1. **Nombre del proyecto:**

CARDIOSIGHT – Calculadora de Riesgo Cardiovascular

1. **Eje rector del proyecto**

El presente proyecto está orientado al área de salud pública en México ya que se enfoca en crear una herramienta que permita disminuir la incidencia de nuevos casos de enfermedades cardiovasculares en el país.

1. **Planteamiento del problema**

Las enfermedades cardiovasculares son un grupo de desórdenes que afectan al corazón y/o los vasos sanguíneos [1], las cuales representan un problema de salud pública en México ya que han mostrado un crecimiento en las últimas décadas, hasta constituir la primera causa de muerte en la población adulta [2].

Estudios epidemiológicos han permitido identificar un conjunto de variables denominadas factores de riesgo de tipo cardiovascular (FRCV), que son todas aquellas características biológicas no modificables y características conductuales modificables (estilos de vida), cuya presencia se relaciona con una mayor probabilidad de sufrir una enfermedad cardiovascular en el futuro [3]. Su identificación temprana es fundamental para implementar cambios en los hábitos de los pacientes con el objetivo de prevenir un evento de enfermedad cardiovascular primario.

En los últimos años México ha implementado varias estrategias enfocadas en el sector salud que prometen buenos resultados, como lo son los programas de detección temprana, modelos preventivos y programas educacionales [4]. A pesar de esto, se han encontrado inconsistencias en los procesos de atención primaria, práctica clínica e instrumentos de vigilancia que limitan la monitorización cardiovascular de la población, ya sea por desconocimiento del personal de salud acerca de la existencia de los programas, deficiencias en la infraestructura y equipamiento básico en los hospitales públicos, o por la falta de difusión de programas preventivos hacia los diferentes grupos de población (niños, adolescentes, adultos mayores). Factores que influyen negativamente en la implementación de prácticas para la prevención de nuevos casos y promoción de hábitos saludables [5, 6, 7, 8].

1. **Descripción de la solución a la problemática detectada.**

Se propone el desarrollo de una aplicación que permita calcular a través de la inteligencia artificial, el factor de riesgo de una persona a sufrir de una enfermedad cardiovascular. Herramienta que ayudará a la población a implementar acciones de naturaleza preventiva y estrategias para tener un estilo de vida más saludable.

Para esto se utilizará un dataset que contenga los FRCV de una muestra de la población, además de una variable que mencione si esta persona sufre de una enfermedad de esta índole.

La aplicación tiene ciertas ventajas:

● La población tendrá la facilidad de conocer su nivel de riesgo sin la necesidad de asistir a una unidad médica de salud. Acelerando de esta manera la identificación de los individuos con un riesgo alto antes de que se desarrolle un evento cardiovascular primario.

● Posterior a la identificación, dependiendo del resultado, se le sugerirá al usuario visitar las páginas disponibles del Instituto Mexicano del seguro social que brindan la información sobre las campañas preventivas actuales relacionadas a este tipo de enfermedades, en las cuales encontrarán la información necesaria para educarse sobre la problemática, además de la posibilidad de encontrar su clínica más cercana para realizar su primer chequeo médico.

1. **Hipótesis.**

Es posible desarrollar un algoritmo de machine learning que permita realizar un cálculo de factor de riesgo de sufrir una ECV y que sea al menos tan preciso como los mencionados en la literatura.

1. **Metodología**

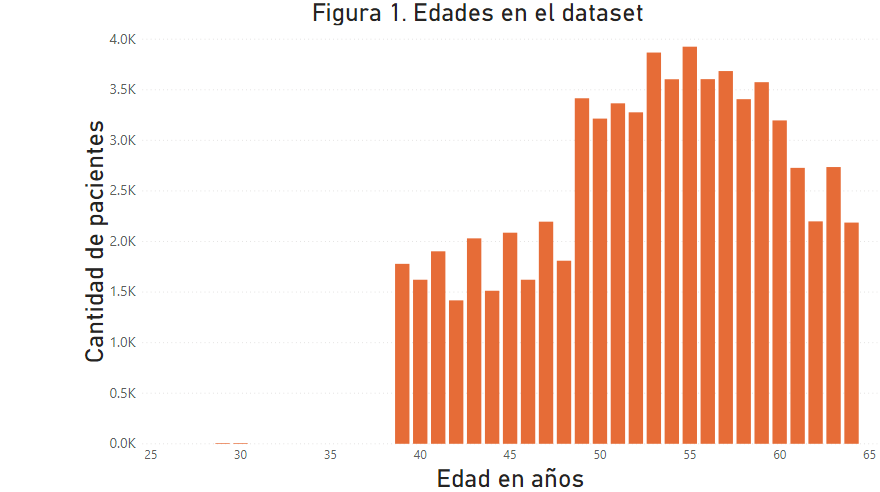
(a) Descripción de los datos:

Se decidió trabajar con el dataset *Cardiovascular Disease* encontrado en la página Kaggle. El cual consiste en 70,000 datos recabados de pacientes, contiene 11 características y una variable objetivo, información recabada por medio de resultados médicos, compartida por el paciente o de forma factual.

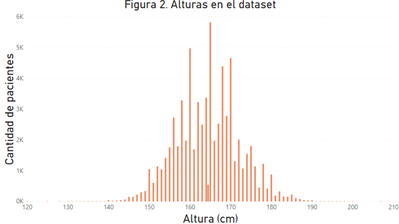
|  |  |
| --- | --- |
| **Características** | **Tipo de variables / Tipo de unidades** |
| Edad | int - Día |
| Altura | int - cm |
| Peso | Float - Kg |
| Género | int - Valor binario |
| Presión arterial sistólica | int - mmHg |
| Presión arterial diastólica | int - mmHg |
| Cholesterol | int - Valor categórico |
| Glucosa | int - Valor categórico |
| Hábito - Fumar | int - Valor categórico |
| Hábito - Tomar alcohol | int - Valor categórico |
| Hábito - Realizar actividad física | int - Valor categórico |
| Presencia o ausencia de ECV | int - Valor categórico |

Lo primero que se realizó fue un análisis exploratorio de los datos, en el cual se logró visualizar de forma gráfica el comportamiento de las variables. Durante este proceso se detectaron registros duplicados los cuales fueron eliminados del dataset, también se encontraron algunas variables que no estaban dentro de los parámetros fisiológicos “normales” de un humano, los cuales fueron modificados, por último, se crearon la variable BMI (índice de masa corporal) con ayuda del peso y la altura. A continuación, se muestra una parte del análisis correspondiente a cada una de las variables:

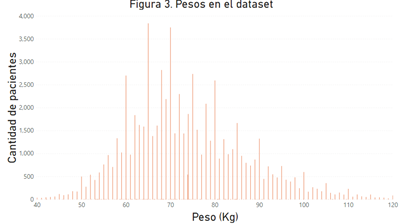
**Edad:** La variable edad representa la edad de todos los pacientes en días. Para efectos prácticos se realizó una conversión a años dividiendo los días entre 365. De esta manera se logró obtener la lista de edad, siendo la menor de 30 y la mayor de 65 años (Figura 1).



