# MODELO PREDICTIVO DE HIPERTENSIÓN PARA EL DIAGNÓSTICO DE PACIENTES CON FACTORES DE RIESGO CARDIOVASCULAR EN EL DEPARTAMENTO DE BOLÍVAR, MEDIANTE TÉCNICAS DE DEEP LEARNING.

**CRISTIAN RUIZ SANCHEZ**
**GISSELLA ROJAS ACEVEDO.**

**UNIVERSIDAD DEL SINÚ ELÍAS BECHARA ZAINÚM**
**FACULTAD DE CIENCIAS EXACTAS E INGENIERÍAS**
**ESCUELA DE INGENIERÍA DE SISTEMAS**
**CARTAGENA-COLOMBIA**
**Enero 2021**

# MODELO PREDICTIVO DE HIPERTENSIÓN PARA EL DIAGNÓSTICO DE PACIENTES CON FACTORES DE RIESGO CARDIOVASCULAR EN EL DEPARTAMENTO DE BOLÍVAR, MEDIANTE TÉCNICAS DE DEEP LEARNING.


**Estudiantes:**
Cristian Ruiz Sánchez
Gissella Rojas Acevedo

**Director del proyecto**
Eugenia Arrieta Rodríguez



**UNIVERSIDAD DEL SINÚ ELÍAS BECHARA ZAINÚM**
**FACULTAD DE CIENCIAS EXACTAS E INGENIERÍAS**
**ESCUELA DE INGENIERÍA DE SISTEMAS**
**CARTAGENA-COLOMBIA**
**Enero 2021**

**Agradecimientos**

Un peldaño muy importante avanzado, gracias primeramente a Dios, a mi familia, a todo el grupo interdisciplinar de docentes de la Universidad del Sinú, que me impulsaron a forjar un carácter sólido como estudiante, como profesional y no menos importante como persona. Al grupo de compañeros que aprendimos a que trabajar en equipo no es fácil, pero dando un poco más de cada uno de nosotros lo pudimos lograr en las jornadas académicas. Agradecimiento muy especial para nuestra asesora y tutora principal en el desarrollo de este proyecto, la ingeniera Eugenia Arrieta Rodríguez, quien fue clave para culminar este proceso en hora buena.

CRISTIAN RUIZ SANCHEZ

**Agradecimientos**

Debo agradecer de forma sincera a la profesora Eugenia Arrieta por aceptarme para hacer este proyecto bajo su dirección. Con su ayuda y confianza. En mi trabajo y su capacidad para dirigir mis ideas fue una contribución inmensa y constante, no solamente en el desarrollo de esta tesis, sino además en mi formación. Las ideas propias, continuamente enmarcadas en su orientación y rigurosidad, fueron la clave del buen trabajo que hemos llevado a cabo, el cual no se puede concebir sin su constantemente apropiada colaboración. Le agradezco además el haberme facilitado constantemente los medios suficientes para realizar cada una de las fases propuestas a lo largo del desarrollo de este proyecto. Muchas gracias profe.

GISSELLA ROJAS ACEVEDO.

## CONTENIDO

## TABLA DE ILUSTRACIONES

## RESUMEN

El presente proyecto de investigación cómo opción de grado del Programa de Ingeniería de Sistemas, es un estudio basado en el desarrollo de una herramienta que permita hacer la predicción temprana de sufrir de la hipertensión arterial, basándose en unas escalas de riesgos que van desde el menos probable de padecer esta patología, pasando por el de mediano nivel y finalmente identificando los pacientes que estén dentro de alto riesgo de contraerla. Para este estudio se utilizaron técnicas de aprendizaje profundo supervisado.

La metodología del proyecto esta soportada por tres fases estructuradas, la primera es la búsqueda de información relacionada con el tema del proyecto la cual permitió identificar que modelos de clasificación eran los más adecuados para la realización del proyecto, las cuales fueron, árbol de decisión y red neuronal, la segunda fase conlleva el tratamiento de los datos obtenidos del hospital municipal de Arjona Bolívar, dichos datos pasaron por varias técnicas de limpieza tales como: verificación y rellenado de datos faltantes, balanceo de los tres tipos de clases en la variable de respuesta (riesgo bajo, riesgo medio y riesgo alto), conversión de datos cadena a numéricos, de  para después realizar la última fase que es el desarrollo del modelo donde se toma primeramente un modelo de predicción muy sencillo convencional, árbol de decisión,  en este se tiene unos datos de entrada independiente que consta de alrededor de 40 variables independientes que permiten definir la correlación con la variable de respuesta como base para la predicción de tipo clasificación. Finalmente, y como modelo central del análisis general de la predicción para encontrar qué tan alto el nivel de riesgo una paciente que sufre de hipertensión arterial. El modelo fue construido con tres capas, donde la primera representa los datos de entradas, conformada por 60 neuronas y una función de activación igual a *relu.* La siguiente parte viene siendo la capa oculta quien cuenta con 30 neuronas y función de activación similar a la anterior capa, *relu*, y por último la de salida posee tres neuronas multiconectadas como las anteriores, finalmente en el resultado que arrojó se mostró una precisión y una actitud alrededor del 87% en la predicción final. Hay que tener en cuenta que para realizar la predicción final y verificación del modelo se tomó el 30% de la muestra de todos los datos disponibles.

## INTRODUCCIÓN

La Hipertensión Arterial es el estado en el cual, el individuo presenta la presión arterial sistémica permanentemente elevada, esta es una enfermedad crónica en el cual se aumenta la presión de la sangre que es bombeada a través del corazón hacia las arterias por todo el organismo, que al circular la sangre los vasos sanguíneos poseen una tensión persistentemente alta, lo cual puede dañarlos. La tensión normal en adultos está en 120 mm Hg1 cuando el corazón está latiendo, a esto se le denomina tensión sistólica y de 80 mm Hg cuando el corazón está en estado de relajación, esto se denomina tensión diastólica. Cuando la tensión sistólica es igual o supera los 140 mm Hg y/o la tensión diastólica es igual o sobrepasa los 90 mm Hg, la tensión arterial se estima alta o elevada.[1]

Además, en el conjunto de las patologías, la hipertensión arterial es el primordial elemento de peligro de muerte y patología[1] internacionalmente, en especial, es causa de infartos de miocardio, accidentes cerebrovasculares, insuficiencia renal, ceguera, vasculopatía periférica e insuficiencia cardiaca. Este puede ser más peligroso si la persona además de tener esta enfermedad posee otra, en particular como la diabetes.

Por otra parte, la hipertensión es la 1ª causa de patología en las naciones desarrollados; la 2ª causa de patología, luego del tabaquismo, en las naciones en desarrollo; la 1ª causa de ataque cerebrovascular e insuficiencia cardiaca; y la 2ª causa de síndrome coronario agudo.[2]

Ahora bien, algunas de las razones específicas que provocan la hipertensión arterial, sí se ha referente con una serie de componentes que acostumbran a estar presentes en la mayor parte de los individuos que la sufren. Conviene dividir esos involucrados con la herencia genética, el sexo, la edad y la raza, y por consiguiente poco modificables, de esos otros que se podrían cambiar al variar los hábitos y el ambiente en el cual viven los individuos, como la obesidad, la sensibilidad al sodio, el consumo desmesurado de alcohol, la utilización de ciertos fármacos y un estilo de vida bastante sedentario.[3]

De acuerdo a esto se realizó una investigación de antecedentes sobre organizaciones o personas que hubieran llevado a cabo proyectos de inteligencia artificial para predecir los riegos de contraer hipertensión arterial, se encontró que en la mayoría tuvieron un nivel de precisión

[1] Patología: Parte de la medicina que estudia los trastornos anatómicos y fisiológicos de los tejidos y los órganos enfermos, así como los síntomas y signos a través de los cuales se manifiestan las enfermedades y las causas que las producen.

alrededor del 80% correcto, porcentaje de alguna forma aceptable pero un poco medio, no alto, de la confiabilidad al momentos de aplicar dichos modelos. Con el ánimo de mejorar los proceso en este ámbito, y aportando una parte muy significativa en la solución del problema de contar con herramienta ágil y que garantice la veracidad de determinar tempranamente el riesgo de un paciente de sufrir de hipertensión arterial, se plantea desarrollar una herramienta que utiliza los datos de diferentes pacientes para hacer un análisis logrando hacer un diagnóstico lo más acertado para predecir si la persona es propensa a tener enfermedades cardiovasculares en un futuro y de dar un posible tratamiento de manera oportuna, estas predicciones se basaran en la toma de algunas variables tales como: la presión arterial, los niveles de azúcar que se pueden presentar en la sangre y niveles de oxígeno, entre otras que podrían influir. Se implementarán técnicas de análisis de variables para determinar cuáles son las que más relevancia e incidencia tiene respecto a la variable de respuesta.

Para la solución se desarrolló un modelo de inteligencia artificial utilizando técnicas de aprendizaje profundo, más concretamente la implementación de una red neuronal que consta de tres capas, capa de entrada, capa oculta y finalmente la capa de salida. Es importante enfatizar que este proyecto está diseñado y dirigido dar solución puntualmente a los pacientes del municipio de Arjona y el departamento Bolívar.

# 1. DISEÑO METODOLÓGICO

## 1.1. Descripción del problema

La Hipertensión Arterial perjudica alrededor al 20% poblacional adulta de la más grande parte de las naciones es la primera causa de morbilidad y motiva el mayor número de consultas en las afecciones del artefacto circulatorio.

Así mismo, es el componente de peligro de mayor relevancia para la patología cardio cerebrovascular, y constantemente se asocia con otros componentes de peligro bien conocidos como por ejemplo dieta, elevación de lípidos sanguíneos, obesidad, tabaquismo, Diabetes Mellitus e inacción física (sedentarismo). Por otra parte, Varias personas que poseen presión arterial alta causada por una patología no diagnosticada. Esta clase de presión arterial alta es llamada hipertensión secundaria, tiende a aparecer de manera frecuente y causa una presión arterial más alta que la hipertensión primaria. Varios trastornos y medicamentos tienen la posibilidad de generar hipertensión secundaria, algunos de ellos son:

- Problemas en los riñones o problemas renales.
- Problemas de tiroides.
- El consumo de sustancias psicoactivas, como la cocaína y las anfetaminas

Aun cuando no existe una causa concreta se conoce que alguna de estas razones, entre otras, juega un papel bastante fundamental en su desarrollo para que se presenten una presión arterial elevada, los siguientes son algunos factores de riesgo:

- La edad. El peligro de sufrir de presión arterial alta aumenta con la edad. Hasta alrededor los 64 años, esta es más común en los hombres. Las mujeres son más sensibles a desarrollar lo luego de los 65 años.
- Antecedentes familiares: esta enfermedad puede ser hereditaria en algunos casos.
- Sobrepeso
- Consumo de tabaco. Fumar tabaco no solo aumenta la presión arterial momentáneamente, sino que los químicos del tabaco tienen la posibilidad de perjudicar el revestimiento de los muros arteriales. Esto puede hacer que las arterias se estrechen y aumente el peligro

de patología cardíaca. El tabaquismo pasivo además puede incrementar el peligro de patología cardíaca.[4]

Acudiendo a la reciente y constante necesidad del área de la salud para tomar decisiones correctas y en los menores tiempos posibles, se sugiere utilizar plataformas, arquitecturas o algoritmos que permitan desarrollar un modelo efectivo para optimizar el trabajo del médico durante el proceso de atención de pacientes.

