevos usuarios.

References

[1] NCD Risk Factor Collaboration (NCD-RisC), «Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019: a pooled analysis of 1201 population-representative studies with 104 million participants.,» The Lancet, vol. 398, pp. 957-980, 2021.

[2] A. Gheorghe, U. Griffiths, A. Murphy, H. Legido, P. Lamptey y P. Perel, «The economic burden of cardiovascular disease and hypertension in low- and middle-income countries: a systematic review,» *BMC Public Health*, vol. 18, nº 975, 2018.

[3] Organización Mundial de la Salud, «Organización Mundial de la Salud,» [En línea]. Available: https://www.who.int/es. [Último acceso: 05 01 2024].

[4] Departamento de Estadísticas e Información de Salud, «Departamento de Estadísticas e Información de Salud,» [En línea]. Available: https://deis.minsal.cl/. [Último acceso: 01 05 2024].

[5] P. Varleta, M. Acevedo, G. Valentino, S. Brienza y M. del Sueldo, «Conciencia de Enfermedad Cardiovascular y Conocimiento sobre Factores de Riesgo y Prevención Cardiovascular: Resultados Chilenos de Encuesta SIAC Cono Sur,» *Revista chilena de cardiología*, vol. 41, nº 2, 2022.

[6] Y. Sethi, N. Patel, N. Kaka, O. Kaiwan, J. Kar, A. Moinuddin, A. Goel, H. Chopra y S. Cavalu, «Precision Medicine and the future of Cardiovascular Diseases: A Clinically Oriented Comprehensive Review,» *Journal of Clinical Medicine*, vol. 12, nº 5, p. 1799, 2023.

[7] J. Campo, K. Parra, F. Mendoza y A. De La Hoz, «Análisis de técnicas de MD en diagnóstico de enfermedades cardiovasculares,» ***EIEI ACOFI***, 2015.

[8] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande y P. Singh, «Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning,» *Computational Intelligence and Neuroscience*, 2021.

[9] A. Gupta, R. Kumar, H. Singh y B. Raman, «MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis,» IEEE Access, vol. 8, pp. 14659-14674, 2020.

[10] E. Sánchez, Y. Hernandez y J. Ortiz, «Breve revisión de la literatura sobre Modelos Predictivos para la Detección de Enfermedades Cardiovasculares,» *Ciencia y Tecnología Aplicada*, vol. 5, nº 1, 2022.

[11] J. Moon, H. Posada y K. Chon, «A literature embedding model for cardiovascular disease prediction using risk factors, symptoms, and genotype information,» *Expert Systems with Applications*, vol. 213, nº A, p. 118930, 2023.

[12] O. Chavez, L. Galindo, A. Barriendos y M. Cuadros, «Aplicación Móvil para Predecir la Probabilidad de Pertenecer al Grupo de Riesgo Cardiovascular Utilizando Machine Learning,» *International Institute of Informatics and Cybernetics,* pp. 158-163, 2022.

[13] S. Timarán, I. Hernández, S. Caicedo, A. Hidalgo y J. Alvarado, «El proceso de descubrimiento de conocimiento en bases de datos,» de *Descubrimiento de patrones de desempeño acádemico con árboles de decisión en las competencias génericas de la formación profesional*, Bogotá: Ediciones Universidad Cooperativa de Colombia, 2016, pp. 63-86.

[14] C. Andersson, M. Nayor, C. Tsao y R. Vasan, «Framingham Heart Study: JACC Focus Seminar,» *J Am Coll Cardiol*, vol. 77, nº 21, pp. 2680-2692, 2021.

[15] «Una revisión de la literatura para la detección y proyección de enfermedades cardiovasculares mediante el aprendizaje automático,» *EAI Endorsed Transactions on Internet of Things,* vol. 10, 2024.

[16] H. Ayatollahi y L. Gholamhosseini, «Predicting coronary artery disease: a comparison between two data mining algorithms,» *BMC Public Health*, vol. 19, nº 448, 2019.

[17] H. Benhar, A. Idri y J. Fernández, «Data preprocessing for heart disease classification: A systematic literature review,» *Comput Methods Programs Biomed*, vol. 195, 2020.

Declaración sobre uso de Inteligencia Artificial (IA)

Los autores aquí firmantes declaramos que no se utilizó ninguna herramienta de IA para la conceptualización, traducción o redacción de este artículo.