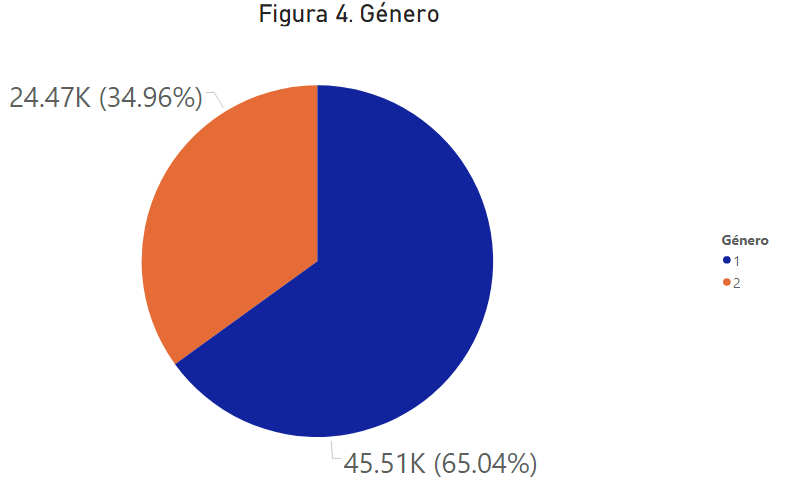
**Altura:** La variable representa la altura de todos los pacientes en centímetros (cm). Se detectó una inconsistencia con algunos datos, ya que no corresponden a la altura “factible” de un humano promedio. Más adelante en la sección de índice de masa corporal se abordará en su procesamiento, su representación gráfica fue producida posterior a la transformación de dichos datos (Figura 2).



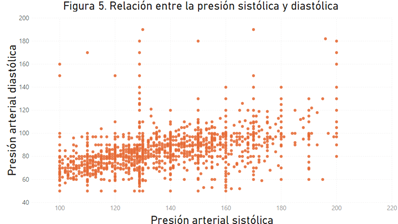
**Peso:** La variable representa el peso todos los pacientes en kilogramos (kg). Se detectó una inconsistencia con algunos datos, ya que no corresponden al peso “factible” de un humano promedio. Más adelante en la sección de índice de masa corporal se abordará en su procesamiento, su representación gráfica fue producida posterior a la transformación de dichos datos (Figura 3).



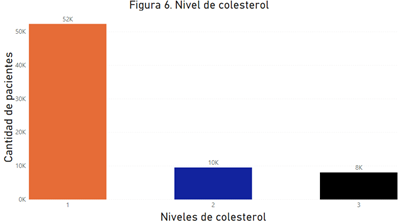
**Género:** La variable representa el género de todos los pacientes. Siendo 1 mujer y 2 hombre, esta variable se encuentra pobremente equilibrada (Figura 4).



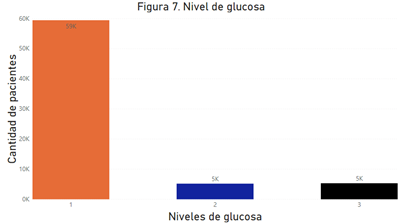
**Presión arterial sistólica y diastólica:** Ambas variables son representadas en el dataset como números enteros, sus unidades son expresadas en milímetros de mercurio (mmHg). Se detectó que los valores de ambas variables no estaban dentro de los rangos “factibles” para un humano promedio. Siendo un rango normal para la presión sistólica de 120 mmHg a 180 mmHg y para la diastólica de 80 mmHg a 100 mmHg. En ambos casos se decidió cambiar los valores que no entraban dentro del rango correspondiente por la media de los valores que sí (Figura 5).



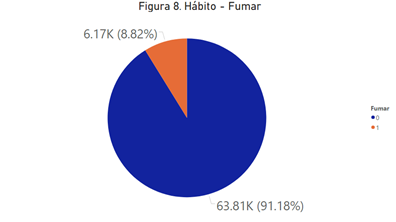
**Colesterol:** La variable colesterol representa el nivel de colesterol del paciente, esta variable fue obtenida por medio de exámenes médicos. Puede tomar tres valores 1 o “normal” (menos de 200 mg/dl), 2 o “arriba de lo normal” (200 - 239 mg/dl) ó 3 o “muy arriba de lo normal” (más de 240 mg/dl). Esta variable no está bien balanceada ya que existen más pacientes con un nivel de colesterol “normal” (Figura 6).



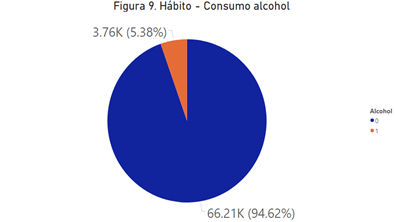
**Glucosa:** La variable glucosa representa el nivel de glucosa en la sangre del paciente, esta variable fue obtenida por medio de exámenes médicos. Puede tomar tres valores 1 o “normal” (menos de 100 mg/dl), 2 o “arriba de lo normal” (100 - 125 mg/dl) ó 3 o “muy arriba de lo normal” (más de 125 mg/dl). Esta variable no está balanceada ya que más pacientes con un nivel de glucosa “normal” (Figura 7).



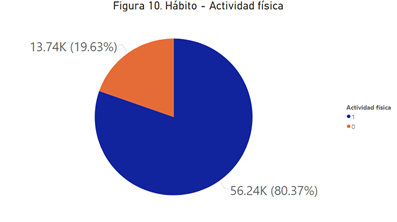
**Hábito - fumar:** La variable fumar representa si el paciente fuma con el número “1” o no fuma con el número “0”. La variable no se encuentra balanceada ya que hay muchos más pacientes que no fuman (Figura 8).



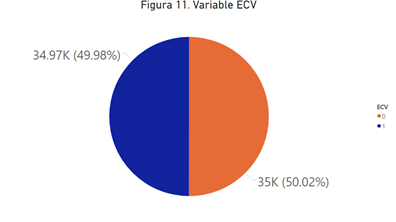
**Hábito - tomar alcohol:** La variable alcohol representa si el paciente consume alcohol con el número “1” o no consume alcohol con el número “0”. La variable no se encuentra balanceada ya que hay más pacientes que no consumen alcohol (Figura 9).



**Hábito - realizar actividad física:** La variable actividad física representa si el paciente realiza actividad física con el número “1” o no realiza actividad física con el número “0”. La variable no se encuentra balanceada ya que hay más pacientes que realizan actividad física (Figura 10).

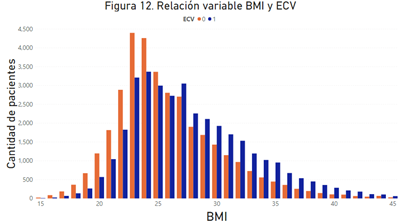


**ECV:** Esta es la variable objetivo, si un paciente tiene de alguna enfermedad cardiovascular se representará con un “1”, de lo contrario se representará con un “0”. La variable objetivo se encuentra equilibrada (Figura 11).



**BMI:** Con ayuda de las variables peso y altura se obtuvo la variable índice de masa corporal (BMI), una vez que se calcularon los valores correspondientes se encontró que muchos de ellos no se encontraban dentro de los parámetros “factibles”, 543 valores eran menores a un BMI de 15 o mayores de 45. Gracias a esta variable se identificó que los datos de peso y altura relacionados con esta métrica no estaban correctamente capturados, provocando de esta manera un cálculo erróneo.

