INSTITUTO TECNOLÓGICO DE BUENOS AIRES – ITBA

ESCUELA DE POSTGRADO

CATEGORIZACION y ANALISIS de la FREQUENCIA CARDIACA de UN INDIVIDUO con INTELIGENCIA ARTIFICIAL

AUTOR/ES: Goldman, Jorge Carlos (Leg. Nro: 44113)

DOCENTE/S TITULAR/ES O TUTOR/ES: Riccillo, Marcela

TRABAJO FINAL PRESENTADO PARA LA OBTENCIÓN DEL TÍTULO DE ESPECIALISTA EN CIENCIAS DE DATOS

BUENOS AIRES

PRIMER CUATRIMESTRE, 2020

Abstract

Este estudio presenta un enfoque novedoso en la aplicación de técnicas de aprendizaje automático para la clasificación de enfermedades del músculo cardíaco. Una detección temprana de arritmias aumenta considerablemente la posibilidad de corrección y sobrevida de los pacientes mediante medicación adecuada indicada por un profesional de la salud. En el siguiente trabajo se evaluarán diversos algoritmos de aprendizaje automático con técnicas de selección de variables, a fin de lograr una clasificación, con cierto grado de exactitud, de diversas enfermedades del músculo cardíaco, basándonos en las mediciones obtenidas mediante dispositivos electrónicos. Los resultados experimentales mostraron que a través del algoritmo de Random Forest, se logra la clasificación de una persona enferma de una sana con casi 94% de exactitud, con selección de las variables más significativas mediante el algoritmo de RFE.

Contenido

CAPÍTULO 1 - Introducción

1.1. Introducción

El siguiente estudio tiene como objetivo evaluar la posibilidad de, mediante algoritmos de aprendizaje automático, ayudar a médicos y especialistas en cardiopatías, a detectar a través de las mediciones recogidas por medio de dispositivos electrónicos, la existencia de afecciones al músculo cardíaco de los pacientes. Bajo ningún concepto se intenta reemplazar al profesional de la salud, sino proveer de una herramienta simple en la cual pueda confiar, con cierto grado de exactitud, para así obtener un diagnóstico rápido y efectivo. Para poder determinar si un individuo se encuentra sano o enfermo, y posteriormente poder proveer un diagnóstico preliminar.

Se utilizarán diversos algoritmos de aprendizaje supervisado tanto para el análisis como para la selección de variables más significativas y para finalmente la predicción de la enfermedad cardiaca.

Las últimas décadas se han caracterizado por cambios dramáticos en la forma cómo la utilización de algoritmos de inteligencia artificial aplicado a la medicina ha influenciado la detección temprana y el descubrimiento de nuevas enfermedades. Durante mucho tiempo, la capacidad tecnológica de la industria médica no permitía a los investigadores y médicos hacer uso efectivo de esta clase de algoritmos, debido a la carencia de información o difícil manipulación de la misma.

Uno de los primeros campos en adoptar este tipo de tecnologías para mejorar sus diagnósticos fue la cardiología, debido a que los electrocardiogramas son métodos baratos, sencillos y no invasivos de obtener información crítica para el diagnóstico de enfermedades del músculo cardíaco.

Hoy en día, se cuenta con equipos de medición de alta exactitud portátiles [Figura 1.1], lo que permiten al médico monitorear a un individuo durante un período de tiempo prolongado. El paciente que es sometido a este estudio tiene indicado continuar con su vida normal, dieta y medicamentos, así como está contraindicado realizar ejercicios.

Estos dispositivos son comúnmente conocidos como Holters, los cuales registran en diversos canales de información distintas mediciones sobre el estado del corazón durante un período definido de tiempo.

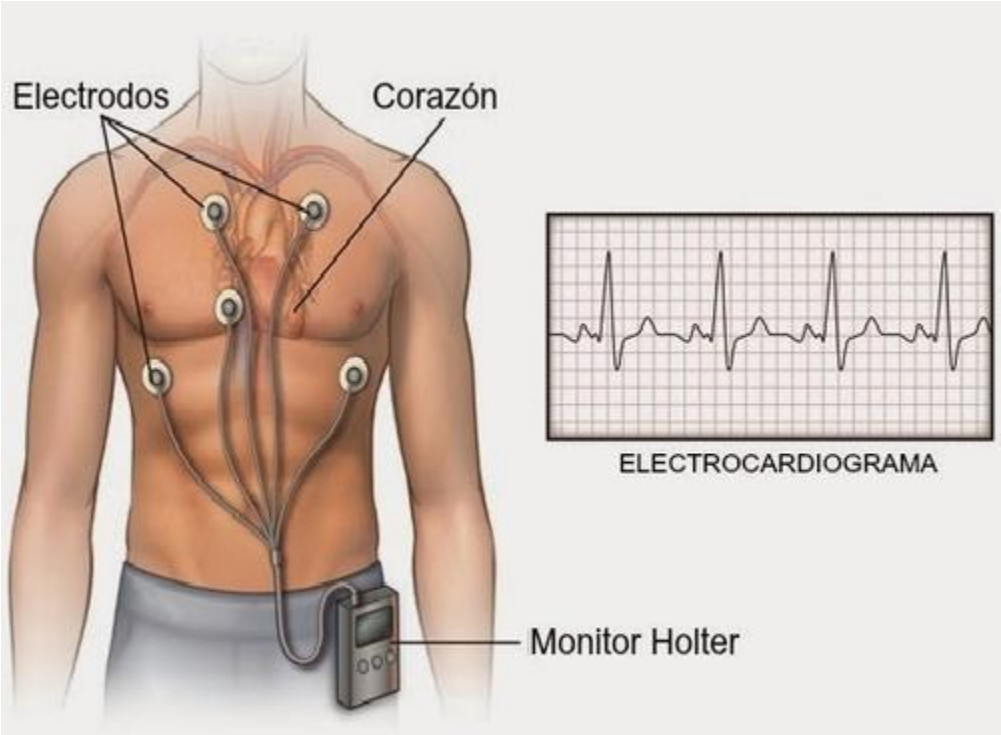

Figura 1.1. Holter colocado en un individuo de estudio. [1]

1.2. Antecedentes

Las primeras investigaciones en este campo fueron realizadas a finales de los años 90 en Suecia donde desarrollaron un algoritmo conocido como VF15. El mismo, era un algoritmo de aprendizaje supervisado el cual intentaría por primera vez categorizar ciertas enfermedades coronarias basándose en lo aprendido de un conjunto de datos de entrenamiento de electrocardiogramas e información adicional de los pacientes como sexo, edad, peso, altura, etc. Este obtuvo mejor performance contra algoritmos más simples como Naive Bayes o Nearest Neighbors. VF15 obtuvo una exactitud de 62% en dicho momento. [H. Altay Guvenir, B. A. (1997)]

[1] https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/holter-monitor

Casi simultáneamente el Dr M. Kukar en su publicación en 1999 en la revista de Inteligencia Artificial aplicada a la medicina, demostró que mediante la utilización de algoritmos de aprendizaje tanto supervisado como no supervisado, utilizando los algoritmos de Naive Bayes y KNN, podía alcanzar exactitudes en el diagnóstico de enfermedades isquémicas coronarias de casi un 75%. [M. Kukar, I. K. (1999)]

Luego de casi una década de incontables trabajos en el área, hacia mediados de 2013, la Dra. M Mitra y la Dra. R. K. Samanta, proponen la utilización del algoritmo Levenberg-Marquardt (LM), el cual consiste en la utilización de ajustes de curvas de mínimos cuadrados de manera iterativa, para la clasificación de arritmias coronarias. Para dicho estudio, debieron realizar una selección de variables más significativas mediante el algoritmo CFS, el cual se basa en correlación de N variables para lograr dicho objetivo. Obteniendo como resultado un algoritmo clasificador de casi un 80% de exactitud con cerca de un 87% de sensibilidad. [M. Mitra, R.K. Samanta. (2013)]

1.3. Definición Del Problema

El monitoreo ambulatorio de presión arterial y ritmo cardíaco es un método no invasivo de obtener información sobre el estado cardiológico de un individuo durante un cierto período de tiempo, generalmente 24 horas.