A continuación, se plantea el desarrollo de un modelo de predicción de riesgos para enfermedades de presión alta. Así mismo, se plantea implementar técnicas de DEEP LEARNING, para realizar predicción del riesgo que puedan tener los pacientes de sufrir estas patologías, en el municipio de Arjona, Bolívar, quienes en su gran mayoría y a raíz de factores como socio-económico, cultural, hereditarios, entre otros, poseen un grado alto de padecer hipertensión arterial. Basándonos en una predicción mediante el uso de variables clínicas, como anteriormente mencionadas, algunos ejemplos: la presión de la sangre, los niveles de azúcar en la sangre y niveles de oxígeno, edad, peso, estilos de vida.

## 1.2. Justificación

La hipertensión arterial es un problema grave de salud pública presente en todos los rincones del mundo. Esta es causada por diferentes factores, como la carga genética, la raza y el sexo, hasta el padecimiento de enfermedades metabólicas como diabetes mellitus y dislipidemia, e inclusive factores comportamentales como el consumo de alcohol, tabaco y sedentarismo.

Según varias organizaciones a nivel mundial y especialistas coinciden en que es una de las enfermedades que causa mayor número de decesos. Esta condición es el principal factor de riesgo para enfermedad cardiovascular, falla renal, mortalidad prematura y discapacidad. De los 17 millones de muertes al año causada por enfermedad cardiovascular, el padecimiento de hipertensión arterial explica hasta 9.4 millones de dichas muertes. Más aún, durante el 2008, la OMS estimó que la esta fue la responsable de al menos el 51% de las muertes debidas a accidente cerebrovascular. Para el 2014, la prevalencia mundial de hipertensión en mayores de 18 años fue de 22.2%.

En Colombia, la Encuesta Nacional de Salud realizada en el 2007 arrojó una prevalencia de hipertensión en la población general de 22.8%. Durante el 2011, la prevalencia se estimó en 7.29%, con una incidencia anual de 191 por cada 100 000 habitantes y una tasa de mortalidad de 13 por cada

100.000 habitantes. Para el 2013, según datos arrojados por la cuenta de alto costo, en Colombia había 2.414.354 personas con hipertensión arterial afiliadas al Sistema General de Seguridad Social en Salud, arrojando una prevalencia de 5.53%. A demás se reportaron una prevalencia de hipertensión de 37.5% en población colombiana urbana y rural de 35-70 años, pertenecientes.

Del mismo modo, como uno de los motivos de emergencias que toman un puesto superior entre los adultos mayores se encuentra la hipertensión, entre los periodos de 2005 hasta 2010 donde se registró la “tasa de mortalidad promedio ajustada por edad” en los departamentos que presentaban mayores índices de mortalidad por hipertensión fuero: Boyacá, Casanare, Meta, San Andrés y Vichada.[2]. Con ánimo de hacer un aporte a la investigación, a la prevención y detección temprana de los factores de riesgos que aumentan la probabilidad de que una persona padezca de hipertensión se plantea, este proyecto de centra en diseña y desarrolla un modelo de DEEP LEARNING capaz de ayudar a diagnosticar pacientes cuyas condiciones de salud, hábitos, entre otros aspectos aumente su vulnerabilidad ante la Hipertensión.

En orden de ideas, este proyecto aportará a la línea de investigación de inteligencia artificial del grupo de investigación DEARTICA. Debido a que este proyecto involucra técnicas de Machine learning y Deep Learning que son unas de las temáticas de gran impacto a nivel de investigaciones en el campo de la ingeniería y ciencias de la computación.

## 1.3. Alcance

Se desarrolló un modelo predictivo tomando como referencias variables para poder hacer el análisis y predicciones en los pacientes. Del total de pacientes encontrados en los registros clínicos proporcionados por el hospital local de Arjona Bolívar, se determinó que el 70% de estos se utilizarían para el entrenamiento del modelo en fase de construcción, el 30% restante se tomó para realizar las pruebas del modelo. Es importante resaltar que el este proyecto también está definido para implementar en el resto del departamento de Bolívar.

## 1.4. Preguntas de investigación

- ¿Cuáles de las técnicas de deep learning permite obtener un predictor de hipertensión con una precisión superior al 80%?
- ¿Cuáles son las variables que se deben tener en cuenta para una mejor predicción?

## 1.5. Objetivos

### 1.5.1. General

Desarrollar un modelo predictivo de hipertensión para realizar el diagnóstico de pacientes con factores de riesgo cardiovascular en el departamento de Bolívar, mediante técnicas de DEEP LEARNING.

### 1.5.2. Específicos

- Búsqueda de información que permita identificar que modelos de clasificación son los más adecuados para el desarrollo del proyecto.
- Realizar un análisis de los datos o las variables más significativas para determinar el riesgo cardiovascular.
- Desarrollar el modelo con las técnicas de clasificación que mejor se adecue al problema.
- Entrenar el modelo con el ánimo de que vaya realizando de aprendizaje automático con segmento de los datos brindados, para ello:
  Se determinan las técnicas para aplicar en el proceso de predicción basado en estudios antes hechos.
- Evaluar el modelo de DEEP LEARNING utilizando métodos de validación de resultados.

## 1.6. Revisión sistemática

Se efectuó una búsqueda de trabajos o investigaciones relacionadas acerca de la predicción de la hipertensión en el buscador académico Google Académico (GOOGLE SCHOLAR), se establecen los términos o palabras claves para las búsquedas asociadas a la investigación. Dichos términos o palabras se resumen en la Tabla 1.

**Tabla 1: términos relacionados con el tema de investigación.**

| TÉRMINOS | VARIABLES |
|---|---|

| | |
|---|---|
| Predicción de la hipertensión | X1 |
| Red neuronal | X2 |
| Diagnóstico de hipertensión arterial. | X3 |

Se generan las búsquedas haciendo las combinaciones de las variables, un ejemplo es siguiente manera:

- (X1 OR X3) AND X2;
- X1 AND X2
- X2 AND X3

**Selección de Trabajos y Criterios de Inclusión y Exclusión**

En la Tabla 2, se presentan los criterios de inclusión y exclusión.

**Tabla 2: criterios que se tomaron en cuenta para la búsqueda de información**

| | |
|---|---|
| Criterios de Inclusión | ● Publicaciones en español e inglés<br>● Trabajos o investigaciones donde se presentarán datos de mucha relevancia en la identificación de variables que influyen en los resultados de la predicción.<br>● Publicaciones del año 2014 en adelante. |
| Criterios de Exclusión | ● Publicaciones que no están disponibles para su revisión completa.<br>● Trabajos de años anteriores a 2014<br>● Publicaciones no digitales.<br>● Que en su tema se involucra a niños menores de 1 año.<br>● Que no tenían procedimientos o modelos a seguir para las predicciones. |

Inicialmente se encontraron 51 publicaciones relacionadas donde se exponía acerca de la temática que se está tratando, de los cuales 43 contenían información o contenían ciertas similitudes, pero no se centraban en el estudio principal, debido a que su enfoque era en puntos de poca relevancia para el proyecto, se excluyeron. 2 de ellos eran solo información duplicada de las anteriores investigaciones, también quedaron por fuera, dejando 6 los cuales se pueden tomar como referencia para el proyecto. La siguiente imagen ilustra el proceso de selección de las publicaciones encontradas y en la tabla 3 las descripciones de las investigaciones finales.

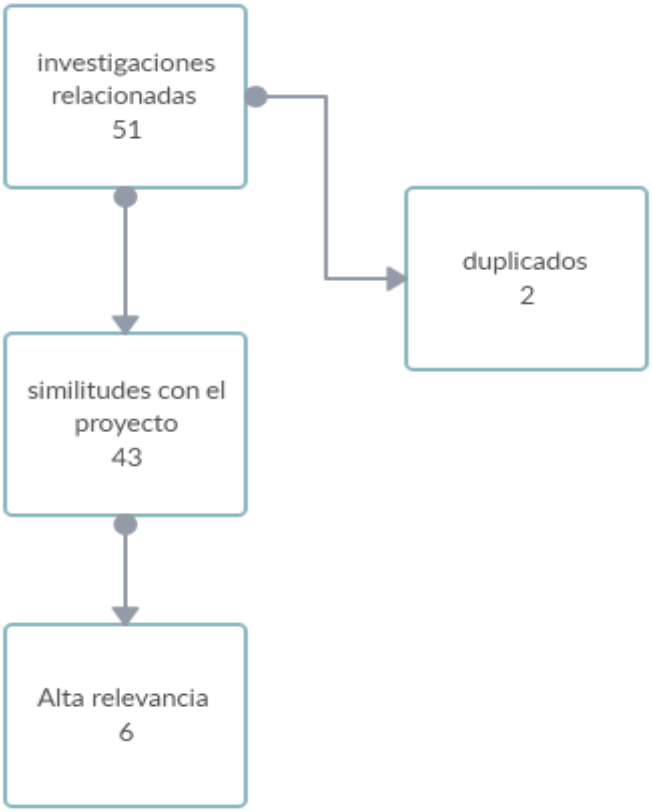



*Ilustración 1 filtrado de información*

**Tabla 3: cuadro descriptivo de las investigaciones encontradas**

| Año | Título de investigación | método |
|---|---|---|
| 2014 | Árbol para predecir el desarrollo de la cardiopatía hipertensiva | Árbol de predicción para el desarrollo de la cardiopatía hipertensiva, a partir de factores hemodinámicos y no hemodinámicos[5]. |
| 2015 | la hipertensión arterial: importancia de su prevención | Se especifica el tratamiento farmacológico para el tratamiento de la HTA, esto con el fin de controlar la presión arterial del paciente y más a largo plazo reducir la morbimortalidad, fundamentalmente de las enfermedades cardiovasculares, cerebrovasculares y renales asociadas a la HTA[6]. |

| | | |
|---|---|---|
| 2016 | Predicción del riesgo cardiovascular e hipertensión arterial según Framingham en pacientes de atención primaria en salud. Estudio FRICC | Estudio de campo observacional, descriptivo y correlacional, en el que se incluyeron 400 adultos de ambos sexos (M: 310 y H: 90), con una edad promedio para ambos géneros de 41,5 ± 16,4. haciendo uso de la escala de Framingham[7]. |
| 2017 | Validación de un modelo de predicción de hipertensión | Se realiza una valoración por expertos del modelo de predicción mediante un árbol de decisiones utilizando la técnica de CHAID (CHI SQUARE AUTOMATIC INTERACTION DETECTOR) para el vaticinio de hipertensión arterial en la adultez desde la adolescencia[8]. |
| 2018 | Red neuronal para el diagnóstico de hipertensión arterial | Uso de distintos métodos, como: los árboles de decisión, RAMDOM FOREST, KNN y RAMDOM TREE. Además del uso de las métricas de evaluación[9]. |
| 2019 | A machine learning approach for the prediction of pulmonary hypertension | En la base de datos de 90 pacientes con presión arterial pulmonar ( PAP ) determinada de forma invasiva con las correspondientes estimaciones ecocardiográficas de PAP obtenidas en 24 horas, se hizo la aplicación de cinco algoritmos ML ( bosque aleatorio de árboles de clasificación, bosque aleatorio de árboles de regresión, regresión logística penalizada por lazo ), árboles de clasificación mejorados, máquinas de vectores de soporte ) usando un esquema de validación cruzada ( CV ) de 10 veces 3 veces[10]. |

En las investigaciones encontradas se utiliza con más frecuencia los árboles de decisión, los cuales permiten hacer comparativas entre varias variables, hacer divisiones de ellas y ayudan a tomar la decisión “más acertada”, desde un punto de vista probabilístico, ante un abanico de posibles decisiones, por tal razón, se implementó la técnica de árbol de decisión junto con la red neuronal.