Al igual que con la presión arterial, se transformó el peso y altura que no estaban dentro del parámetro a la media aritmética de la muestra que, si lo estaba, posteriormente se calculó nuevamente su BMI y se comprobó que los resultados estuvieran dentro de los parámetros correspondientes (Figura 12).



(a) Descripción del modelo:

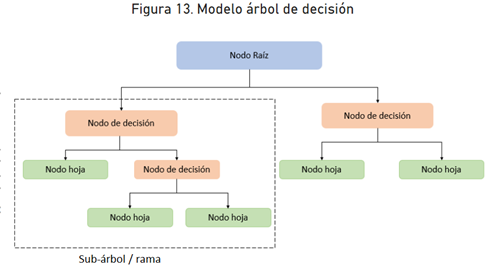
Debido a que el propósito del proyecto es detectar si una persona será propensa o no a sufrir de una enfermedad cardiovascular, se decidió trabajar con el algoritmo de clasificación *“Random forest”.*

**Árbol de decisión**

No se puede hablar del modelo “Random Forest” sin antes explicar el modelo de árbol de decisión. El cual es un método de aprendizaje supervisado usado para resolver problemas de clasificación (o regresión), es una técnica que divide iterativamente el dataset en dos o más muestras. El objetivo de este método es predecir la clase valor de la variable objetivo. Este modelo ayuda a segregar la información y construye el modelo de decisión para predecir la clase etiqueta desconocida. Puede ser construido con variables categóricas o continuas. El árbol de decisión de forma óptima encuentra el nodo raíz basándose en el valor de entropía más alto. Esto le proporciona al modelo una ventaja para seleccionar la hipótesis más consistente entre el dataset de entrenamiento.

La entrada del modelo son los atributos del dataset, y la salida el modelo de decisión (estructura de árbol). Los problemas que se presentan al construir un árbol de decisión es seleccionar el atributo de división, las divisiones, el criterio de parada, la muestra de entrenamiento, orden de división, entre otros.

El modelo de decisión tiene una estructura de árbol que incluye una colección de nodos: los nodos de decisión (nodo dividido en base a la condición), y los nodos hoja. Entre los varios atributos del dataset, seleccionar aquel que será el nodo raíz para empezar las divisiones en un trabajo difícil. Los nodos de decisión pueden tener dos o más ramas. Para empezar con el primer nodo raíz, el modelo predice el mejor atributo o el mejor nodo predictor del conjunto de nodos disponibles. Hay muchas maneras para seleccionar el mejor atributo como nodo raíz, basado en el grado de impureza de los nodos hijo. Las medidas de desempeño son entropía, índice Gini y el error de clasificación. Estas medidas son calculadas para todos los atributos y se realiza una comparación para seleccionar la mejor división (Figura 13).



**“Random Forest”**

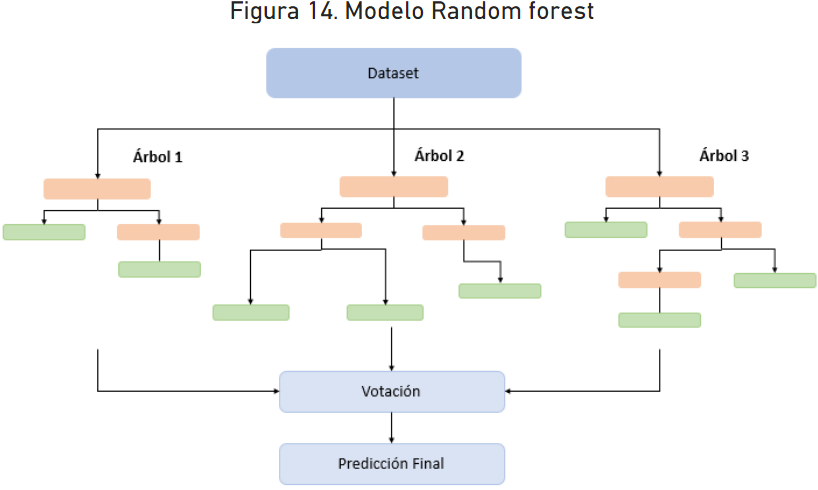
“Random Forest” es un modelo de aprendizaje supervisado, que puede ser utilizado para clasificación o regresión (Figura 14). La lógica detrás de este modelo es que se construye combinando las predicciones de varios clasificadores de árbol simple, los cuales son entrenados de forma aislada. Las predicciones de los árboles se combinan mediante un promedio. Pasos para aplicar el modelo de “Random Forest”:

1. Se carga la información que contiene “m” características que representan el comportamiento del dataset.
2. El algoritmo del modelo selecciona “n” de las “m” características para crear muestras de forma aleatoria.
3. Se calcula el nodo usando la mejor división. Se divide el nodo en sub-nodos.
4. Se repite el proceso hasta llegar a “n” números de árboles.
5. Se calcula el número total de votos de cada árbol para la variable objetivo. La clase con más votos es el predictor final del modelo.

Existen tres parámetros necesarios para construir un *“Random Forest”*: El método para dividir las hojas del árbol, tipo de predictor a usar en cada hoja, método para inyectar aleatoriedad [9].

Algunas de las ventajas de este modelo son las siguientes:

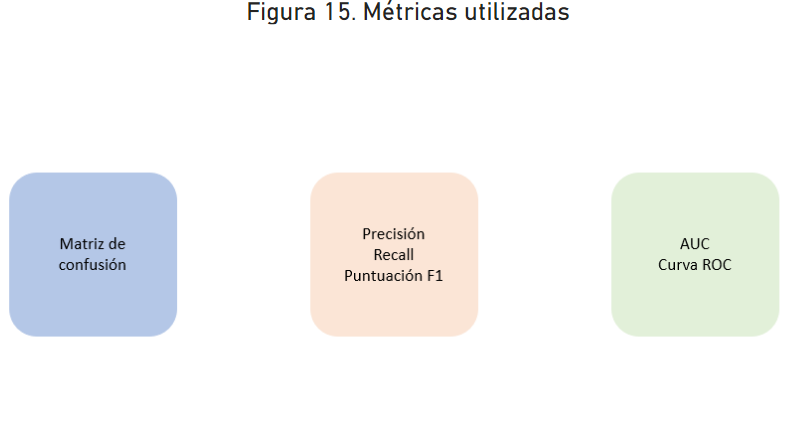
1. Útil para una interpretación visual, fácil de entender incluso por no expertos.
2. Robusto a ruido y overfitting.
3. Es un modelo de rápida ejecución.
4. Produce sus propias estimaciones de error, fuerza, correlación, e importancia de variables [10].