Se trata de una técnica muy efectiva y ofrece una visión global del estado de un individuo, sin embargo, carece de información sobre la actividad que se encuentra haciendo el mismo en los momentos de las mediciones y requiere personal altamente entrenado para procesar los datos capturados por el dispositivo.

Se define como cardiopatía a una familia de enfermedades del músculo cardíaco. Estas representan la principal causa de muerte registrada en Estados Unidos, tanto en mujeres como hombres.[2]

Al ser una familia de enfermedades, cada cual, con su caracterización específica, las cuales requieren de entrenamiento médico a fin de poder discernir una de otras, en este estudio se prefirió seleccionar las más comunes y frecuentes:

- Enfermedad Arterial Coronaria: Esta enfermedad se produce cuando las arterias y demás vasos que proveen al músculo cardíaco de nutrientes y oxígeno se ven afectadas por placas o coágulos. Desencadenando en síntomas como anginas de pecho o falta de aliento.
- Infarto: Se produce cuando el bloqueo en las arterias y vasos es completo, causando muerte de tejido cardíaco.
- Alteración del ritmo cardíaco (Taquicardia o Bradicardia): Se caracterizan por una frecuencia cardiaca menor o mayor de lo habitual y se ocasionan por fallos en la formación del impulso eléctrico o en la conducción del mismo. Pueden ser asintomáticas. Y muchas veces el stress es un factor importante en estos casos.
- Bloqueos: Se producen cuando el estímulo eléctrico no se conduce adecuadamente desde las aurículas a los ventrículos
- Otros: Incluyendo Fibrilaciones e Hipertrofias del músculo cardíaco

[2] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

En los conjuntos de datos reales como los utilizados para este estudio, la cantidad de instancias consideradas como “anormales” tiende a ser baja en comparación con las instancias consideradas “normales”. En este caso, la cantidad de información sobre los individuos que presentan anomalías en su ritmo cardíaco, es considerablemente menor en comparación con las de personas sanas

Los conjuntos de datos desbalanceados representan un problema para la mayoría de los algoritmos de aprendizaje automático.

1.4. Justificación Del Estudio

Esta investigación tiene como objetivo evaluar la posibilidad de categorizar previo al diagnóstico médico, las anomalías y enfermedades en las mediciones realizadas que no corresponden a una persona sana, a fin de mejorar el diagnóstico final.

1.5. Limitaciones De La Investigación

Este estudio solo se va a centrar en la detección de anomalías y enfermedades cardiacas características mencionadas previamente, mediante el uso inteligencia artificial, procesando las señales y parámetros obtenidos a través de las señales electrocardiográficas y datos del paciente.

1.6. Alcance De La Investigación

El personal médico especializado en cardiología contará con información previa al diagnóstico sobre el estado de salud del paciente y si hubo una anomalía en su electrocardiograma.

1.7. Hipótesis

A través de algoritmos de Inteligencia Artificial, se puede inferir con una exactitud superior al 80% de certeza la detección de anomalías en el ritmo cardíaco correspondientes con cardiopatías conocidas.

1.8. Objetivo General

Utilizar diferentes técnicas de aprendizaje automático para predecir el estado del músculo cardíaco (Sano o Enfermo) de un individuo al momento de realizar una medición de la presión arterial y ritmo cardíaco mediante un estudio de Holter. En el caso de que el músculo cardíaco reporte alguna anomalía durante el estudio, incluso se le podría dar un indicio al médico de ante que cuadro o que familia de cardiopatía se estaría enfrentando.

1.9. Objetivos Específicos

1. Evaluar diferentes técnicas y algoritmos de aprendizaje automático a fin de determinar el mejor predictor de cardiopatía existente del individuo.
2. Seleccionar la técnica más adecuada de aprendizaje automático y sus parámetros.

CAPÍTULO 2 - Técnicas

Los algoritmos descritos a continuación serán los utilizados para la predicción de actividades basadas en validación cruzada.

2.1. Árboles de Decisión

Los Árboles de Decisión [Hastie, T., Tibshirani, R. ISLR] son un mecanismo basado en aprendizaje automático el cual intenta por sus propios medios identificar las variables más significativas de un determinado conjunto de sujetos de estudio.

Estos son fáciles de comprender y modificar a fin de dar un modelo de decisión que se adapte a las necesidades del negocio o estudio.

Estos algoritmos se adaptan bien a grandes números de variables a fin de proporcionar una estructura binaria arbórea como resultado con reglas fáciles y sencillas para predecir la categorización de futuros individuos.

2.2. Random Forest

Basada en la técnica descrita previamente, los Bosques Aleatorios o Random Forest [Hastie, T., Tibshirani, R. ISLR], son un mecanismo de aprendizaje automático supervisado en el cual se crean N árboles de Decisión independientes y se utilizan en conjunto a fin de predecir una clasificación de individuos nuevos de manera más certera.

A continuación, se describen las características de esta técnica:

Pros:

- Es un mecanismo con mejores probabilidades de clasificación que los árboles de decisión.
- Puede manejar de manera eficiente grandes números de variables de entrada sin excluir ninguna
- Proporciona un análisis sobre las variables más relevantes a la clasificación y da un estimativo de probabilidad en su clasificación

Contras:

- Categorizaciones dudosas o ruidosas pueden hacer que el modelo falla al predecir correctamente nuevos individuos
- La clasificación hecha por random forest es difícil de interpretar por el humano.

2.3. Redes Neuronales

Las redes neuronales artificiales [Hastie, T., Tibshirani, R. ESLR] son un conjunto de algoritmos basados en la naturaleza de la neurona humana, las cuales buscan que un conjunto de ellas, interconectadas de manera específica y con funciones de activación, disparen resultados similares ante estímulos similares aprendidos previamente.

Son algoritmos pertenecientes a la categoría de aprendizaje automático supervisado, en el cuál mediante un conjunto de entrenamiento, logra que la "red" aprenda las características que definen a cada individuo a partir de sus variables de entrada.

La unidad básica que conforma la "red" es conocida como neurona, el cual se conecta con las demás neuronas de la red a fin de crear conexiones ponderadas.

Uno de los principales inconvenientes de las redes neuronales es el sobre-entrenamiento, en el cual su capacidad de predicción basada en el conjunto de entrenamiento es mayor que su capacidad de generalización de nuevos individuos no pertenecientes al mismo.

2.4. Support Vector Machines (SVM)

Esta familia de algoritmos conocida como Support Vector Machines [Hastie, T., Tibshirani, R. ESLR] funciona de manera muy diferente a las descritas previamente en este trabajo. Si bien forma parte de los algoritmos de aprendizaje supervisado, el objetivo de este es construir un hiperplano o conjunto de estos de un orden superior, al de la cantidad de variables que conforman a cada individuo del conjunto de entrenamiento, para dividir el espacio muestral en N categorías previamente asignadas.

Su funcionamiento interno está basado en Kernels (funciones Núcleo) las cuales utiliza a fin de formar los hiperplanos descritos previamente.

Esta familia de algoritmos se caracteriza por ser extremadamente robusta en comparación con las redes neuronales una vez entrenado, además de haber demostrado mejores resultados a través del tiempo en casos complejos de clasificación.