## 1.7. Estado del arte

### Proyectos mayormente relacionados

A continuación, se explica de manera un poco más detallada las investigaciones que tienen mayor relación con el proyecto, detallando de igual manera los métodos en que se realizaron dichas investigaciones, los proyectos están basados en estudios internacionales.

La investigación “Árbol para predecir el desarrollo de la cardiopatía hipertensiva” [5], tiene como objetivo la predicción para el desarrollo de la cardiopatía hipertensiva, a partir de factores hemodinámicos y no hemodinámicos[2]. Se realizó el diseño y validación de un árbol de predicción del desarrollo de la cardiopatía hipertensiva mediante el procedimiento de descubrimiento de conocimientos en bases de datos, conocido internacionalmente como proceso KDD ( del inglé Knowledge Discovery in Database ) y minería de datos ( Data Mining - DM ), en pacientes hipertensos atendidos en la consulta especializada de HTA de la Policlínica de Especialidades del Hospital General Universitario " Carlos Manuel de Céspedes " del municipio Bayamo, provincia Granma, Cuba, desde el 1ro de enero de 2004 hasta el 31 de diciembre de 2009. Como resultado de este método el árbol predijo el riesgo de desarrollar la cardiopatía hipertensiva a 82,598 % de los pacientes; con un área bajo la curva ROC de 0,861 y una tasa de verdaderos positiva de 0,733 y de 0,921 para las clases 1 y 2, respectivamente. El factor más importante lo constituyó la proteína C reactiva, seguida en orden de importancia por la glucemia, el ácido úrico, el colesterol y la microalbuminuria.

En [6] la autora explica la importancia de la prevención de la hipertensión arterial (HTA), exponiendo los diagnósticos y las causas que llegan a provocar esta enfermedad, además del tratamiento farmacológico para la HTA:
Los medicamentos más usados para el tratamiento de la HTA los podemos agrupar en:
Diuréticos: se denominan a veces «píldoras de agua». Se utilizan para tratar la insuficiencia cardíaca congestiva (ICC), la presión arterial alta (hipertensión) o el edema (retención de líquidos). Los diuréticos también se recetan para ciertos tipos de enfermedades del riñón o hígado. Disminuyen

[2] La hemodinámica es aquella parte de la biofísica que se encarga del estudio de la dinámica de la sangre en el interior de las estructuras sanguíneas como arterias, venas, vénulas, arteriolas y capilares, así como también la mecánica del corazón. https://ocw.unican.es/mod/page/view.php?id=537

la cantidad de Na y por tanto el volumen sanguíneo, disminuyendo la carga cardiaca por vasodilatación.

Alfabloqueantes: Los alfa bloqueadores, alfas bloqueantes, antagonistas alfa adrenérgicos o bloqueantes α son agentes farmacológicos que actúan como antagonistas de los receptores alfa adrenérgicos. Bloquean de manera selectiva y competitiva los receptores alfa1 adrenérgicos postsinápticos vasoconstrictores, produciendo vasodilatación arteriovenosa, reducción de las resistencias vasculares periféricas y de la PA.

Betabloqueantes: bloquean competitiva y reversiblemente los receptores betaadrenérgicos, disminuyendo la frecuencia y el gasto cardiaco además de bloquear la liberación de renina.

Antagonistas del calcio: se fijan a los canales de calcio tipo L voltajes dependientes eliminando la corriente de calcio que provoca la contracción muscular, produciendo la relajación del músculo liso vascular

Agentes que bloquean la producción o acción de la angiotensina:

- Inhibidores de la enzima convertidora de la angiotensina (IECAs): bloquean la síntesis de angiotensina II por inhibición competitiva de la enzima convertidora de la angiotensina (ECA) produciendo vasodilatación arteriovenosa además de nutrieres.

Antagonistas de los receptores de la angiotensina (ARA II): Bloquean de forma competitiva y selectiva los receptores AT1 inhibiendo la acción de la angiotensina II.

El objetivo es especificar el tratamiento farmacológico para el tratamiento de la HTA, esto con el fin de controlar la presión arterial del paciente y más a largo plazo reducir la morbimortalidad, fundamentalmente de las enfermedades cardiovasculares, cerebrovasculares y renales asociadas a la HTA.

Otro trabajo donde el enfoque está basado en la prevención del riesgo cardiovascular haciendo uso de la escala de Framingham[3] [11], El método fue un estudio de campo observacional, descriptivo y correlacional, en el que se incluyeron 400 adultos de ambos sexos (M: 310 y H: 90), con una edad La estratificación del riesgo permitió determinar el porcentaje de riesgo para enfermedad cardiovascular e hipertensión Se tomaron medidas antropométricas según la OMS para el IMC y perímetro abdominal. Y así mismo, se identificaron los factores de riesgo cardiovascular en la población de estudio.

[3] El Estudio de Framingham o Estudio Framingham del Corazón (en inglés Framingham Heart Study) es un estudio de cohortes de larga duración sobre el riesgo cardiovascular, que todavía se encuentra en marcha, realizado entre los residentes de Framingham.

Los resultados luego de la investigación, el 10% de los estudiados presentó algo de padecer la enfermedad los sus próximos 10 años, 14% y 75% presentó moderado y bajo. La clasificación para el riesgo de hipertensión bajo es de 1,2 y 4 años. Llevar una vida sedentaria aparece como un factor frecuente, El 21% presentó HTA; 7% Diabetes; 6,7% Tabaquismo; y el 89% bebedores.

con lo anterior, teniendo en cuenta de que las investigaciones encontradas, entre ellas la técnica que es común es usar arboles de decisión para el análisis de los datos. Teniendo en cuenta esto, se opta por usar redes neuronales para hacer el algoritmo que permita hacer la predicción con los datos de los pacientes obtenidos del hospital local de Arjona Bolívar.

## 1.8. Marcos de referencia

### 1.8.1. Marco teórico

Podemos definir como DEEP LEARNING la rama de Aprendizaje Máquina basada en un conjunto de algoritmos que intentan modelar abstracciones de alto nivel en los datos usando múltiples capas de procesamiento con estructuras complejas, o bien compuestas de múltiples transformaciones no lineales.

Entonces se considera parte de una familia más amplia de métodos de Aprendizaje Máquina basados en aprendizaje de representaciones de los datos[12]. En la ilustración 2 se muestra como está compuesto el concepto de MACHINE LEARNING y cuáles son sus derivaciones.

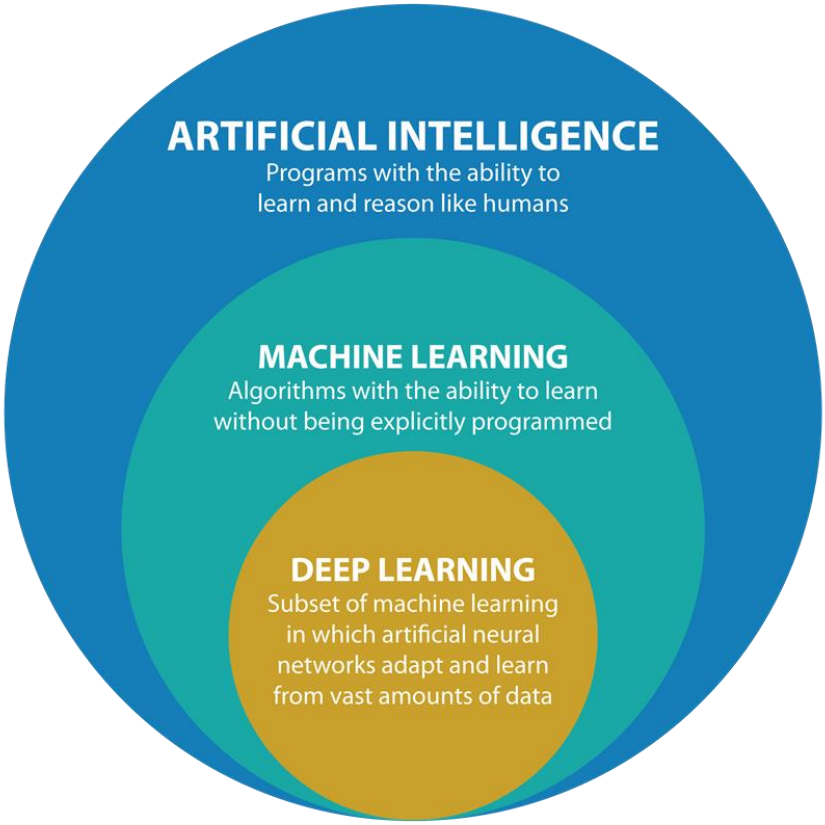



*Ilustración 2 (Oracle, E. (2018, septiembre 14). Diferencias entre la Inteligencia Artificial y el Machine Learning.) Composición de MACHINE LEARNING*

Un de ejemplo de DEEP LEARNING es el siguiente, una imagen puede ser representada de muchas maneras: como un vector de intensidades por píxel, o como un conjunto de regiones con una forma o curvatura particular. Algunas representaciones hacen más fácil el aprender tareas como el reconocimiento de rostros o de voces

Teniendo en cuenta lo anterior es pertinente afirmar que DEEP LEARNING representan modelos con múltiples capas o etapas de procesamiento de información no lineal y/o Métodos para el aprendizaje supervisado o no supervisado para la representación de características usando capas sucesivas y más abstractas.