(b) Aplicación del modelo:

Se decidió aplicar el modelo de *“Random Forest”* en el dataset de tres maneras diferentes para realizar una comparación de resultados y seleccionar el modelo con el mejor desempeño. Para empezar con este procedimiento se dividió el dataset en un conjunto de entrenamiento y otro de prueba con una relación 70:30. Posteriormente se realizaron los siguientes entrenamientos:

i. Modelo “Random Forest” con parámetros por defecto: Se entrenó el modelo utilizando los parámetros por defecto, se obtuvo su exactitud, matriz de confusión, reporte de clasificación que incluye precisión, *recall* y puntuación F1. Además de su *AUC* y curva *ROC*. Estas métricas desempeño sirvieron como línea base de comparación (Figura 15).



ii. Modelo “Random Forest” con tuneo de hiper parámetros: Se utilizó el mismo conjunto de prueba y entrenamiento. Después de consultar la bibliografía existente sobre este modelo, se seleccionaron los hiperparámetros con los cuales se deseaba experimentar las técnicas de tuneo.

|  |  |
| --- | --- |
| **Hiperparámetro** | **Descripción** |
| n\_estimators | Número de árboles de decisión |
| max\_features | Máximo número de características consideradas para el nodo de decisión |
| max\_depth | Máximo número de niveles en cada árbol de decisión |
| min\_samples\_split | Mínimo número de puntos de datos en un nodo antes de que el nodo sea dividido |
| min\_samples\_leaf | Mínimo número de puntos de datos permitidos en un nodo hoja |
| bootstrap | Método de muestreo de puntos de datos |

Se definió una lista con los valores deseados de cada hiperparámetro, con 30 valores diferentes, 100 iteraciones y 3 divisiones de validación cruzada. Para esto se utilizó la función de Python *“RandomizedSearchCV”*, con el objetivo de disminuir el tiempo de análisis.

Al final se compararon los modelos obtenidos, se seleccionó el de mejor desempeño, posteriormente se entrenó con nuestros datos, se obtuvieron las predicciones y métricas mencionadas en el punto anterior.

iii. Modelo “Random Forest” con segundo tuneo de hiper parámetros: Se realizó un segundo tuneo con diferentes opciones de hiperparámetros, en esta ocasión con 19 valores diferentes, conservando la misma cantidad de iteraciones y divisiones en la validación cruzada. Se obtuvieron las predicciones y métricas anteriormente mencionadas.

(c) Punto de corte:

El modelo de “Random Forest” no tiene como salida etiquetas binarias, lo que produce son probabilidades en una escala de 0 a 1. Estas probabilidades sólo son transformadas a predicciones binarias por medio del uso de puntos de corte. El punto de corte es la probabilidad arriba de la cual todas las predicciones son declaradas como positivas, y debajo de esta negativas. El punto de corte predefinido por la librería *scikit-learn* es de 0.5. Esto significa que cada muestra con una probabilidad predicha de al menos 0.5 será declarada como positiva, y aquellas debajo se clasificará como negativas [11].

Se decidió hacer una comparativa de métricas entre el modelo de mejor desempeño con punto de corte predefinido y un punto de corte calculado utilizando el índice de Youden, prueba estadística que permite un balance entre sensibilidad y especificidad.

(d) Limitaciones:

i. De datos:

El objetivo del presente proyecto fue el de desarrollar una herramienta para calcular el nivel de riesgo de una persona a sufrir de una enfermedad cardiovascular, para esto era necesario contar con un dataset que contuviera los FRCV de una muestra de la población mexicana. Lamentablemente no se pudo encontrar un dataset público o privado que con este tipo de información.

Es por ello que se decidió trabajar con el dataset el dataset *Cardiovascular Disease* que se puede encontrar en la página Kaggle. A pesar de que no es de población originaria de México, sirve como antecedente para nuestra metodología y comprobar si es funcional.

ii. Temporales:

Debido a que se está trabajando con un dataset público no se tiene la fecha exacta en la que estos fueron recabados. Sería preferible trabajar con datos actualizados, para conocer de forma concreta como las enfermedades cardiovasculares afectan a la población considerando todos los factores actuales de alimentación, económicos y sociales, de esta manera se podría realizar un estudio más minucioso.

iii. Teóricas:

La mayoría de los datos pertenecen a la categoría médica, al no tener conocimiento profundo sobre la materia se tuvieron algunos problemas al tratar de transformar los datos, un ejemplo de esto fue la presión arterial, altura, peso e índice de masa corporal, los cuales no entraban en un rango considerado “normal” para un individuo promedio. La transformación de estos datos fue llevada a cabo consultando la bibliografía existente sobre información fisiológica y antropométrica.

Además de esto, al no tener un amplio conocimiento de machine learning se invirtió considerable tiempo en el estudio de los modelos más utilizados, así como la información referente a sus parámetros y el ajuste de estos. El reto más grande fue el de interpretar los resultados obtenidos, no solo de forma visual sino también de forma estadística. Se logró superar estos retos gracias a la bibliografía existente y proporcionada, además del apoyo por parte de nuestro mentor.

1. **Marco teórico**

Consultando la bibliografía existente se pueden encontrar dos tipos de estudios realizados: Los enfocados en comparar el desempeño estadístico de las puntuaciones de riesgo más comunes como *SCORE* y *Framingham*, y aquellos que comparan el desempeño de diferentes algoritmos de machine learning por medio de diferentes métricas. La métrica de comparación común entre ambos casos en el *AUC (área bajo la curva ROC)*.

En uno de los artículos considerados relevantes para este proyecto Álvarez, Díaz, López y Prieto (2005) hacen una comparativa entre las puntuaciones de riesgo *SCORE* y *Framingham*. Su estudio se enfoca en una muestra de población de varones con edades de 45 a 65 años. Recabaron información de 379 participantes, que incluía edad, antecedentes familiares de enfermedad coronaria o muerte súbita, antecedentes personales de patología cardiovascular, presión arterial, diabetes, hábito tabáquico y perfil lipídico. Posteriormente obtuvieron las puntuaciones de riesgo con los métodos anteriormente mencionados.

La información más importante en este estudio son sus conclusiones, ya que menciona que no se puede aplicar a la población general, debido al rango y género que seleccionaron, además de que no se puede comparar el riesgo calculado por *SCORE* (mortalidad cardiovascular) y *Framingham* (Riesgo coronario) ya que su propósito difiere. Concluyen que lo ideal sería conocer el riesgo cardiovascular general, pero que es difícil elegir un punto de corte para el alto riesgo cardiovascular. Ambos métodos mostraron sobre estimación de los resultados, y recomiendan que se debe esperar al desarrollo de otros métodos de cálculo de riesgo adaptados a la población de su país.

Se consultó un segundo artículo para encontrar más información relacionada a la comparación de desempeño entre estos dos métodos de cálculo mundialmente utilizados, donde Yuksel, Karagoz, Bektas, Kaya, Kiris, Erdogan e Isik (2016) realizaron un estudio con 227 personas con una edad media de 63.3±9.2, en el cual clasificaron a personas con un alto y bajo riesgo con tres métodos, *Framingham*, *SCORE* y *SYNTAX*, método predictivo de muerte e infarto del miocardio obtenido por medio de una angiografía coronaria.

En este documento se mencionan las diferentes características que los primeros métodos de cálculo utilizan, siendo *SCORE* el que requiere de una menor cantidad. A pesar de esto, se realizó en análisis de importancia de estas características, siendo para el modelo de *Framingham* la hipertensión, colesterol total y colesterol HDL, y para *SCORE* la hipertensión y colesterol total. Se concluyo que ambos métodos son efectivos en predecir la presencia de enfermedad de arteria coronaria y que ninguno de los modelos tiene un desempeño superior al otro ya que para *Framingham* se obtuvo un *AUC* de 0.819 y para *SCORE* de 0.811.