2.5. Cross-Validation

La validación cruzada o Cross-Validation [Hastie, T., Tibshirani, R. ISLR], es una técnica utilizada comúnmente para evaluar los resultados y el error de prueba estadísticos con el fin de garantizar que los mismos son independientes de la partición realizada para entrenamiento y prueba.

Dicha técnica consiste en realizar una serie de K particiones sobre el conjunto de datos y alternar los mismos entre los utilizados para entrenamiento de los modelos y la validación de los mismos.

A fin de dar una mejor explicación se provee el siguiente ejemplo:

Suponiendo una población de 100 individuos distribuidos uniformemente, se realizan un Cross-Validation de 10 particiones, comúnmente conocidas como folds.

Por lo que, en la primera ronda, se utilizaran los 9 primeros folds para entrenar el modelo y el último para validación de este.

En la segunda iteración, se utilizarán los 8 primeros folds y el último para entrenamiento, dejando el número 9 para prueba.

Así sucesivamente, hasta que el primer fold sea el de prueba y los 9 restantes de entrenamiento.

Una vez recolectados la exactitud y el error de los 10 modelos, se tomará como válido el promedio de estos.

2.6. Undersampling

La técnica, como su nombre lo dice, consiste en balancear clases desbalanceadas removiendo individuos al azar uniformemente en la clase mayoritaria hasta alcanzar la misma cantidad de instancias.

CAPÍTULO 3 - Resultados

3.1. Herramientas

Se utilizará el lenguaje R[3] para procesar los datos obtenidos del conjunto de Arritmias del Instituto de Ciencias de la Computación de la Universidad de California, Irvine (https://archive.ics.uci.edu/ml/datasets/arrhythmia).

3.2. Análisis Exploratorio de los Datos

Se comenzó por hacer un análisis exploratorio de los datos disponibles, los cuales cuentan con 451 observaciones de 279 variables correspondientes a pacientes evaluados.

[3] https://www.r-project.org/

Variables dependientes para la detección de la cardiopatías o irregularidades cardiacas:

- Categorización del estado del paciente (Sano o Enfermo)
- Categorización de la cardiopatía presentada por el paciente en caso de estar enfermo (Enfermedad Arterial Coronaria, Infarto, Alteración del ritmo cardíaco, Bloqueo de la rama derecha, Otros)

Variable independiente para la detección de la cardiopatías o irregularidades cardiacas:

- Edad
- Sexo
- Altura
- Peso
- Mediciones de señales electrocardiográficas simultáneas en 274 canales.

Primero se observó, que si bien el conjunto de datos se encuentra altamente desbalanceado respecto de cada cardiopatía específica respecto de la clase mayoritaria que es la de individuos sanos, este no se encuentra significativamente desbalanceado respecto de si los individuos categorizados entre sanos o enfermos cómo se puede observar en la *Figura 3.1*.

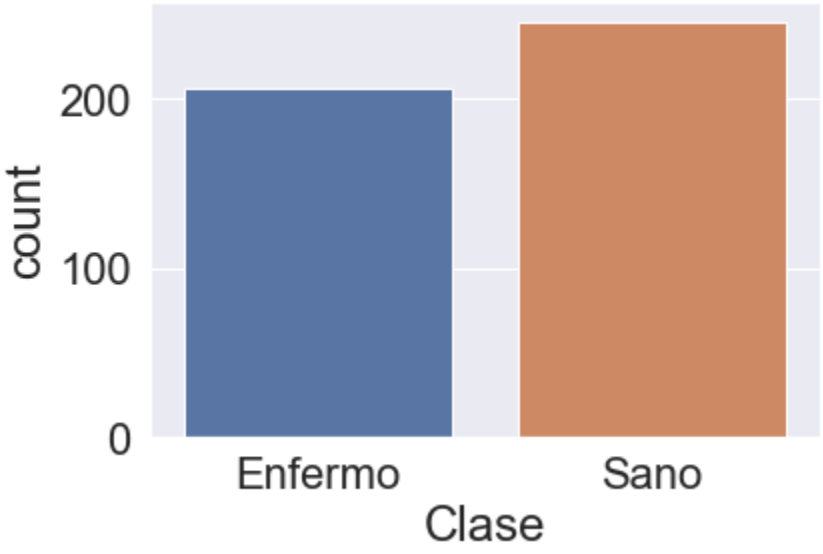

Figura 3.1. Cantidad de objetos de estudio.

Luego se procedió a evaluar las demás variables que resultaron relevantes a los individuos de estudio, como por ejemplo el rango etario, peso, altura, sexo y ritmo cardíaco.

La *Figura 3.2*, provee un histograma de edades, donde la media se encuentra en 46.41 años con un desvío estándar de 16.4 años.

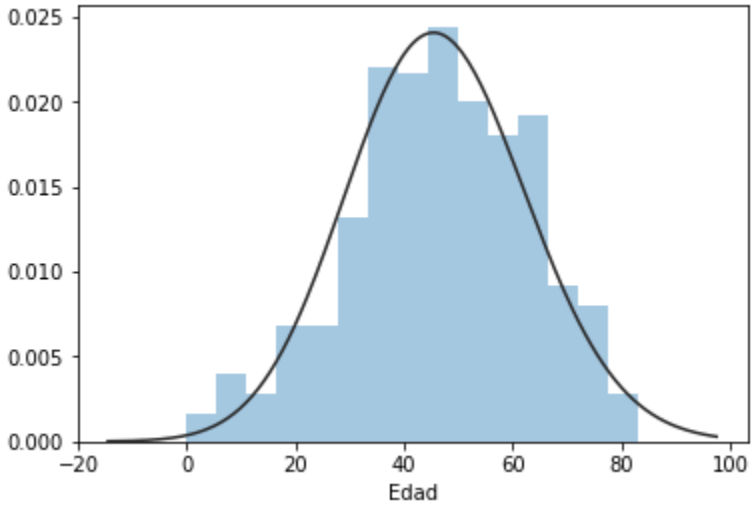

Figura 3.2. Histograma Edades Individuos de estudio.

La *Figura 3.3.* provee un histograma de la altura de los pacientes, la cual cuenta con una media de 166.1cms y un desvío estándar de 37.19 Cms.

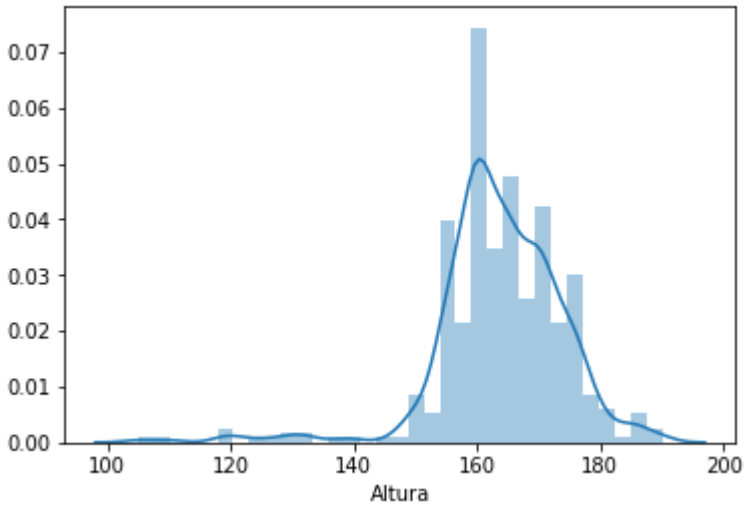

Figura 3.3. Histograma Altura Individuos de estudio.

La *Figura 3.4.* nos permite observar la cantidad de individuos del sexo femenino como la de masculino, siendo 202 y 249 respectivamente.