#### 1.8.1.1. Inicios

Hace algunos años se viene hablando sobre DEEP LEARNING, dicen que es un área nueva con importante potencial y que incluso es relativamente fácil (en comparación de otras ´áreas) hacer aportaciones nuevas a La disciplina, pues hay problemas relativamente sencillos que no han sido abordados Debido a la falta de especialistas en el tema. Otros dicen que no es un área nueva y Que simplemente antes no tenía nombre, pero que DEEP LEARNING se viene estudiando Desde hace 20 años, con las redes

neuronales. Si vemos a DEEP LEARNING dentro del contexto de redes neuronales artificiales podemos Identificar tres corrientes de investigación:

- Cibernética (1940-1960): se logra entrenar una sola capa de neuronas.
- Conexionismo (1980-1995): se logra entrenar una red neuronal con una o dos capas Ocultas.
- Deep Learning (2006): se logran entrenar redes neuronales muy profundas, de 4 o Más capas.

Como se puede observar, los antecedentes de DEEP LEARNING han reflejado diferentes Enfoques de investigación y en ocasiones han sido populares y en otras no tanto, pero es Claro que DEEP LEARNING es heredero de las redes neuronales. Por otro lado, es importante mencionar que el trabajo de investigación que logró el resurgimiento de las redes Neuronales en 2006 fue el de “A fasto LEARNING algoritmo foro Deep belio nets “de Hilton y Asidero. Este trabajo mostró que, en contra de lo que se creía anteriormente, Si era posible entrenar eficientemente un algoritmo de DEEP LEARNING.[13]

Existen muchas formas de aplicar el aprendizaje profundo, pero podemos detallar un grupo de estas que las sintetizan, por ejemplo:

**Clasificación:** cada individuo tiene una categoría asignada, por ejemplo, si queremos clasificar el tipo de estudiante (aplicado/desaplicado); también podemos nombrar si queremos ver que tan avanzada está la enfermedad una persona. El objetivo es predecir a que clase pertenece un elemento o sujeto dado.[14]

**Regresión:** el objetivo es predecir un número real para un individuo. Dicha variable real guarda una relación funcional con otras variables dadas. Ejemplos típicos son predecir el ingreso promedio de una familia colombiana, los niveles de colesterol de un enfermo de diabetes. Otro ejemplo son las series temporales, en donde también la variable a predecir es una variable real que guarda una dependencia temporal con ella misma, como cuando a partir de un histórico de ventas queremos realizar una predicción de esta en tiempo determinado.[15]

**Ranking:** ordenar elementos según algún criterio. Un ejemplo típico es regresar las páginas que son relevantes dada una consulta de búsqueda. Este tipo generalmente surge para diseñar algoritmos extracción de información y de procesamiento de lenguaje natural.[16]

**Clustering:** estos algoritmos generan conglomerados dentro de los cuales los individuos son similares y fuera de ellos no. Por ejemplo, en análisis de redes sociales se aplican estos algoritmos para encontrar comunidades de usuarios con intereses en común. También en investigación de mercados se buscan conglomerados de clientes que se parezcan, se estudian las

preferencias de dichos conglomerados y posteriormente se dirigen campañas publicitarias a los mismos.[17]
**Reducción de dimensionalidad:** dada una representación mediante variables de un individuo, se busca poder representarlo usando menos variables preservando la mayor cantidad de información sobre los individuos. En ocasiones esta técnica se usa para obtener índices. Un ejemplo son los índices de marginación que publica el DANE.

#### 1.8.1.2. Herramientas para trabajar con DEEP LEARNING

Lenguajes de programación más usados:
**Phyton:** es un lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional.
**R:** es un entorno y lenguaje de programación con un enfoque al análisis estadístico. R nació como una reimplementación de software libre del lenguaje S, adicionado con soporte para alcance estático.

IDE de desarrollo:
**Entorno Anaconda:** es una distribución libre y abierta de los lenguajes Python y R, utilizada en ciencia de datos, y aprendizaje automático. Esto incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputos científicos.

**Jupyter**: es un software de código abierto, estándares abiertos y servicios para computación interactiva en docenas de lenguajes de programación, entre ellos Julia, Phyton y R.

**Spyder:** es un entorno de desarrollo integrado multiplataforma de código abierto para programación científica en lenguaje Python.

**Azure IA:** Plataforma de Microsoft para el desarrollo de modelos de MACHINE LEARNING y DEEP LEARNING en la nube. Muy robusta y por ello es necesario pagar por utilizar licencia.

**Google Colab:** es un servicio cloud, basado en los Notebooks de Jupyter, que permite el uso gratuito de las GPUs y TPUs de Google, con librerías como: Scikit-learn, PyTorch, TensorFlow, Keras y OpenCV. Todo ello con bajo Python 2.7 y 3.6, que aún no está disponible para R y Scala.

**1.8.1.3.** Bibliotecas y APIS

**TensorFlow:** es una biblioteca de código abierto que se basa en un sistema de redes neuronales. Esto significa que puede relacionar varios datos en red simultáneamente, de la misma forma que lo hace el cerebro humano. Por ejemplo, puede reconocer varias palabras del alfabeto porque relaciona las letras y fonemas.

**Keras:** es una biblioteca de Redes Neuronales de Código Abierto escrita en Python. Es capaz de ejecutarse sobre TensorFlow, Microsoft Cognitive Toolkit o Theano. Está especialmente diseñada para posibilitar la experimentación en más o menos poco tiempo con redes de Aprendizaje Profundo.

**Scikit Learn:** Scikit-learn es una biblioteca de aprendizaje automático gratuita para Python. Cuenta con varios algoritmos como máquina de vectores de soporte, bosques aleatorios y vecinos k, y también admite bibliotecas numéricas y científicas de Python como NumPy y SciPy.

**Shogun:** es una biblioteca de software de aprendizaje automático de código abierto escrita en C ++. Ofrece numerosos algoritmos y estructuras de datos para problemas de aprendizaje automático. Ofrece interfaces para Octave, Python, R, Java, Lua, Ruby y C # usando SWIG.

### 1.8.2. Marco conceptual

Hipertensión: La hipertensión arterial es una enfermedad frecuente que afecta a un tercio de la población adulta. Se produce por el aumento de la fuerza de presión que ejerce la sangre sobre las Arterias de forma sostenida. Es una enfermedad que no da síntomas durante mucho tiempo y, si no se trata, puede desencadenar complicaciones severas como infarto de corazón, accidente cerebrovascular, daño renal y ocular, entre otras complicaciones. Se puede evitar si se controla adecuadamente.[18]

Variable: Una variable refiere, en una primera instancia, a cosas que son susceptibles de ser modificadas (de variar), de cambiar en función de algún motivo determinado o indeterminado. El término variable alude a las cosas de poca estabilidad, que en poco tiempo pueden tener fuertes alteraciones o que nunca adquieren una constancia (muy frecuentemente sucede esto con el clima, o el humor de una persona).[19]

Red Neuronal: Las redes neuronales artificiales son un modelo computacional vagamente inspirado en el comportamiento observado en su homólogo biológico. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales

Árbol de decisión: Los árboles de decisión son una técnica de aprendizaje automático supervisado muy utilizada en muchos negocios. Como su nombre indica, esta técnica de MACHINE LEARNING toma una serie de decisiones en forma de árbol, estos pueden usarse para resolver problemas tanto de clasificación como de regresión.

Regresión logística: es un método estadístico para predecir clases binarias. El resultado o variable objetivo es de naturaleza dicotómica. Dicotómica significa que solo hay dos clases posibles. Por ejemplo, se puede utilizar para problemas de detección de cáncer o calcular la probabilidad de que ocurra un evento. La Regresión Logística además es uno de los algoritmos de MACHINE LEARNING más simples y utilizados para la clasificación de dos clases. Es fácil de implementar y se puede usar como línea de base para cualquier problema de clasificación binaria. La Regresión Logística describe y estima la relación entre una variable binaria dependiente y las variables independientes.

MACHINE LEARNING: Es una derivación de inteligencia artificial que crea sistemas que aprenden de manera automatizada, es decir, identificar patrones complejos en millones de datos, predecir comportamientos todo mediante un algoritmo y que además de todo son capaces de mejorarse de manera independiente con el tiempo.

### 1.8.3. Marco legal y consideraciones éticas

El presente proyecto tiene sus bases legales sobre los siguientes pilares de normas, decretos y leyes del estado colombiano:

I. Decreto 846 de 2016; Por el cual se modifica la estructura del Departamento Administrativo de Ciencia, Tecnología e Innovación - COLCIENCIAS.
II. Decreto 591 del 26 de febrero de 1991 por el cual se regulan las modalidades específicas de contratos de fomento de actividades científicas y tecnológicas.
III. Decreto 585 del 26 de febrero de 1991 por el cual se crea el Consejo Nacional de Ciencia y Tecnología, se reorganiza el Instituto Colombiano para el Desarrollo de la Ciencia y la Tecnología-Colciencias- y se dictan otras disposiciones.

IV. Decreto 584 del 26 de febrero de 1.991, por el cual se reglamentan los viajes de estudio al exterior de los investigadores nacionales.

V. Decreto 393 del 26 de febrero de 1991 por el cual se dictan normas sobre asociación para actividades científicas y tecnológicas, proyectos de investigación y creación de tecnologías.

VI. Decreto 1467 del 2018 por el cual adiciona y modifica el Decreto 1082 de 2015 con el objeto de reglamentar la Ley 1923 de 2018 y se dictan otras.

VII. Decreto 293 del 2017; Por el cual se reglamenta el artículo 7 de la Ley 1753 de 2015 en lo relacionado con los Planes y Acuerdos Estratégicos Departamentales en Ciencia, Tecnología e Innovación y se dictan otras.

VIII. Ley 1581 de 2012 y el Decreto 1377 de 2013, se desarrolla el derecho constitucional que tienen todas las personas a conocer, suprimir, actualizar y rectificar todo tipo de datos personales recolectados, almacenados o que hayan sido objeto de tratamiento en bases de datos en las entidades del públicas y privadas.

## 1.9. Metodología

### Línea de Investigación

La Universidad del Sinú Seccional Cartagena cuenta con varios Grupos de Investigación que trabajan con el fin mostrar avances tecnológicos a la optimización de procesos y generación de nuevos conocimientos. Este proyecto aportará a la línea de investigación de inteligencia artificial del grupo de investigación DEARTICA. Debido a que este proyecto involucra técnicas de Machine learning y Deep Learning que son unas de las temáticas de gran impacto a nivel de investigaciones en el campo de la ingeniería y ciencias de la computación.