Este estudio hace hincapié en que se necesita una evaluación más completa de ambos métodos considerando las características de la población a la cual se requiere aplicar. Además, muestra evidencia de que en algunos países su utilización puede producir una sobreestimación de los resultados.

Por último, se consideró el estudio realizado por Weng, Reps, Kai, Garibaldi y Qureshi (2016), en el cual se trabajó con la información de 378,256 pacientes, con edades de 30 a 84 años a los que se les dio un seguimiento durante 10 años. Se utilizaron cuatro modelos de *machine learning*: regresión logística, *random forest*, *gradient boosting machines* y redes neuronales. Se obtuvieron las métricas de exactitud, *AUC*, sensibilidad, especificidad, positivos verdaderos y positivos falsos. Se realizó una comparación de los resultados obtenidos con los algoritmos y el método de cálculo de riesgo *ASCVD*.

Se demostró que los algoritmos de *machine learning* utilizados tuvieron un mejor desempeño en la predicción que *ASCVD*. Siendo el *AUC* de: 0.728 para *ASCVD*, 0.745 para *Random Forest*, 0.760 para regresión logística, 0.761 para *gradient boosting machines* y 0.764 para redes neuronales.

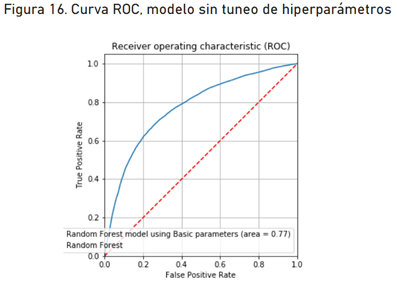
1. **Resultados**

(a) Aplicación del modelo:

Al aplicar el modelo de *“Random Forest”* sin ajuste de hiperparámetros se obtuvieron las siguientes métricas para clase 1 (ECV detectada) y clase 0 (ECV no detectada):

|  |  |
| --- | --- |
| **Métricas** | **Valor** |
| Exactitud | 0.71 |
| Precisión - Clase 1 | 0.71 |
| Recall - Clase 1 | 0.7 |
| Puntuación F1 - Clase 1 | 0.71 |
| Precisión - Clase 0 | 0.71 |
| Recall - Clase 0 | 0.72 |
| Puntuación F1 - Clase 0 | 0.72 |

|  |  |
| --- | --- |
| **Matriz de confusión** | |
| 7623 | 2917 |
| 3138 | 7315 |



Basándonos en los resultados obtenidos con este modelo podemos decir que tiene un mayor desempeño prediciendo los verdaderos negativos que los verdaderos positivos. Como se mencionó anteriormente estos resultados se utilizaron como línea base para comparar los siguientes modelos.

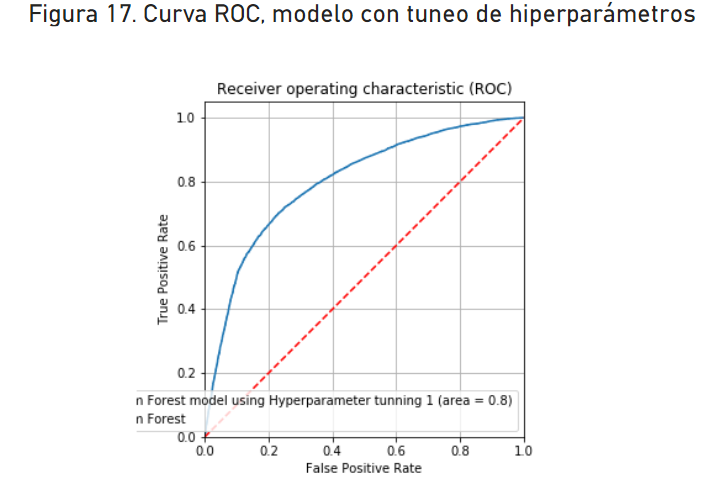
En el primer ajuste de hiperparámetros se encontró que el modelo con un mejor desempeño fue el siguiente:

|  |  |
| --- | --- |
| **Hiperparámetro** | **Valor** |
| n\_estimators | 100 |
| max\_features | ‘sqrt’ |
| max\_depth | 10 |
| min\_samples\_split | 10 |
| min\_samples\_leaf | 2 |
| bootstrap | *‘True’* |

Las métricas obtenidas son las siguientes:

|  |  |
| --- | --- |
| **Métricas** | **Valor** |
| Exactitud | 0.73 |
| Precisión - Clase 1 | 0.76 |
| Recall - Clase 1 | 0.69 |
| Puntuación F1 - Clase 1 | 0.72 |
| Precisión - Clase 0 | 0.72 |
| Recall - Clase 0 | 0.78 |
| Puntuación F1 - Clase 0 | 0.75 |

|  |  |
| --- | --- |
| **Matriz de confusión** | |
| 8227 | 2313 |
| 3266 | 7187 |



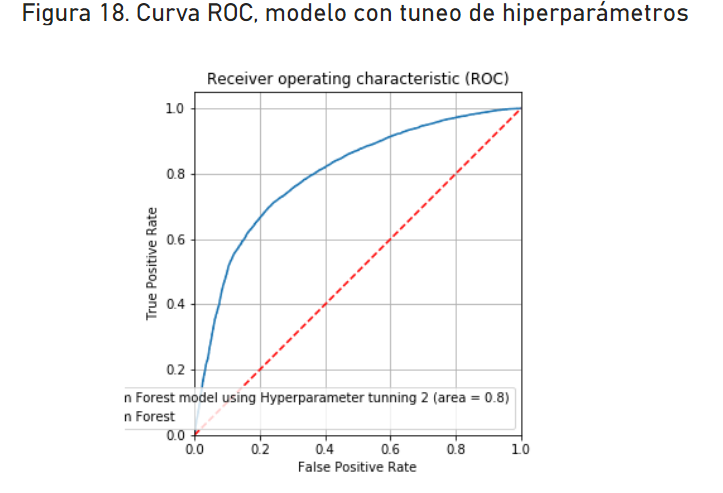
Se puede observar un aumento en la exactitud de +0.2 esto significa que el porcentaje total de elementos clasificados correctamente fue mejor este modelo. Se puede decir que este caso, que la exactitud si proporciona un significado de desempeño importante, ya que la variable objetivo está correctamente balanceada. Del mismo modo hay un aumento de Recall con lo cual podemos decir que este modelo puede identificar a un mayor número de individuos que verdaderamente no tienen una enfermedad cardiovascular. Gracias a la curva ROC podemos ver que el AUC aumentó, esto significa que el modelo funciona mejor para la clasificación en comparación con el primero.