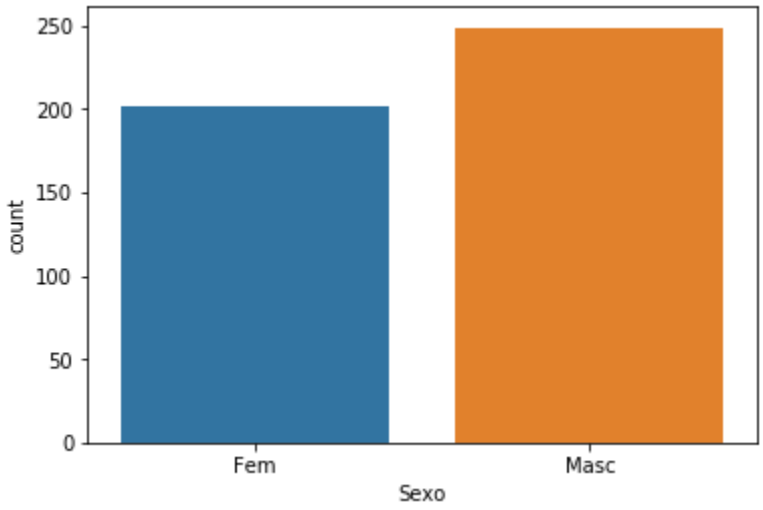

Figura 3.4. Histograma Sexo Individuos de estudio.

La *Figura 3.5.* provee un histograma del peso de los individuos de estudio, con una media de 68.14 Kg y un desvío estándar de 16.6 Kg.

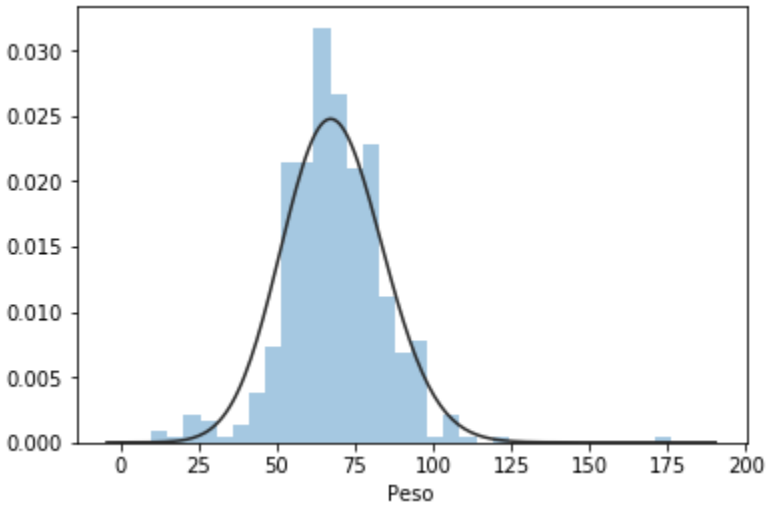

Figura 3.5. Histograma Peso Individuos de estudio.

Por último, la *Figura 3.6.* provee un histograma del ritmo cardíaco de los pacientes estudiados, con una media de 74.46 pulsaciones por minuto y un desvío estándar de 13.87.

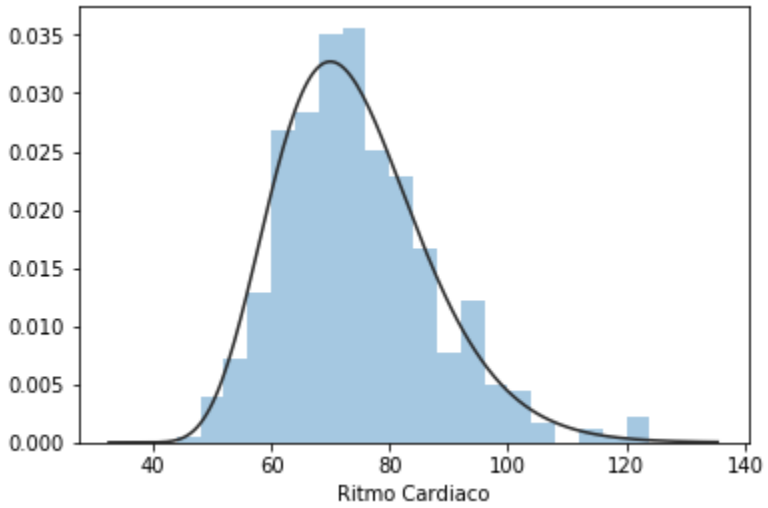

Figura 3.6. Histograma Ritmo Cardíaco Individuos de estudio.

3.3. Resultados

Utilizando las técnicas mencionadas en la sección de Técnicas, se procedió a evaluar la exactitud de varios algoritmos de clasificación con el conjunto de datos original. A continuación, se detallan los resultados obtenidos en base a la experimentación realizada.

3.3.1. Análisis de la cantidad de particiones sobre la Validación Cruzada

Siguiendo lo que se conoce como la técnica del codo, se va a utilizar 10 cómo el número de particiones para Cross-Validation a partir de este punto en adelante. La cual consiste en utilizar el punto de inflexión de la *Figura 3.7* a partir del cual el error no presenta una varianza significativa respecto de mediciones anteriores.

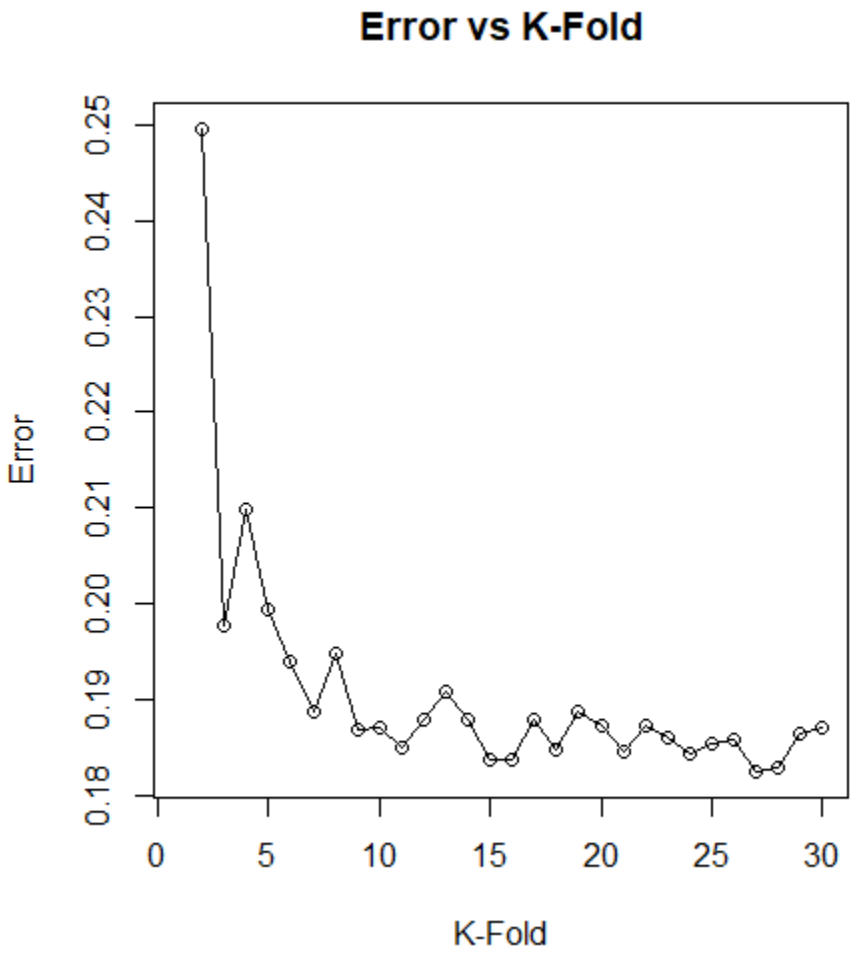

Figura 3.7. Error vs Cantidad de folds realizados

3.3.2. Análisis de la cantidad de árboles en el algoritmo de Random Forest

Primeramente, se procedió a evaluar la cantidad de árboles en el algoritmo de Random Forest mínima a partir de la cual la varianza del error no era significativa. Cómo se puede observar en la *Figura 3.8*, se decidió utilizar 250 árboles cómo configuración básica para el algoritmo mencionado.