### Tipo de investigación

Este trabajo corresponde a una investigación aplicada, de cohorte retrospectiva en el cual los sujetos se estudian posteriormente de haberse producido la enfermedad. Los datos se obtendrán de los pacientes atendidos en el hospital local de Arjona y el diagnóstico de estas realizadas por un profesional de la salud; en el que se aplican los conocimientos y las técnicas de inteligencia artificial para contribuir en la solución de un problema de la vida real, como es el apoyo diagnóstico en el departamento de cardiología, específicamente aplicado a la hipertensión. En este tipo de investigación el énfasis del análisis está en la aplicación efectiva de las

técnicas de inteligencia artificial para obtener resultados positivos en términos de sensibilidad y especificidad.

**Muestra y Población**

Se tomaron los registros clínicos de los pacientes proporcionados por el hospital local de Arjona Bolívar, estos permitieran encontrar las variables que mayor determina riesgos cardiovasculares.

Estos datos fueron llevados a un término de pacientes anónimos para salvaguardar la identidad real de estos, respetando los principios de ética profesional. Del total de pacientes encontrados en los registros clínicos se determinó que el 70% de estos se utilizarían para el entrenamiento del modelo en fase de construcción, el 30% restante se tomó para realizar las pruebas del modelo.

**Variables**

En la tabla 5 se muestra las descritas algunas de las variables presentes en el estudio de la hipertensión.

**Tabla 5: variables de estudio**

| Nombre | Descripción | Unidad de mediad | Tipo de variable | Valores normales |
|---|---|---|---|---|
| Edad | Edad de la persona | Años | Numérica | 0-120 años |
| Talla | Medida de la persona en estatura | Centímetros | Numérica | 1-200 centímetros |
| Peso | fuerza que ejerce un determinado cuerpo sobre el punto en que se encuentra apoyado | kilogramos | Numérica | 0-1000 kg |
| Temperatura | Grado o nivel térmico de un cuerpo o de la atmósfera. | Celsius | Numérica | 0 a -273,15 ºC |
| Pulso | cantidad de veces que el corazón late durante un minuto. | Frecuencia | Numérica | 60 y 100 latidos por minuto |

| | | | | |
|---|---|---|---|---|
| Tensión arterial | Cantidad de presión que se ejerce en las paredes de las arterias al desplazarse la sangre | mmHg (milímetros de mercurio) | Numérica | Presión arterial normal: (máxima) están entre 120-129 mmHg, (mínima) entre 80 y 84 mmHg. Presión arterial normal-alta: (máxima) están entre 130-139 mmHg, (mínima) entre 80-89 mmHg |

**Selección de la metodología**

En la tabla 4 se muestra las fases con las que se llevó a cabo el proyecto de investigación, haciendo una descripción de las tareas en cada una de las fases.

**Tabla 4: fases del proyecto**

| Fase | Objetivo | Actividad |
|---|---|---|
| Búsqueda de la documentación acerca del tema de investigación | Adquirir conocimiento del tema en cuestión | Escogencia de las bases de datos donde se alojan los documentos informativos. |
| | | Identificación de variables para hacer la búsqueda sistemática. |
| | | Filtrado de información, análisis y comparación de los documentos encontrados. |
| Limpieza de datos | Obtener un conjunto de datos limpio para utilizarlo posteriormente en el modelo. | Búsqueda de columnas faltantes y eliminación de las columnas que tengan más de 60% de pérdida. |
| | | Visualizar de manera gráfica los datos y como están estructurados |
| | | Rellenado de columnas con datos faltantes menor a 60% de perdida. |

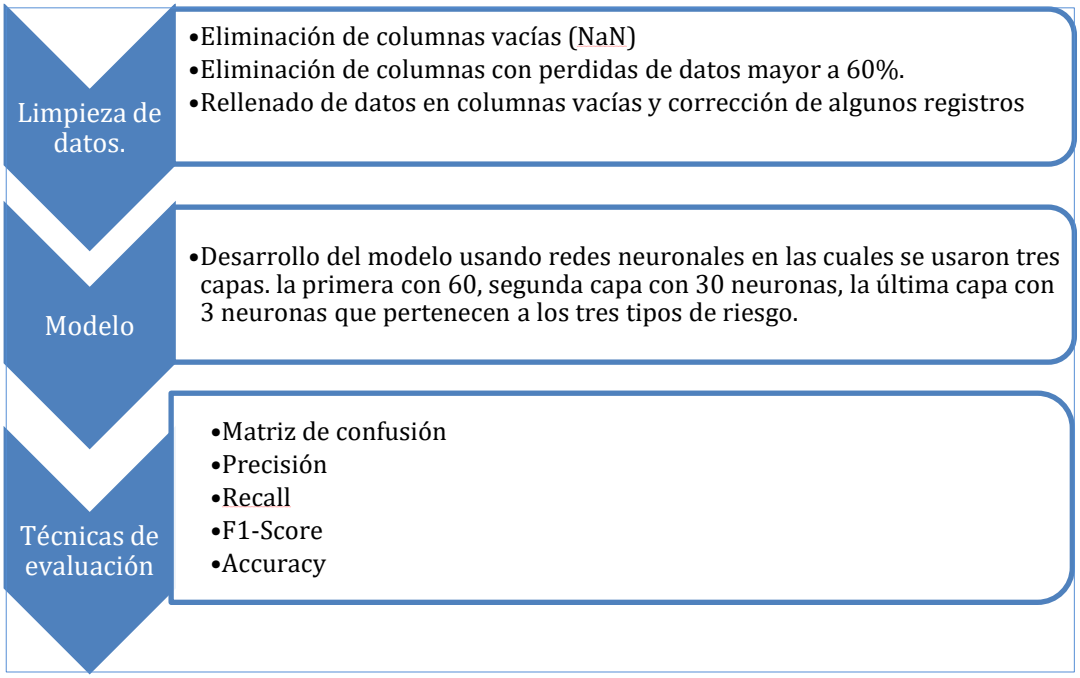
| | | |
|---|---|---|
| | | Convertir variables de columnas categóricas a nuevas columnas binarias. |
| | | Verificar datos de la variable de respuesta. |
| Construcción del modelo | Desarrollar un modelo implementando las técnicas de MACHINE LEARNING para el diagnóstico de la hipertensión arterial con los datos obtenidos de los pacientes. | Recolección de los datos de los pacientes del hospital local de Arjona |
| | | Organizar los datos para la aplicación del modelo. |
| | | Escogencia de la data set para el entrenamiento y pruebas del modelo |
| | | Usar técnicas para la normalización de los datos. |
| | | Comprobar el modelo construido usando el 30% de los datos de pruebas, hacer la comparación con el error obtenido en entrenamiento vs el error de pruebas. |
| | | Redacción de los capítulos que corresponden a la construcción del modelo con relación, además de la documentación de los escenarios de experimentación y pruebas efectuadas en la identificación de la técnica de DEEP LEARNING. |
| Evaluación | Evaluar el funcionamiento del modelo | Verificar la funcionalidad y la exactitud del modelo utilizando el conjunto de datos destinados para las pruebas. |
| | | Haciendo uso de las métricas, evaluar que tan preciso es a la hora de arrojar los resultados de las predicciones. |
| | | Redacción de los capítulos finales de la monografía en la relación a los resultados obtenidos productos de la investigación |

# 2. ANÁLISIS DE DATOS

En este capítulo se muestra la realización de la exploración de los datos y el modelo que se implementó para hacer el estudio de los pacientes.

En primera medida se realizó la extracción de los datos de los pacientes para su posterior limpieza, en los cuales se eliminaron gran parte de columnas que se encontraban vacías (NaN) y con pérdidas que superan los 60% de datos perdidos. Seguido a este proceso se pasó a ver el comportamiento de los datos, las variables categóricas, transformándolas de variables a columnas y pasó después realizar el rellenado de datos y corrección de algunos registros. Habiendo ya concluido el proceso de limpieza de datos se seleccionó las columnas óptimas para hacer el estudio y así se desarrolló el modelo usando redes neuronales en las cuales se usaron tres capas. La primera con 60, segunda capa con 30 neuronas, la última capa con 3 neuronas que pertenecen a los tres tipos de riesgo. Al

modelo se le aplicaron las respectivas técnicas de evaluación para medir la precisión de este.




Para el desarrollo del modelo se hizo uso de las siguientes librerías:

- Scikit-Learn
- Keras
- Tensor Flow
- Pandas

## 2.1. Entendimiento del negocio

La Hipertensión Arterial Sistémica perjudica alrededor al 20% poblacional adulta de la más grande parte de las naciones es la primera causa de morbilidad y motiva el mayor número de consultas en las afecciones del artefacto circulatorio.

La Hipertensión Arterial es el componente de peligro de mayor relevancia para la patología cardio cerebrovascular, y constantemente se asocia con otros componentes de peligro bien conocidos como por ejemplo dieta, elevación de lípidos sanguíneos, obesidad, tabaquismo, Diabetes Mellitus e inacción física (sedentarismo).

Acudiendo a la reciente y constante necesidad del área de la salud para tomar decisiones correctas y en los menores tiempos posibles, se sugiere utilizar plataformas, arquitecturas o algoritmos que permitan desarrollar un modelo efectivo para optimizar el trabajo del médico durante el proceso de

atención de pacientes. A continuación, planteamos el desarrollo de un modelo de predicción de riesgos para enfermedades de presión alta. Asíí mismo, se plantea implementar técnicas de Deep Learning para realizar predicción del riesgo que puedan tener los pacientes de sufrir estas patologías. Con base en una predicción mediante el uso de variables clínicas, como, por ejemplo: la presión de la sangre, los niveles de azúcar en la sangre y niveles de oxígeno, edad, peso, estilos de vida.

## 2.2. Adquisición, entendimiento y exploración de los datos

Los datos para este proyecto fueron proporcionados por un hospital del departamento de Bolívar, el cual cuenta con base de datos de tipo SQL Server. Para el proceso de obtención de datos el primer paso fue el entendimiento de las estructuras de la base de datos. Para esto proporcionaron una base de datos de prueba, luego de conocerlas un poco pudimos realizar una consulta en lenguaje SQL, está consulta fue enviada al área de sistema del hospital en donde se validó y luego se realizó dicha consulta y los resultados obtenidos se enviaron en archivos con extensión CSV.

Se encontraron 188 columnas las cuales fueron tomadas de los registros clínicos relacionado con la hipertensión teniendo en cuenta datos como los diagnósticos, los signos vitales y demás datos de la persona como la edad, los antecedentes familiares, entre otros. Al tiempo también se realizó un análisis de datos para verificar los porcentajes de datos faltantes a través de las siguientes líneas de código:

```
for col in pacientes_df.columns:
    pct_missing = np.mean(pacientes_df[col].isnull())
    print('{} - {}%'.format(col, round(pct_missing*100)
))
```

El cual recorre y hace una búsqueda en todas las columnas, este arroja como resultado que el 9% de los campos presentan un porcentaje de perdida de dato mayor al 60%, siendo un total de 17 columnas. En la tabla 6 se puede ver a detalle el porcentaje de datos perdidos por tipo de datos de todos los campos.