En el último tuneo de hiperparámetros se obtuvieron los siguientes resultados:

|  |  |
| --- | --- |
| **Hiperparámetro** | **Valor** |
| n\_estimators | 100 |
| max\_features | ‘sqrt’ |
| max\_depth | 10 |
| min\_samples\_split | 10 |
| min\_samples\_leaf | 7 |
| bootstrap | *‘True’* |

|  |  |
| --- | --- |
| **Métricas** | **Valor** |
| Exactitud | 0.73 |
| Precisión - Clase 1 | 0.76 |
| Recall - Clase 1 | 0.69 |
| Puntuación F1 - Clase 1 | 0.72 |
| Precisión - Clase 0 | 0.72 |
| Recall - Clase 0 | 0.78 |
| Puntuación F1 - Clase 0 | 0.75 |

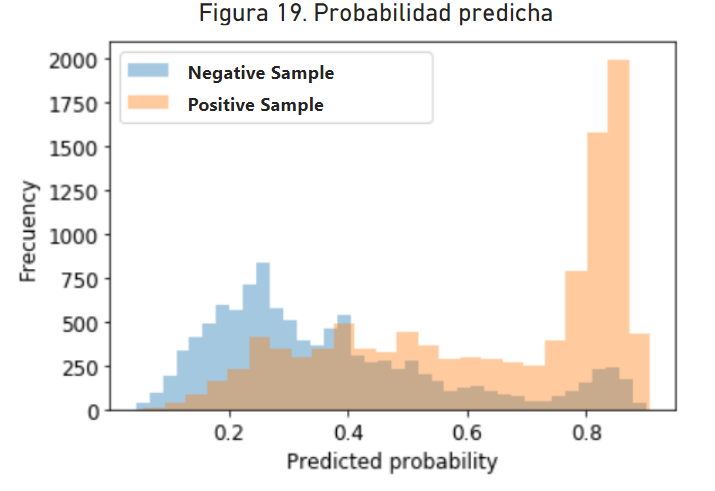
|  |  |
| --- | --- |
| **Matriz de confusión** | |
| 8222 | 2318 |
| 3259 | 7194 |



Debido a que el último modelo obtuvo como salida los mismos resultados que el segundo, se decidió trabajar con el segundo modelo y profundizar más en el comportamiento de sus probabilidades.

(b) Modificación de punto de corte

En la sección de punto de corte se mencionó que la librería *scikit-learn* tiene un punto de corte preestablecido de 0.5. Se creó una gráfica para realizar un análisis sobre el comportamiento de las probabilidades contra este punto y los valores reales de la variable objetivo (Figura 19).



Lo que está gráfica nos dice es que nuestro modelo es bueno para distinguir entre ambas categorías de 0 y 1 debido a que los verdaderos positivos se encuentran sesgados hacia el 1 y los verdaderos negativos sesgados hacia el 0. A pesar de esto, cuando se habla de una predicción médica se desea obtener un balance entre la sensibilidad y la especificidad, a través de la modificación del punto de corte. Para ello se utilizó el índice de Youden, el cual le proporciona un mismo peso a los falsos positivos y falsos negativos en todas las pruebas, entre más cercano es el valor a 1, el modelo tendrá un desempeño perfecto, diferenciando por completo ambas clases, y entre más cercano a 0 se tiene un desempeño pobre.

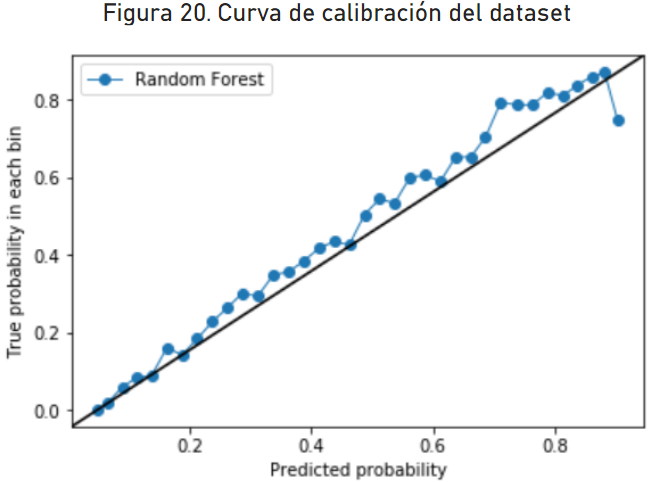
El índice de Youden de nuestro modelo fue de **0.469**, con el punto de corte de **0.475**. Al etiquetar las predicciones del modelo seleccionado con este punto de corte se obtuvieron las siguientes métricas:

|  |  |
| --- | --- |
| **Métricas** | **Valor** |
| Exactitud | 0.73 |
| Precisión - Clase 1 | 0.74 |
| Recall - Clase 1 | 0.71 |
| Puntuación F1 - Clase 1 | 0.73 |
| Precisión - Clase 0 | 0.73 |
| Recall - Clase 0 | 0.76 |
| Puntuación F1 - Clase 0 | 0.74 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Matriz de confusión 1** | |  | **Matriz de confusión 2** | |
| 8227 | 2313 |  | 7977 | 2563 |
| 3266 | 7187 |  | 3005 | 7488 |

Basándonos en el resumen de las métricas, se puede observar un equilibrio entre la sensibilidad y la especificidad, se puede decir que el índice de Youden funcionó como se esperaba.

Al momento de analizar las matrices de confusión del modelo seleccionado con punto de corte automático (Matriz de confusión 1) y del modelo con el punto de corte obtenido por medio del índice de Youden (Matriz de confusión 2), se puede apreciar un mejor balance entre los falsos negativos y falsos positivos de la segunda, provocando el mismo efecto entre los verdaderos positivos y los verdaderos negativos.



Cuando se resuelven problemas de clasificación muchas veces se desea no solo predecir las etiquetas (Ej. 0 ó 1), si no también obtener su puntuación de probabilidad, lo cual nos proporciona una especie de confianza y mayor interpretación en la predicción que sólo obtener 0 ó 1 como salida. Es en estas situaciones en las que las curvas de calibración o de fiabilidad son de utilidad, en las cuales se realiza una gráfica lineal de la frecuencia de probabilidad pronosticada (eje x) contra la frecuencia relativa del valor de probabilidad real (eje y).

Se dice que un modelo está bien calibrado si las probabilidades predichas tienen una proporción cercana a las probabilidades reales.

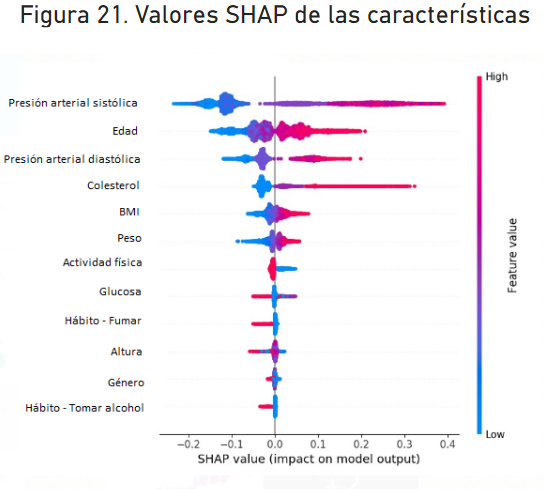
Para conocer el desempeño de las predicciones del modelo seleccionado se graficó su curva de calibración (Figura 20). Se puede observar que en la parte media de la gráfica los resultados están más separados de la línea diagonal que la parte inicial y final, se infiere que a pesar de que los puntos son muy cercanos a la línea, cuentan con un error de predicción que podría ser mejorado por medio de técnicas de calibración.