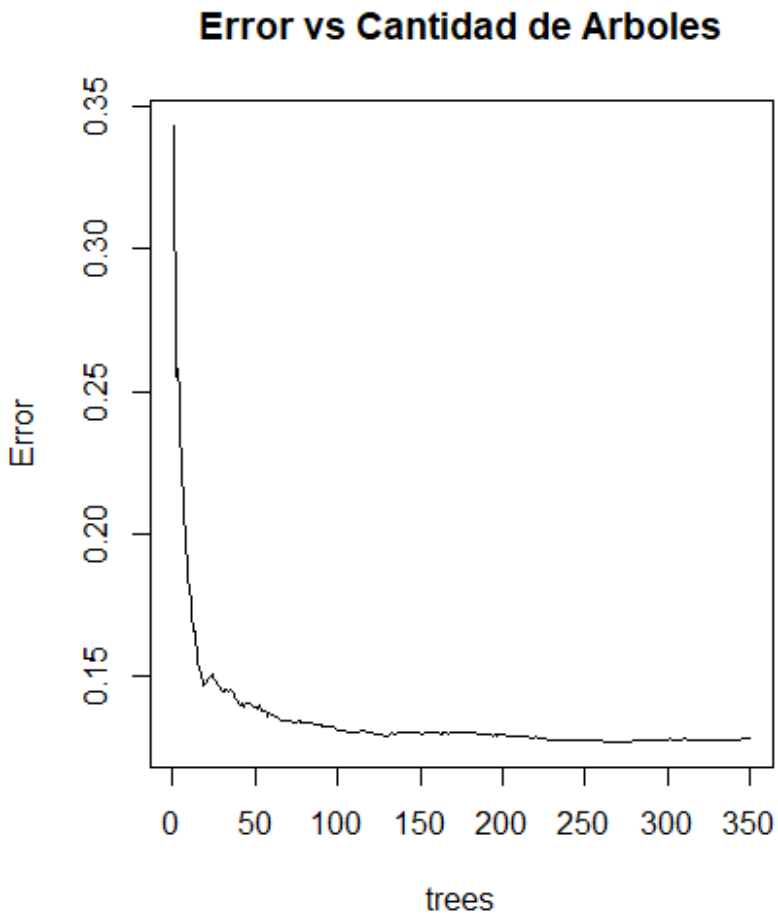

Figura 3.8. Error en Random Forest vs Cantidad de Árboles

3.3.3. Análisis en la determinación del número de neuronas a utilizar para el algoritmo de Redes Neuronales

Para determinar el número de neuronas necesarias en la capa oculta. En este caso se decidió utilizar 35 neuronas en la capa oculta a fin de tener un balance aceptable entre velocidad de entrenamiento y exactitud, la cual también fue verificada a través de Cross-Validation. Ya que el aumento de la cantidad de neuronas en la capa oculta conlleva a 2 problemas, la matriz de pesos interna se multiplica de manera geométrica consumiendo así más memoria y a su vez más tiempo de cálculo de operaciones

3.3.4. Clasificación Multiclase

Luego de Aplicar las Técnicas de Aprendizaje Automático descriptas en el capítulo 2:

- Arboles de Decisión (A. de D.)
- Random Forest
- Redes Neuronales
- SVM con Kernel Polinomico (SVM poly)
- SVM con Kernel Radial (SVM radial)

Se puede observar en la *Tabla 3.1* y en la *Figura 3.9,* el algoritmo de Random Forest fue capaz de realizar una clasificación multiclase con una exactitud del 86%.

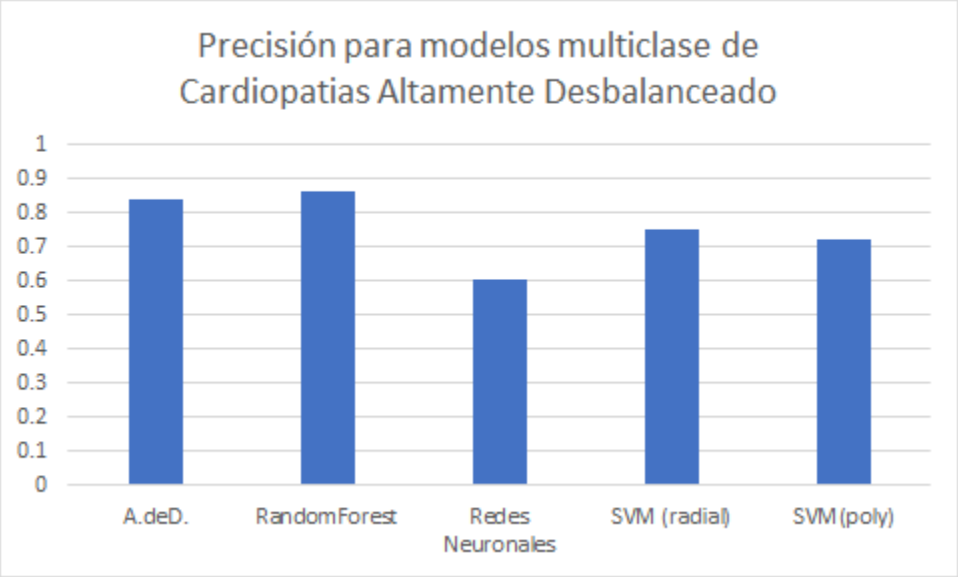

Figura 3.9. Exactitud de Algoritmos de Clasificación con un conjunto de datos altamente desbalanceado

	A.de D.	RandomForest	Redes Neuronales	SVM (radial)	SVM(poly)
Exactitud	0.8426	0.8611	0.6019	0.75	0.7222

Tabla 3.1. Todas las Clases de Cardiopatías. (Altamente Desbalanceado)

3.3.5. Rebalanceo de clases

No obstante, al contar con multiclases altamente desbalanceadas, con una clase mayoritaria de personas sanas, pudimos observar mediante matrices de confusión, que el algoritmo solamente era efectivo en la categorización de dicha clase mayoritaria y no en las cardiopatías o personas enfermas.

También, se pudo observar que, debido a la distribución de los datos, cuando se realizan la separación de datos entre conjuntos de entrenamiento y de evaluación, el último contiene muy

pocos individuos con cada cardiopatía. Lo cual condujo a que estos modelos puedan ser muy efectivos en detectar la clase de persona sana, pero muy pobres al detectar las clases minoritarias.

3.3.5.1. Aplicación de la Técnica de Undersampling

A fin de solucionar el problema mencionado anteriormente, se procedió a aplicar la técnica de Undersampling sobre la clase mayoritaria, a fin de reducir el número de observaciones de la misma y que quede en similar proporción con las otras clases en el conjunto de datos. En el cual se dejó alrededor de entre 30 y 50 individuos de cada clase.

Cómo se puede observar en la *Figura 3.10* y la *Tabla 3.2*, la exactitud de los algoritmos se ve seriamente afectada, perdiendo aproximadamente un 12% y en algunos casos hasta el 20%.