**Tabla 6: Porcentaje de pérdida de datos por tipo de datos**

| | Cadena | | Numéricos | | Total # Campos | Total % Part. |
|---|---|---|---|---|---|---|
| **Rangos perdidos** | **# Campos** | **% Part.** | **# Campos** | **% Part.** | | |
| Perdida Alta | 3 | 1,61% | 14 | 7,53% | 17 | 9,14% |
| Perdida Media | 4 | 2,15% | 10 | 5,38% | 14 | 7,53% |
| Perdida Baja | 122 | 65,59% | 22 | 11,83% | 144 | 77,42% |
| Sin Perdida | 5 | 2,69% | 6 | 3,23% | 11 | 5,91% |
| **Total, general** | **134** | **72,04%** | **52** | **27,96%** | **186** | **100,00 %** |

Los datos que presentan alta perdida son los siguientes:

| Column | Non-Null Count | Dtype | % perdidos | Tipo Datos | Rango_perdidos |
|---|---|---|---|---|---|
| Ateroesclerosis | 0 non-null | float64 | 100,00% | numéricos | Perdida Alta |
| Ecografia_carotida | 0 non-null | float64 | 100,00% | numéricos | Perdida Alta |
| Ecocardiografia | 77 non-null | float64 | 99,20% | numéricos | Perdida Alta |
| EKG | 78 non-null | float64 | 99,20% | numéricos | Perdida Alta |
| Glicemia_ayunas | 82 non-null | float64 | 99,10% | numéricos | Perdida Alta |
| Creatinina.1 | 99 non-null | object | 98,90% | Cadena | Perdida Alta |
| Diabetes_mellitus_diagnosticada | 274 non-null | float64 | 97,00% | numéricos | Perdida Alta |
| LOB | 520 non-null | object | 94,30% | Cadena | Perdida Alta |

| | | | | | | |
|---|---|---|---|---|---|---|
| Riesgo_perimetro | 658 | non-null | float64 | 92,80% | numéricos | Perdida Alta |
| Condiciones_clinicas_asociadas | 722 | non-null | object | 92,10% | Cadena | Perdida Alta |
| 3_intervenciones_sobre_factore_riesgo_especÃfico | 890 | non-null | float64 | 90,30% | numéricos | Perdida Alta |
| Tabaquismo | 2378 | non-null | float64 | 74,10% | numéricos | Perdida Alta |
| Riesgo_edad | 2649 | non-null | float64 | 71,20% | numéricos | Perdida Alta |
| PrevenciÃ³n_cancer_prostata | 2710 | non-null | float64 | 70,50% | numéricos | Perdida Alta |
| Perimetro_abdominal | 2908 | non-null | float64 | 68,40% | numéricos | Perdida Alta |
| Violenciaintrafamiliar | 3413 | non-null | float64 | 62,90% | numéricos | Perdida Alta |
| Abusosexual | 3432 | non-null | float64 | 62,70% | numéricos | Perdida Alta |

## 2.3. Ingeniería de características

### 2.3.1. Selección

**Visualización de la estructura del Data Frame**

En primero lugar, se realizó un análisis de datos para verificar los porcentajes de datos faltantes, arrojando como resultado que el 9% de los campos presentan un porcentaje de perdida de datos mayores al 60%, siendo un total de 17 columnas. Teniendo esto en cuenta se procede a eliminar los campos con datos faltantes. A continuación, se muestra cuales fueron:

- Ecografia_carotida,
- Ecocardiografia
- EKG
- Glicemia_ayunas, Creatinina.1
- Diabetes_mellitus_diagnosticada
- LOB
- Riesgo_perimetro
- Condiciones_clinicas_asociadas
- 3_intervenciones_sobre_factore_riesgo_especÃ-fico
- Tabaquismo
- Riesgo_edad
- PrevenciÃ³n_cancer_prostata
- Perimetro_abdominal
- Violencia intrafamiliar
- Abuso sexual

En segundo lugar, se hicieron la representación de los datos en graficas de algunas columnas. Esto con el fin de saber cómo están distribuidos los datos. En la ilustración 3, vemos que hay mayor cantidad de pacientes subsidiados, a comparación de los no afiliados, contributivo y especial que no presentan muchos datos.

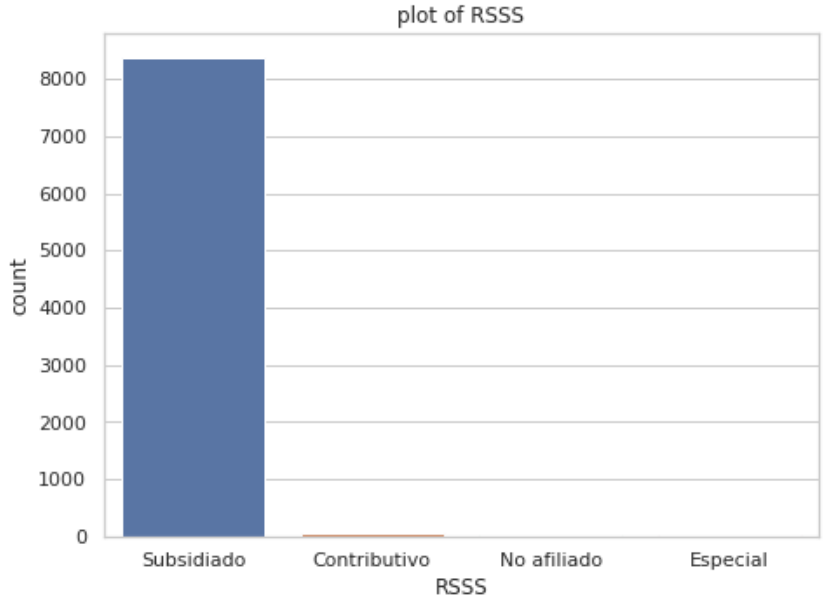



*Ilustración 3 Gráfica de RSSS*

Además, en la ilustración 4 se puede observar una distribución diferente donde los datos que más sobresalen o la gran mayoría de los datos encontrados pertenecen a pacientes solteros mientras que separadas presenta el menor número de datos.

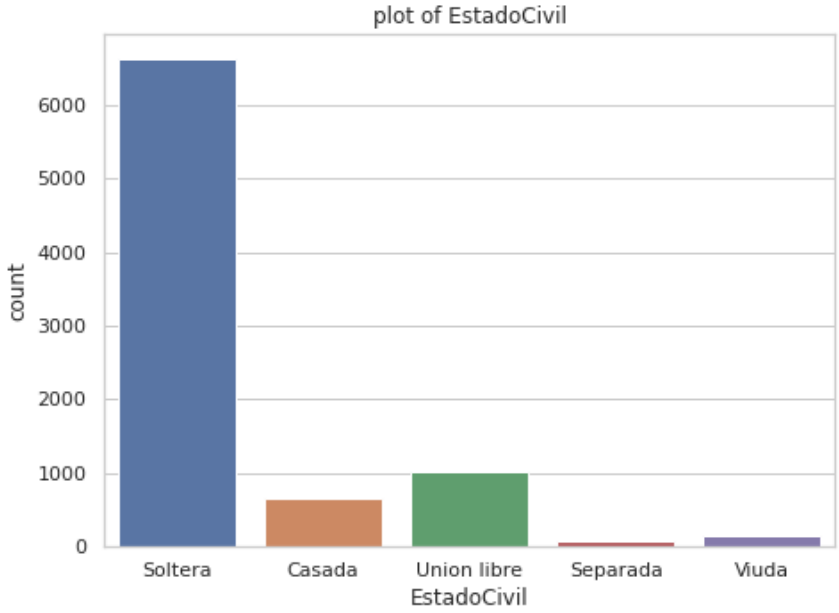



*Ilustración 4 Gráfica de Estado civil*

Por otro lado, la siguiente ilustración, la numero 5 presenta los datos de los pacientes que tienen actividad física haciendo una comparación de los pacientes que no la tiene, mostrando así que la mayoría de los datos encontrados son de pacientes que no tienen actividad física.

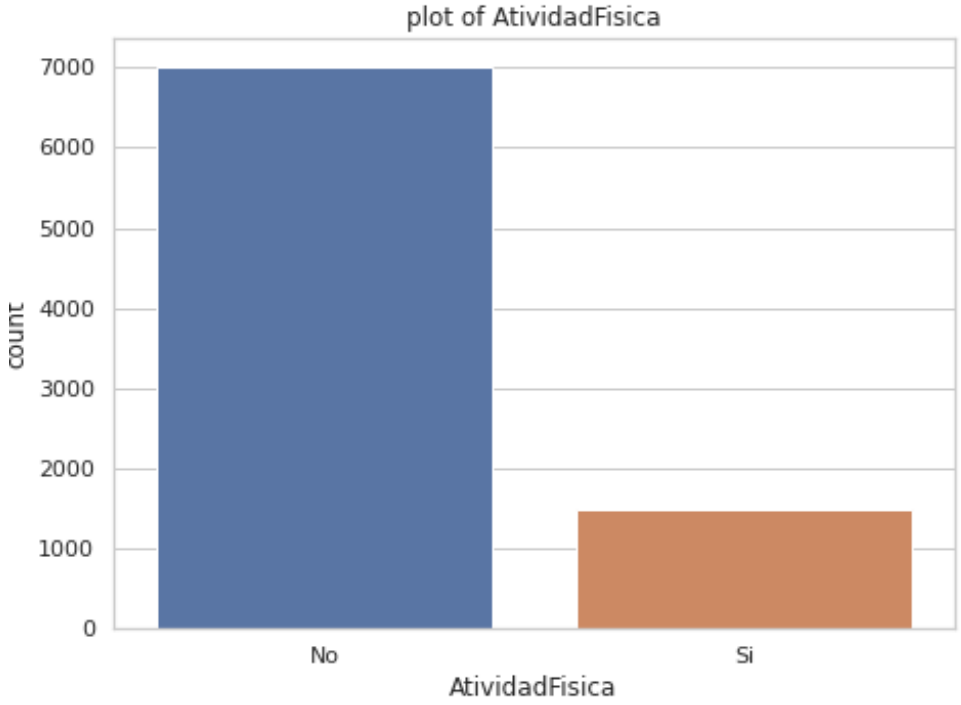



*Ilustración 5 Gráfica de Actividad física*

En la siguiente, la ilustración 6 muestra los datos de los pacientes con palpitaciones, mostrando así que la mayoría de los pacientes no las tiene.