(c) Valores SHAP

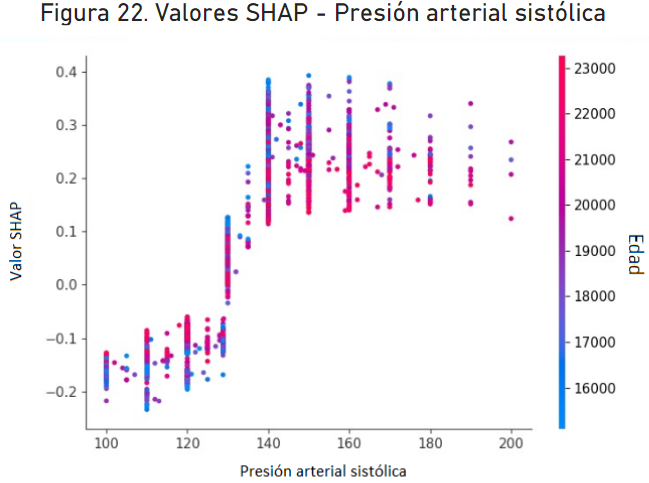
Con el objetivo de interpretar la importancia de cada característica en la predicción de la variable objetivo, se utilizó SHAP, un enfoque de la teoría del juego para explicar el resultado de cualquier modelo de aprendizaje automático. Uno de los beneficios de utilizarlo es que nos puede decir que tanto las características contribuyen de forma negativa o positiva a la predicción.

Se graficó el valor SHAP para todas las características utilizadas (Figura. 21), donde se muestran las variables más significativas para la predicción de forma descendente. Las primeras contribuyen más al modelo y tienen un mayor poder predictivo que las del fondo.

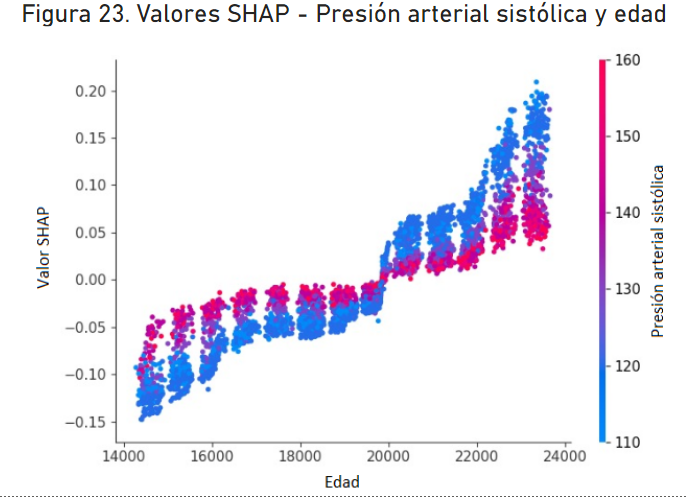
Se infiere que la presión arterial sistólica es la característica que posee el mayor impacto de forma positiva a las predicciones, siguiendo su paso se tiene a la edad, presión arterial diastólica, colesterol y el índice de masa corporal. Además, se puede decir que las características de altura, genero, hábito tabáquico y la ingesta de alcohol no tienen un efecto positivo en la predicción y que están negativamente correlacionadas a la característica objetivo.



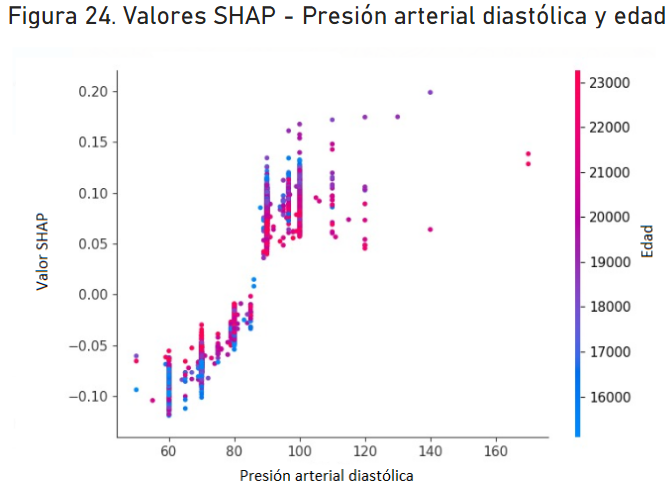
Para entender mejor el comportamiento del valor SHAP y la presión sistólica, se obtuvo su gráfica incluyendo a la edad (Figura 22). Se puede observar que a mayor presión arterial el valor de SHAP tiende a aumentar y la característica tiene un mayor impacto en la predicción, en estos puntos hay más personas que son mayores a 50 años.



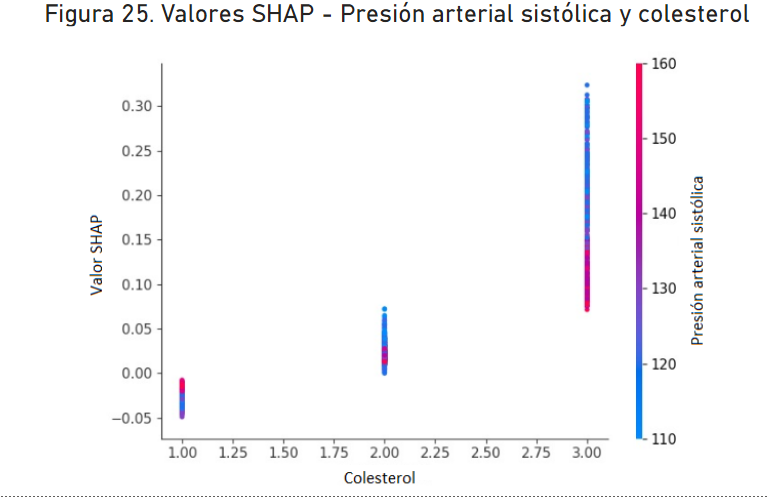
El valor de SHAP con la característica de edad (Figura 23) tiene un patrón marcado. Se observa que, a menor edad, la presión arterial sistólica tiende a ser menor, en estos puntos la característica tiene un poder predictivo negativo. A partir de los 54 años de edad la característica tiene un poder predictivo alto y positivo, también se infiere que la cantidad de personas con presión arterial sistólica tiende a incrementar.



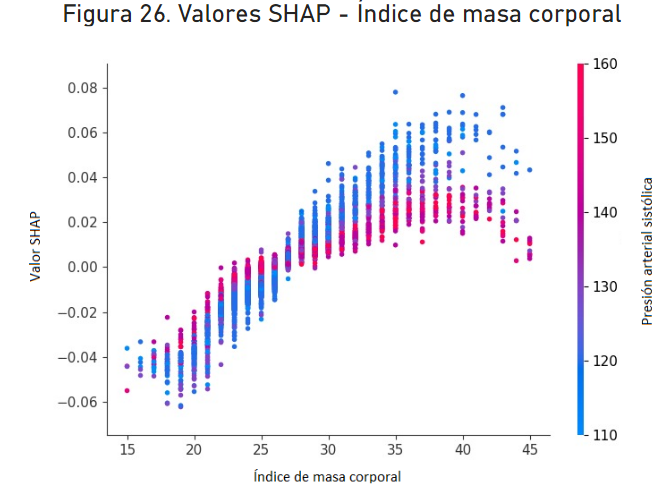
Al comparar el valor SHAP con la presión diastólica (Figura 24) se encuentra que a mayor presión diastólica la variable tiene un valor predictivo alto y positivo, además de que se puede observar una relación con la edad, debido a que en esta área hay una mayor cantidad de personas mayores a 54 años.