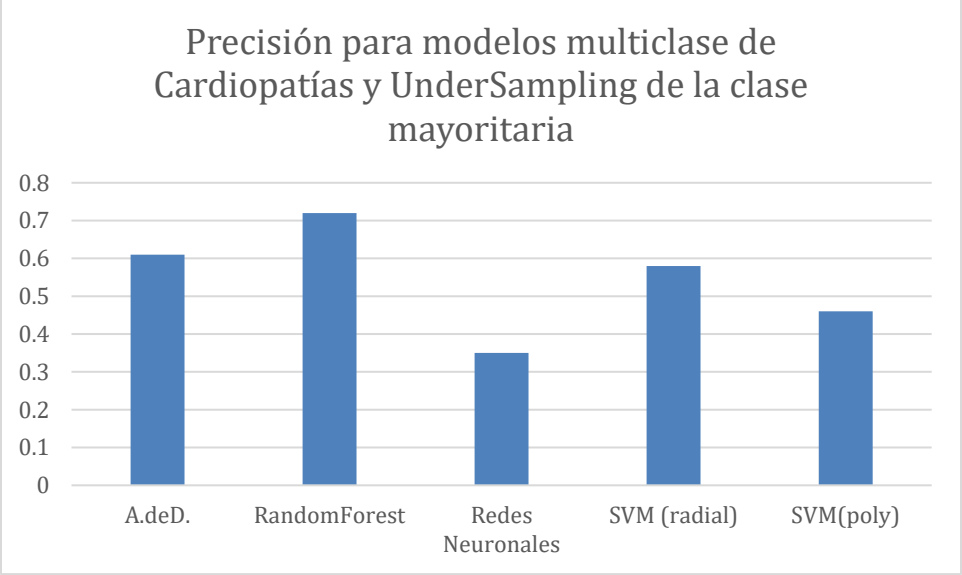

Figura 3.10. exactitud de Algoritmos de Clasificación con un conjunto de datos balanceado

	A.de D.	RandomForest	Redes Neuronales	SVM (radial)	SVM(poly)
Exactitud	0.61	0.72	0.35	0.58	0.46

Tabla 3.2. Todas las Clases de Cardiopatías y Underbalance de la clase mayoritaria

3.3.6. Recategorización en clases Binarias

A partir de los resultados obtenidos previamente, y en base a recomendaciones de trabajos anteriores mencionados en la sección de Antecedentes, se optó por encauzar este estudio en la detección de personas enfermas o sanas y delegar en especialistas en el área de la salud la clasificación de las cardiopatías existentes.

Para llevar adelante lo expuesto previamente, se transformó las clases minoritarias en una sola clase que abarcara a todas las cardiopatías.

Con el objetivo de calcular la Sensibilidad y Especificidad de los algoritmos, el 10% del conjunto de datos originales va a ser separado del resto y conservado para examinar el comportamiento final de los mismos. El 90% restante va a ser dividido en conjuntos de entrenamiento y evaluación.

Ya que el conjunto de datos restantes se encuentra ligeramente desbalanceado, se aplicaron los mismos algoritmos utilizados anteriormente, a fin de evaluar su performance para predecir si una persona se encuentra enferma o sana.

La *Figura 3.11* y la *Tabla 3.3* muestran los resultados de la aplicación de los algoritmos utilizados previamente y de la recategorización en clases binarias (Enfermo o Sano).

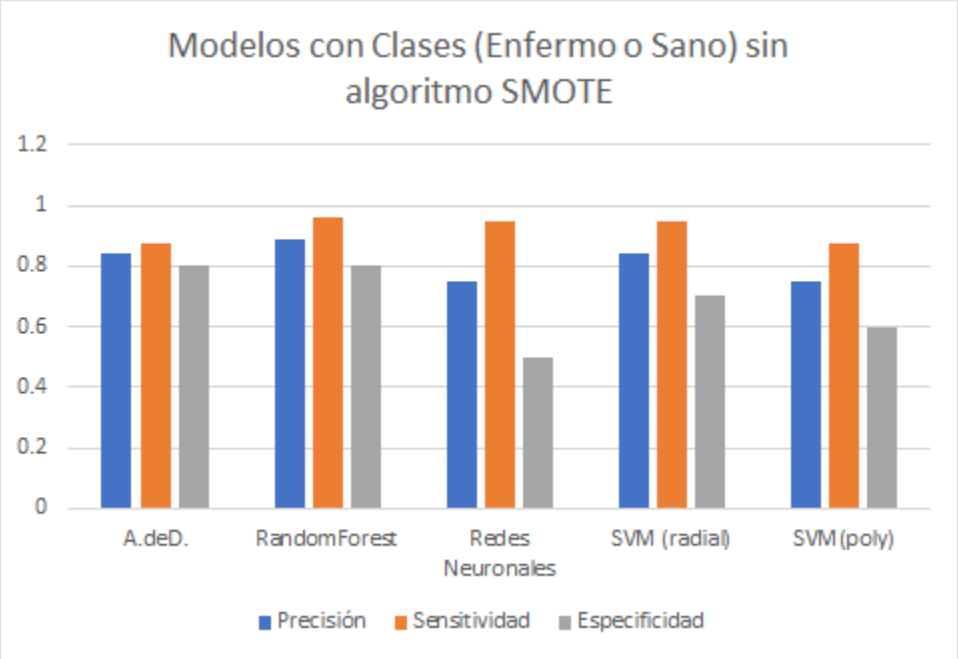

Figura 3.11. exactitud, Especificidad y Sensitividad en clases binarias (Enfermo o Sano)

	A.de D.	RandomForest	Redes Neuronales	SVM (radial)	SVM(poly)
Exactitud	0.8409	0.8864	0.755	0.84	0.75
Sensitividad	0.875	0.9583	0.95	0.956	0.875
Especificidad	0.8	0.8	0.5	0.7	0.6

Tabla 3.3. 2 Clases (Enfermo o Sano) sin algoritmo SMOTE

3.3.7. Optimización de clases Binarias

Con el objetivo de mejorar la exactitud de los algoritmos utilizados previamente, se procedió a aplicar diversos algoritmos de optimización del conjunto de datos.

3.3.7.1. Synthetic Minority Over-sampling Technique

Con el objetivo de mejorar la exactitud de los algoritmos utilizados previamente, se utilizó el algoritmo SMOTE [Bonaccorso, G.], el cual genera a partir de la clase minoritaria, nuevas observaciones con el objetivo de balancear las clases existentes.

Se conoce como algoritmo SMOTE a la técnica aplicada sobre un conjunto de datos desbalanceado a fin de generar individuos sintéticos basados en la información existente de la clase minoritaria.

A fin de poder generar los individuos ficticios, este algoritmo necesita reducir el problema a clases binarias.

Una vez que están definidos las categorías binarias, se busca mediante técnicas matemáticas definir “bordes” suaves que van a separar a las categorías en todos sus espacios dimensionales para todas sus variables.

A partir de los individuos de la clase minoritaria o anormal, se generan artificialmente nuevos individuos que caigan dentro de los “bordes” de las variables generados en el paso anterior hasta lograr la misma cantidad de individuos en ambas categorías.