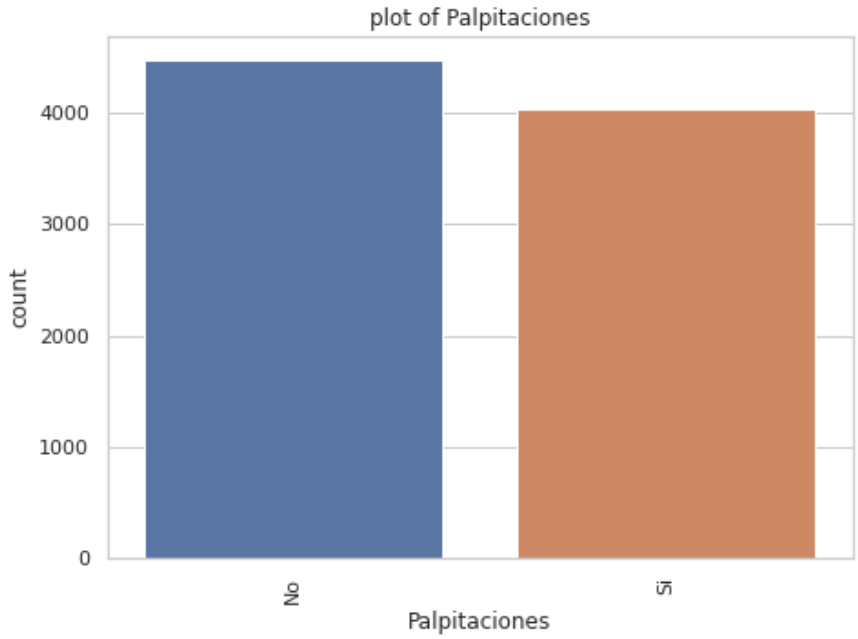



*Ilustración 6 Gráfica de pacientes con palpitaciones*

Así mismo, podemos ver la distribución de los datos en la ilustración 7, aquí se muestra el estado nutricional donde la mayoría de los pacientes tienen el peso adecuado (barra de color rojo), en segundo lugar están los que tienen sobrepeso (barra de color naranja), los siguientes son obesidad en primer grado (verde), variable no identificada y nombrada así como “selecciones” que representa una cantidad por debajo de los 1000 datos hallados pertenecientes a esa categoría, también se puede relacionar a datos vacíos o faltantes (color café), obesidad en segundo grado se encuentra por encima de los 500 datos (color azul), pacientes que presentan un peso bajo que está por debajo de los 500 datos (barra de color morado), y los dos últimos que son obesidad en tercer grado (color rosa) que está por encima de muy bajo peso (color gris).

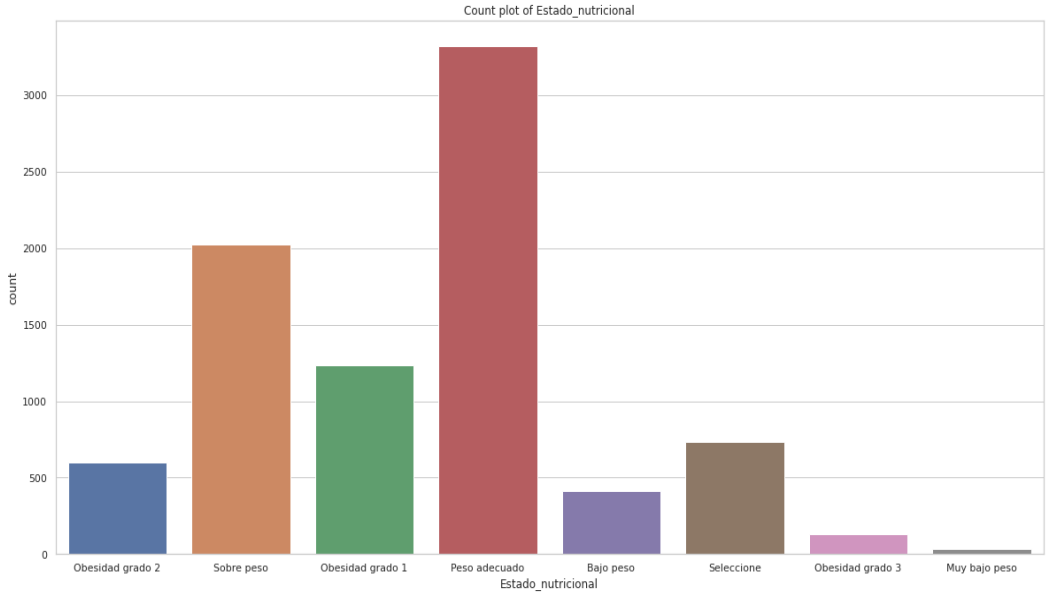



*Ilustración 7 Gráfica de estado nutricional*

De la misma manera se realizó la gráfica para los datos que pertenecen al diagnóstico principal, la ilustración 8 muestra los resultados.

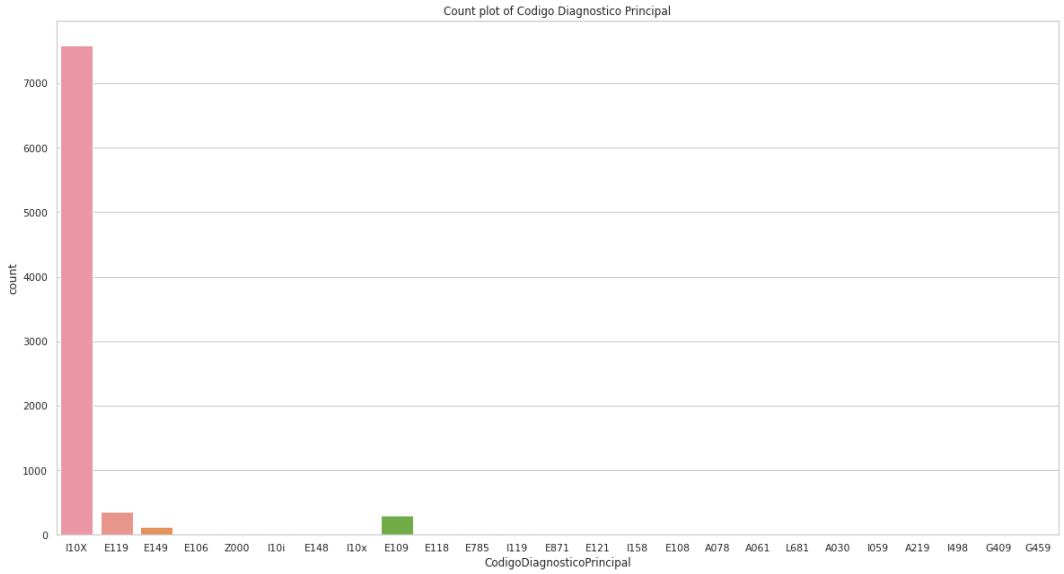



*Ilustración 8 Diagnostico principal*

En la ilustración anterior se puede observar que la mayoría de los datos del diagnóstico principal pertenecen a I10X (representado por la barra de color rosa), seguido a este están los datos de E119 (representado por el color zapote), los dos restantes que son E109 (color verde) y E149 (por el color naranja).

**Verificado y rellenado de datos faltantes**

Entes de hacer el rellenado en las columnas que presentaban datos faltantes, se realizó una representación gráfica para ver de forma visual que columnas presentan la mayor parte de datos faltantes. En la ilustración 9 se puede ver las columnas que presentan datos faltantes, estas se pueden diferenciar porque presentan mayor número de franjas de color amarillo.

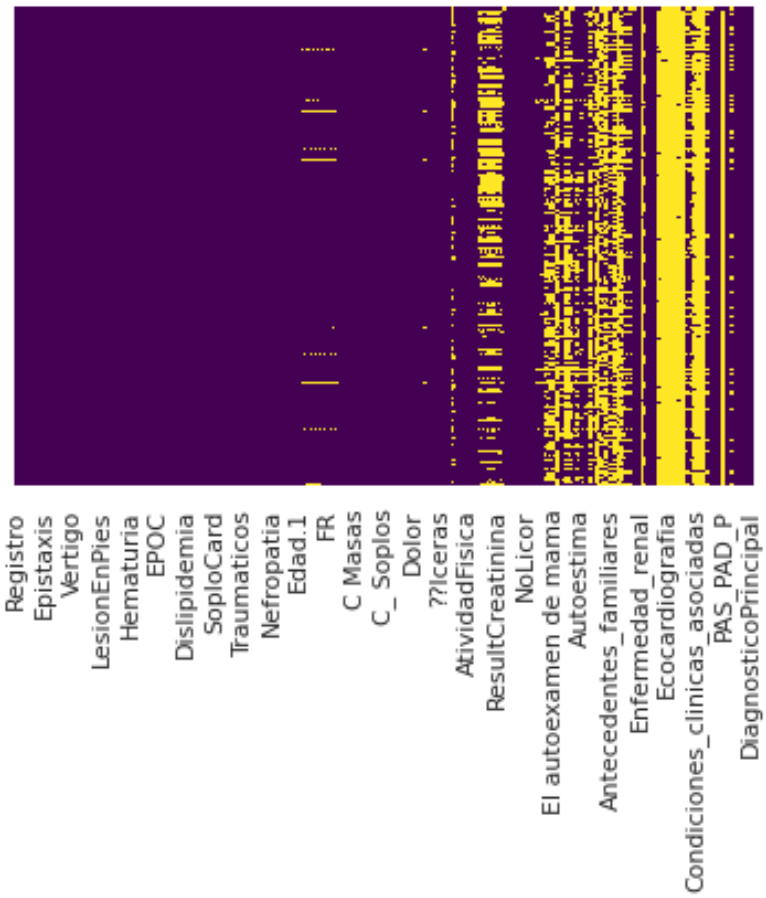



*Ilustración 9 Gráfica de datos faltantes*

A continuación, se realizó el llenado de datos faltantes en las columnas del DATASET, basados en los métodos interpolación lineal para aquellas variables de tipo numéricas y para las categorías el método PAD, se hizo las respectivas correcciones de igual manera las que presentaban errores en los datos, por ejemplo, en las columnas:

- **DiagnosticoPrincipal**: se encontraron espacios vacíos en algunos de los registros (NaN) y se rellenaron con: “NO ESPECIFICADA SIN MENCION DE COMPLICACION”.
- **PAS_PAD_T1_1:** error en algunos registros que mostraban “oct-80”, los cuales se reemplazaron por el numero “80”.
- **ResultCreatinina:** presentó errores en algunos registros, estos llevaban caracteres que no pertenecían al conjunto de datos y los cuales

presentaban dificultades a la hora de procesarlos: “1-0, 0-7, 1-0|, 1-T, 0’.7, .0.8”, estos datos se reemplazaron por: “1.0, 0.7, 1.0, 1.0, 0.7 y 0.8” respectivamente.

- **ResultLDL:** se reemplazaron los valores vacíos por “0”

Al finalizar, el rellenado de datos nuevamente se realizó una representación gráfica para ver si se efectuaron los cambios en las columnas que contenían datos faltantes. en la ilustración 10 se muestra que las columnas que antes aparecían con datos faltantes se han corregido.