El comportamiento de la característica colesterol contra el valor SHAP (Figura 25) también muestra un patrón marcado, debido a que la característica tiene un nivel predictivo bajo y negativo cuando los individuos tienen un nivel de colesterol dentro de los parámetros normales, pero, cuando el nivel se encuentra arriba de éstos, la característica tiene un valor predictivo alto y positivo.



Por último, se tiene a la variable índice de masa corporal contra el valor de SHAP (Figura 26), se puede inferir que, a mayor grado de obesidad, la característica tiende a tener un poder predictivo más alto y positivo. Así mismo, en esta zona se puede encontrar a un mayor número de personas con presión arterial sistólica alta.



1. **Conclusiones**

Se logró obtener un algoritmo de inteligencia artificial capaz de predecir si una persona es más propensa a sufrir de una enfermedad cardiovascular, las métricas obtenidas son muy cercanas a las que Weng et al. (2016) reporta en su investigación, con un valor de 0.73 de exactitud para un punto de corte determinado.

Se puede observar una relación entre la edad (>50 años) y el aumento de presión arterial, tanto sistólica como diastólica, lo cual está relacionado con la hipertensión, en muchos estudios esta variable es también usada como factor de riesgo para las predicciones. En este caso se puede concluir que tener un grado de hipertensión y ser mayor de 50 años, puede influir a que seas un individuo con alto riesgo. El colesterol y el índice de masa corporal también son características a considerar, ya que, si sus niveles son más altos de lo normal, también pueden influir a un riesgo mayor.

Es interesante ver cómo estas variables de alto colesterol, sobrepeso/obesidad, hipertensión y edad, están relacionadas entre sí. Una identificación temprana de estas variables de estas variables de riesgo es fundamental para poder prevenir un episodio de enfermedad cardiovascular primario.

Se consultaron varios artículos referentes al problema de sobreestimación de los métodos de cálculo de riesgo actualmente utilizados. Este efecto también puede observarse en nuestro estudio, pero tras hacer una reflexión se considera que lo ideal es tener el menor número posible de falsos negativos, ya que el propósito de este algoritmo es el de la identificación primaria. Por lo tanto, es preferible identificar un mayor número de personas que puedan ser de alto riesgo con el fin de aplicar medidas preventivas como cambio de alimentación y hábitos, lo cual podría beneficiar a cualquier individuo con o sin riesgo. Los falsos positivos serán tarde o temprano desestimados ya sea porque presentan alguna mejoría con una mayor rapidez o debido a resultados de laboratorio complementarios.

Se infiere que el mayor problema no es la sobreestimación, más bien es la selección del punto de corte para determinar que una persona es de alto riesgo o no. Los métodos de diagnóstico tienen que ser diseñados de acuerdo a la población a la cual se desean aplicar, debido a que el comportamiento de las características y la variable objetivo pueden ser sensibles a la alimentación, etnia, estado social, económico y educacional.

A pesar, del buen desempeño del algoritmo, no está listo para ser aplicado en el mundo real, la poca información explicativa sobre el dataset utilizado referente a la población de la cual fue extraído y la descripción sobre la naturaleza de las pruebas diagnósticas, hace que este estudio sea de poca confianza. Pero, sirve como precedente para saber si la metodología utilizada es útil y para saber qué tipo de recursos son necesarios para realizar una investigación más elaborada enfocada a la población mexicana.

Por último, se puede mencionar que los algoritmos de machine learning son útiles para este tipo de problemas de diagnóstico médico. Debido a la dificultad para hacer una réplica del método de framingham y SCORE, no se pudo comparar nuestras métricas contra la de estos métodos. Aun así, basándonos en el estudio realizado por Yuksel et al. (2016) el *AUC* obtenido en nuestro modelo es muy cercano al obtenido con estos dos métodos, y en el caso del estudio Weng et al. (2016) nuestro *AUC* es mayor al reportado por ellos usando *Random Forest*.

Cómo siguientes pasos, es indispensable trabajar con un dataset enfocado en la población mexicana, con él se podría hacer un análisis más profundo respecto a cómo las variables de alimentación, sociales, económicas y educacionales afectan a las predicciones. Además de que se podría crear un algoritmo de cálculo de factor de riesgo, enfocado a esta población. Se desea trabajar más con el algoritmo creado, para mejorar su desempeño de diferenciación entre verdaderos positivos y negativos aplicando *feature selection* y otros algoritmos de machine learning.

1. **Bibliografía**

1.- Mohammad, B., Mostafa G. y Sahar B. (2016). Cardiovascular diseases in the mirror of science. *J Cardiovasc Thorac Res*, *8*(4), 158-163.

2. - Cerecero, P., Hernández, B., Aguirre, D., Valdés, R y Huitrón G. (2009). Estilos de vida asociados al riesgo cardiovascular global en trabajadores universitarios del Estado de México. *Salud pública de México*, *51*(6), 465-473.

3.- Juárez, C., Cabrera, C., Ramírez, S., García, A., Morales, L y Ramírez, Heidi. (2018). Factores de riesgo para enfermedad cardiovascular en adultos mexicanos. *Revista Médica MD*, *9*(2), 152-162.

4.- Rosas, M., Borray, G., Ramirez, Erick., Jiménez, G., Hernández, M., Barraza, R., Betacourt L., Camacho, R., Parra R., Martinez, H y Arriaga J. (2018). Cardiovascular risk reduction: Past,present and future in Mexico.*Ann Clin Hypertens*, *2*, 038-047.

5.-Córdova, J., Barriguete, J., Lara, A., Rosas, M., Hernández, M., de León, M., Aguilar, C. (2008).Chronic non-communicable diseases in Mexico: epidemiologic synopsis and integral prevention. *Salud pública Méx, 50*(5), 419-27.

6.- Mendoza, K., Pedroza, A., Alcaraz, C., Ávila, L., Aguilar, C., Barquera, S. (2019). Attributable Burden and Expenditure of Cardiovascular Diseases and Associated Risk Factors in Mexico and other Selected Mega-Countries. *Int J Environ Res Public Health, 16*(20), 4041.

7.- Ramón, C., Priego, H., Carballo, A. (2018). Análisis del componente marketing mix del programa “checate, midete, muevete” en comunidades de tenosique, tabasco, méxico. *Revista de Administração da Unimep, 16*(1).

8.- Salazar, A., Martinez, B., Mundo, V, Gómez, I., Uribe, R. (2018). Conocimiento y nivel de comprensión de la campaña Chécate, Mídete, Muévete en adultos mexicanos*. Salud pública Méx, 60*(3).

9.-Sneha, N., Gandil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big Data*, *6(13).*

10.- Fawagreh, K., Medhat, M., Elyan, E. (2013) Random forest: from early developments to recent advancements. *Systems Science & Control Engineering*, 2*(1)*, 602-609.