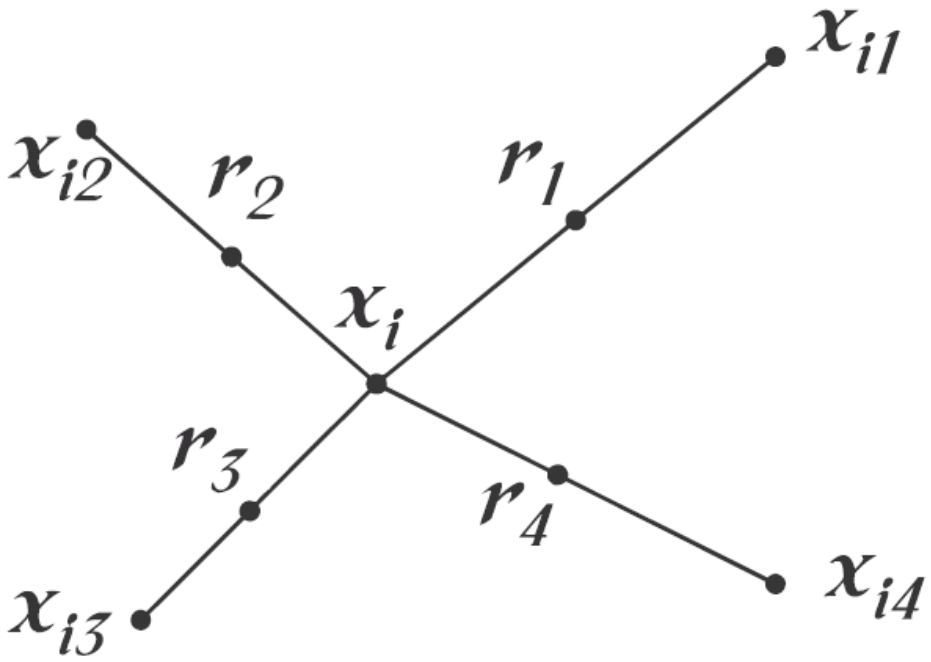

Figura 3.12. Nuevos Puntos Generados mediante SMOTE

A pesar de haber balanceado las clases existentes, podemos observar en la *Tabla 3.4* y *Figura 3.13*, que la utilización del algoritmo SMOTE, solo genera un peor desempeño de los modelos existentes, al crear observaciones “parecidas” a partir de las existentes en la clase minoritaria.

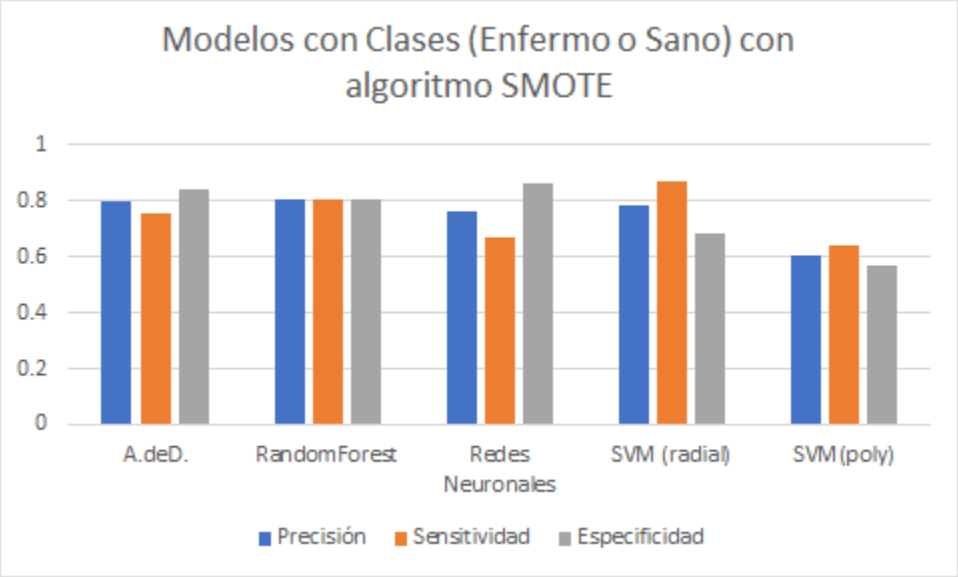

Figura 3.13. exactitud, Especificidad y Sensitividad en clases binarias (Enfermo o Sano) con algoritmo SMOTE

	A. de D.	RandomForest	Redes Neuronales	SVM (radial)	SVM(poly)
Exactitud	0.7946	0.8036	0.7589	0.7857	0.6071
Sensitividad	0.7541	0.8033	0.6721	0.8689	0.6393
Especificidad	0.8431	0.8039	0.8627	0.6863	0.5686

Tabla 3.4. 2 clases (Enfermo o Sano) con algoritmo SMOTE

3.3.8. Selección de Variables más significativas

3.3.8.1. RFE (Recursive Feature Elimination)

El método de eliminación de variables recursivamente consiste en utilizar un modelo (SVM, Random Forest o Redes Neuronales) para intentar determinar cuáles son las variables más significativas o que aportan mayor varianza al mismo con el objetivo final de reducir la dimensionalidad.

A fin de mejorar la exactitud en la predicción de si un individuo se encuentra sano o enfermo, se utilizó el algoritmo de RFE [Fontaine, A.] el cuál utiliza combinaciones de las variables disponibles en el conjunto de datos y los algoritmos de aprendizaje supervisado a fin de determinar la mejor combinación de estas y encontrar la más significativa, maximizando así la exactitud.

Cómo observamos en la *Tabla 3.5* y *Figura 3.14*, la mejor exactitud de casi 94% con una sensibilidad para detectar personas sanas del 97% y una especificidad para detectar personas enfermas del 90%, se alcanzó mediante el algoritmo de Random Forest y la eliminación de variables poco significativas utilizando el algoritmo de RFE.

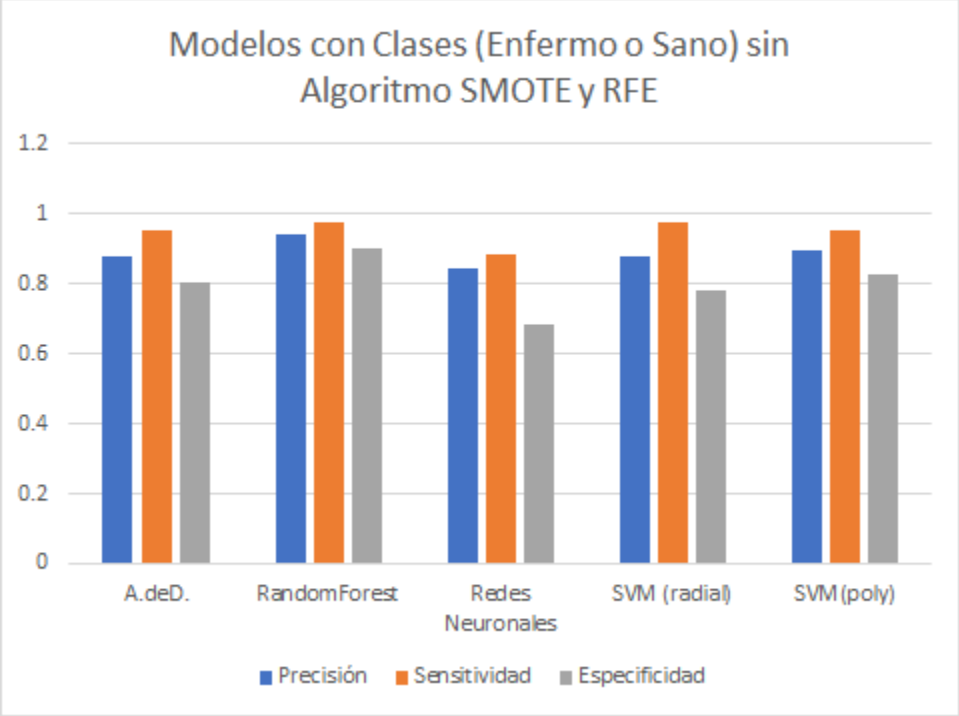

Figura 3.14. Exactitud, Especificidad y Sensibilidad en clases binarias (Enfermo o Sano) sin Algoritmo SMOTE y con RFE

	A. de D	RandomForest	Redes Neuronales	SVM (radial)	SVM(poly)
Exactitud	0.8789	0.9390	0.8415	0.8780	0.8946
Sensitividad	0.9512	0.9756	0.8852	0.9756	0.9512
Especificidad	0.8049	0.9024	0.6843	0.7805	0.8293

Tabla 3.5. 2 clases (Enfermo o Sano) sin Algoritmo SMOTE y con RFE

En la *Figura 3.15* se puede observar la cantidad y significatividad de las variables seleccionadas por dicho algoritmo.

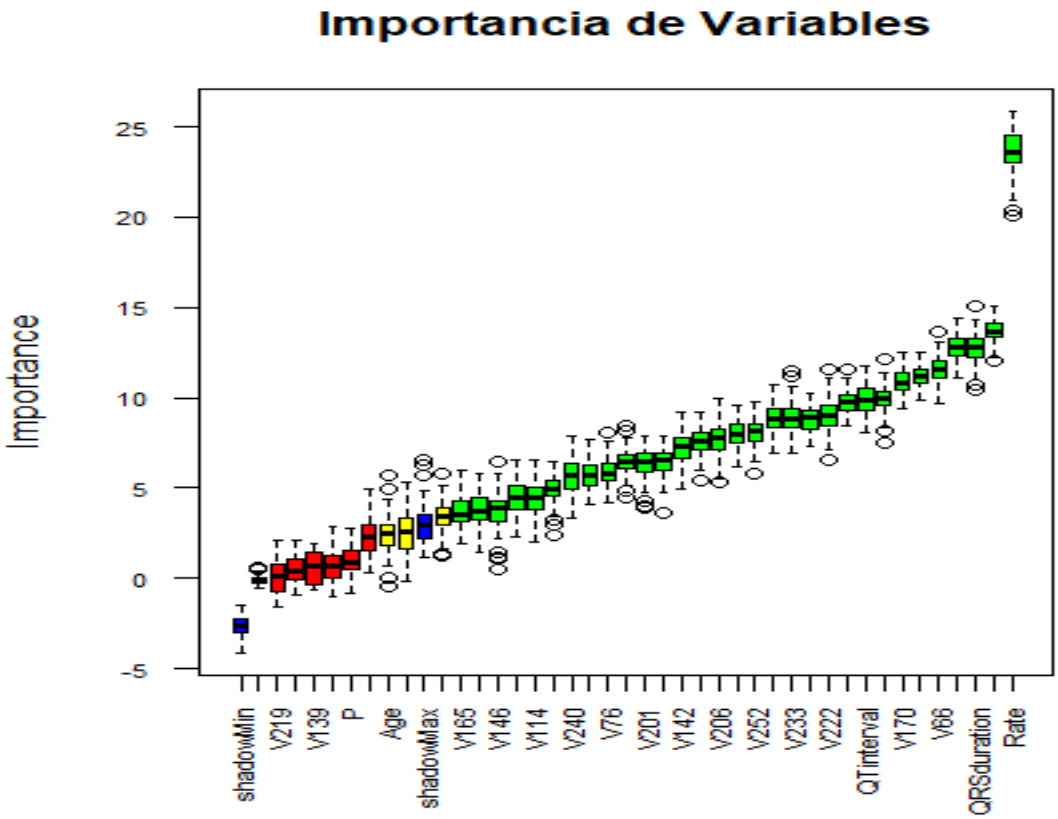

Figura 3.15. Importancia de las Variables seleccionadas por el Algoritmo RFE

CAPÍTULO 4 - Conclusiones

4.1. Conclusiones

En este trabajo se investigó la exactitud de diversos algoritmos de aprendizaje automático para la detección de afecciones cardiacas y la rápida diferenciación de una persona sana de una enferma, en base a sus parámetros corporales y mediciones realizadas por dispositivos electrónicos.

Se analizaron diversos problemas de balanceo de clases en conjuntos de datos altamente desbalanceados o donde las clases existentes no cuentan con suficientes individuos para entrenar algoritmos de manera eficiente.

Las técnicas aplicadas para este estudio fueron las de Arboles de Decisión, Random Forest, Redes Neuronales y Support Vector Machines, junto con técnicas de balanceo de datos como SMOTE y UnderSampling, para posteriormente aplicar algoritmos de selección de variables como RFE (Recursive Feature Elimination).

La utilización de algoritmos de aprendizaje automático ha demostrado ser muy eficiente en la categorización de una persona sana de una enferma, siendo la técnica de Random Forest la que mejores resultados ha dado luego de seleccionar las variables más significativas utilizando la técnica de RFE (Recursive Feature Elimination).

Por otra parte, la utilización de algoritmos de balanceo de clases ha demostrado ser inefectiva en este caso de estudio debido a la baja cantidad de individuos con los que se cuenta en ambas clases.

Random Forest por su versatilidad y adaptabilidad a tanto problemas de clasificación como de regresión ha probado ser el más simple, rápido y efectivo en categorizar con una exactitud

aceptable si una persona presenta o no una cardiopatía en base a sus parámetros físicos y mediciones tomadas por un dispositivo electrónico.

4.2. Futuras extensiones

Como futuras extensiones, se podría utilizar los giroscopios integrados en el dispositivo de Holter, a fin de determinar la actividad que se encuentra realizando la persona al momento de tomar las mediciones de la presión arterial y ritmo cardíaco, los cuales se ven afectados por la postura del individuo (ej: Hipertensión Ortostática). Se podría utilizar técnicas de aprendizaje automático, con dicha información de los giroscopios, a fin de determinar la posición y la actividad que se encuentra haciendo el individuo (Caminar, estar sentado, acostado, etc.), lo cual en la actualidad no está siendo contemplado en este tipo de estudios. Existen diversos estudios enfocados en determinar a través de inteligencia artificial la postura de un individuo basado en los giroscopios de su celular con una exactitud muy alta [Davide Anguita, A (2012).]

CAPITULO 5 – Referencias Bibliográficas

5.1. Referencias Bibliográficas

Altay Guvenir, B. A. (1997). A Supervised Machine Learning Algorithm for Arrhythmia Analysis. *Proceedings of the Computers in Cardiology Conference*. Lund, Sweden.

Davide Anguita, A. G.-O. (2012). Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine. *4th International Workshop of Ambient Assisted Living*, 216-223.

G. Guidi, M. C. (Nov 2014). A Machine Learning System to Improve Heart Failure Patient Assistance. *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1750-1756.

H. Altay Guvenir, B. A. (1997). A Supervised Machine Learning Algorithm for Arrhythmia Analysis. *Proceedings of the Computers in Cardiology Conference.* Sweden.

J Li, J. R. (October 26, 2016). Deep neural networks improve atrial fibrillation detection in Holter. *European Journal of Preventive Cardiology*.

Matja Kukar, I. K. (May 1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial Intelligence in Medicine*, Volume 16, Issue 1, Pages 25-50.

M. Mitra, R.K. Samanta. (2013). Cardiac Arrhythmia Classification Using Neural Networks with Selected Features. *Procedia Technology,Volume 10, 2013, Pages 76-84*

Resul Das, I. T. (2009). Effective diagnosis of heart disease through neural networks ensembles. *Expert Systems with Applications*, Volume 36, Issue 4, May 2009, Pages 7675-7680.

Rine Nakanishi, D. D. (March 2018). MACHINE LEARNING IN PREDICTING CORONARY HEART DISEASE AND CARDIOVASCULAR DISEASE EVENTS. *Journal of the American College of Cardiology*, Volume 71, Issue 11 Supplement.

Shawe-Taylor, N. C. (2000). An introduction to support vector machines and other kernel-base learning methods. *Cambridge University Press*.

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York: *Springer.*

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: *Springer.*

Huang, G.: Learning Capability and Storage Capacity of Two-Hidden-Layer Feedforward Networks. IEEE Trans. on Neural Networks 14(2), 274-281

Bonaccorso, G. (2018) Machine Learning Algorithms - Second Edition: *Packt Publishing*

Fontaine, A. (2018) Mastering Predictive Analytics with scikit-learn and TensorFlow: *Packt Publishing*