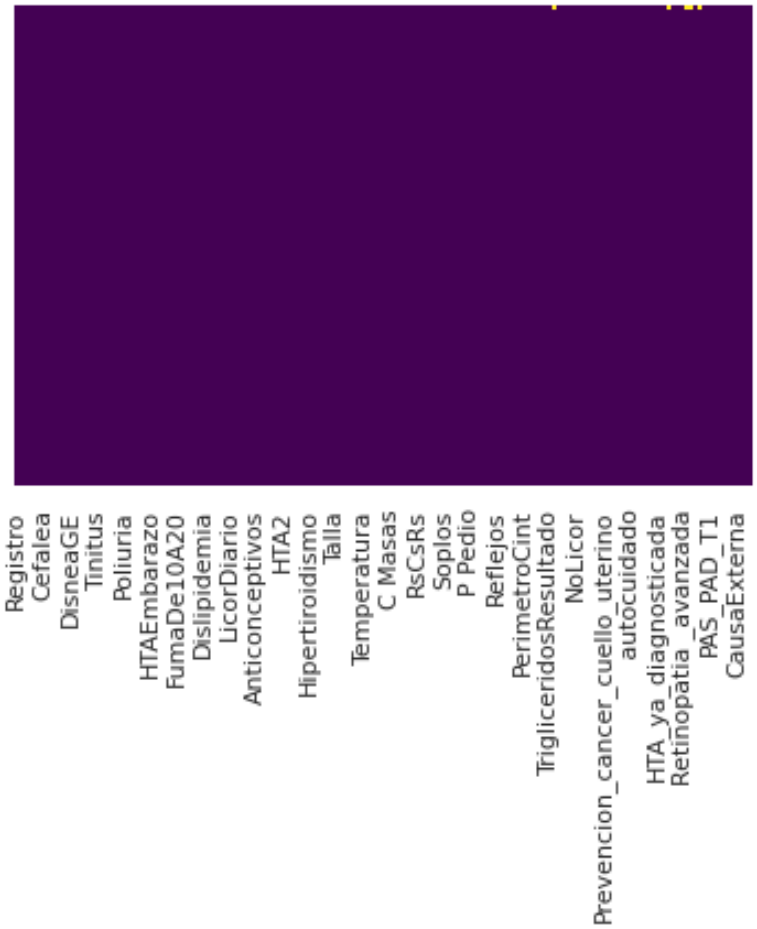



*Ilustración 10 Segunda gráfica de datos faltantes*

### Variables categóricas

Se hace una recopilación en esta sección de las variables categóricas y se muestran la cantidad de categorías que tienen cada una.

| Variable | Cantidad |
|---|---|
| RSSS | 4 |
| Estado civil | 5 |

| | |
|---|---|
| Reflejos | 3 |
| Resultados.1 | 3 |
| Diabetes_mellitus_(DM2) | 47 |
| TipoDiagnostico | 3 |
| CodigoDiagnosticoPrinciapal | 25 |
| DiagnosticoPrincipal | 23 |
| PAS_PAD_T2_1 | 26 |
| PAS_PAD_P_1 | 7 |
| PAS_PAS_P_2 | 7 |

Las columnas que presentaban dos categorías se le realizó la conversión a binario si=1 y no=0, una muestra de algunas columnas a las cuales se les efectuó dicha transformación se muestra en la ilustración 11.

| Cefalea | Epistaxis | DisneaME | Disuria | Lipotimia | Palpitaciones | DisneaGE | Edemas | Vertigo | Precordialgia |
|---|---|---|---|---|---|---|---|---|---|
| No | No | No | No | No | Si | Si | Si | No | No |
| Si | No | No | No | No | Si | No | No | Si | No |
| Si | No | No | No | No | Si | Si | No | Si | Si |
| Si | Si | No | No | No | Si | Si | No | No | No |
| Si | No | Si | No | No | Si | Si | Si | Si | Si |

*Ilustración 11 Muestra de algunas columnas categóricas*

Por otra parte, se convirtió los valores que contenía la variable estado_nutricional a numéricas, quedando de la siguiente manera:

| Peso | Valor |
|---|---|
| Muy bajo | 0 |
| Bajo | 1 |
| Adecuado | 2 |
| Sobre peso | 3 |
| Obesidad grado 1 | 4 |
| Obesidad grado 2 | 5 |
| Obesidad grado 3 | 6 |

**Implementación de Dummies a variables categóricas**

Se utiliza la función *get_dummies()* para la manipulación de datos. Convierte datos categóricos en variables ficticias o indicadoras, dando como

resultado que las categorías que antes pertenecían a variables sean convertidas a columnas. Si un paciente le corresponde una de estas nuevas columnas, este tomará el valor de 1, de lo contrario 0. La ilustración 12 da una muestra de las columnas.

| EstadoCivil_Casada | EstadoCivil_Separada | EstadoCivil_Soltera | EstadoCivil_Union libre | EstadoCivil_Viuda | CodigoDiagnosticoPrincipal_A030 |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |

*Ilustración 12 Dummies a variables categóricas*

**Verificación y análisis de variable de Respuesta**

Además, se realizó un análisis de la variable de respuesta (ClasificacionFinalRiesgoCardiovascular), la cual presentó un total de 8494 filas. La cantidad de categorías de la variable de respuesta es de 3, y estas son:

- Riesgo bajo.
- Riesgo moderado.
- Riesgo alto.

Estas categorías se clasificaron de la siguiente manera:

| Riesgo | Valor |
|---|---|
| Alto | 1 |
| Moderado | 2 |
| Bajo | 3 |

A continuación, se pasa graficar la variable de respuestas, ilustración 13. Donde se puede observar que a la mayoría de los datos se sitúan en el valor 2, el cual pertenece a riesgo moderado, el siguiente es el valor 3 al cual le pertenece a riesgo bajo y por último el valor 1 que le pertenece a riesgo alto.

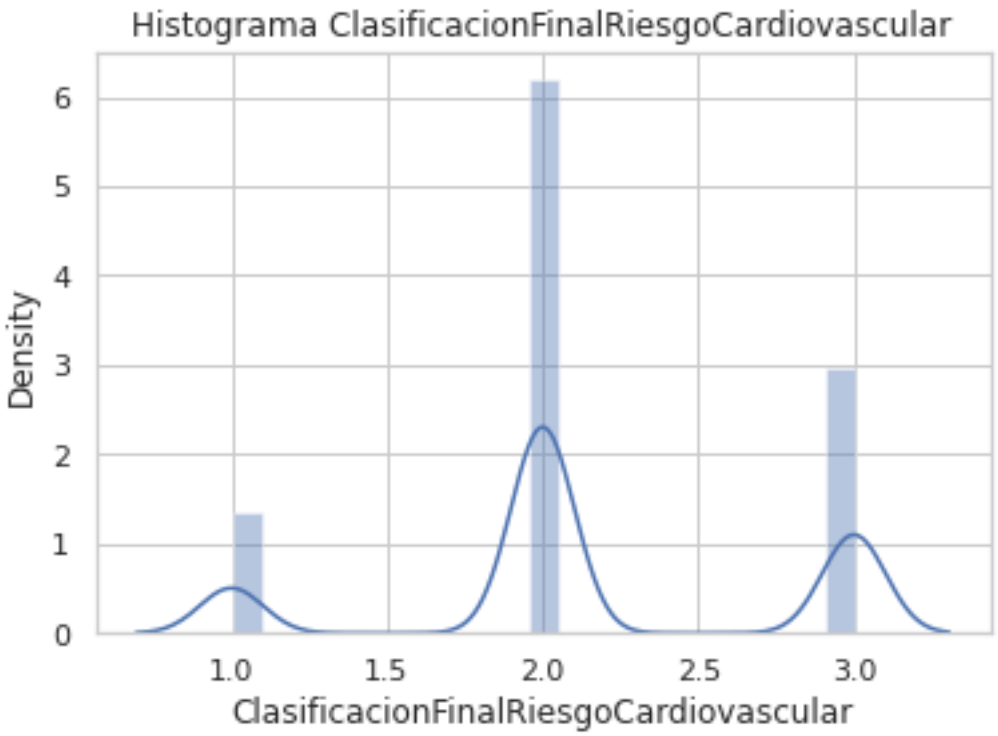



*Ilustración 13 valores de la variable de respuesta*

**Analizando matriz de Correlación entre variables**

Para poder graficar la matriz de correlación se necesitó dividir los datos a fracciones más pequeñas para poder graficar, obteniendo así 10 secciones de los datos agrupados en unidades más pequeñas. La ilustración 14 muestra la distribución de los datos en las 10 secciones.

```
[ ]  df_1 = pacientes_df_dummies.iloc[:,0:20]
     df_2 = pacientes_df_dummies.iloc[:21,:40]
     df_3 = pacientes_df_dummies.iloc[:,41:60]
     df_4 = pacientes_df_dummies.iloc[:,61:80]
     df_5 = pacientes_df_dummies.iloc[:,81:100]
     df_6 = pacientes_df_dummies.iloc[:,101:120]
     df_7 = pacientes_df_dummies.iloc[:,121:140]
     df_8 = pacientes_df_dummies.iloc[:,141:160]
     df_9 = pacientes_df_dummies.iloc[:,161:180]
     df_10 = pacientes_df_dummies.iloc[:,181:195]
```

*Ilustración 14 División de los datos*

Tal como se muestra en la ilustración, a las variables se les asigno una sección de los datos.

| Nombre | Rangos |
|---|---|
| df_1 | 0 hasta 20 |
| df_2 | 21 hasta 40 |
| df_3 | 41 hasta 60 |
| df_4 | 61 hasta 80 |
| df_5 | 81 hasta 100 |
| df_6 | 101 hasta 120 |
| df_7 | 121 hasta 140 |
| df_8 | 141 hasta 160 |
| df_9 | 161 hasta 180 |
| df_10 | 181 hasta 195 |

De esta manera, los datos ya agrupados en secciones más pequeñas se pueden graficar para así mostrar si hay correlación entre las variables. cabe destacar que las variables que, en este caso si tienen una correlación eso significa que las dos variables están directamente relacionadas (son representados en la gráfica por colores cercanos a blanco o más claros, los valores van de 0 a 1), de otra forma si se produce una correlación negativa es que están inversamente relacionadas (colores más oscuros, los valores van de 0 a -1), es decir, en las directamente relacionadas, si una aumenta la otra hará de igual forma, y si disminuye la otra lo hará de igual manera, en caso de las negativas, si una variable disminuye la otra aumenta, y si una aumenta la otra disminuye de igual medida. En caso de que la covarianza sea cero, la otra variable no se puede predecir, además de que si tienen el valor de 1 es que las variables tienen una correlación positiva perfecta. Las ilustraciones 15, 16 y 17 muestran los matices de correlación de la primera, segunda y tercera agrupación.

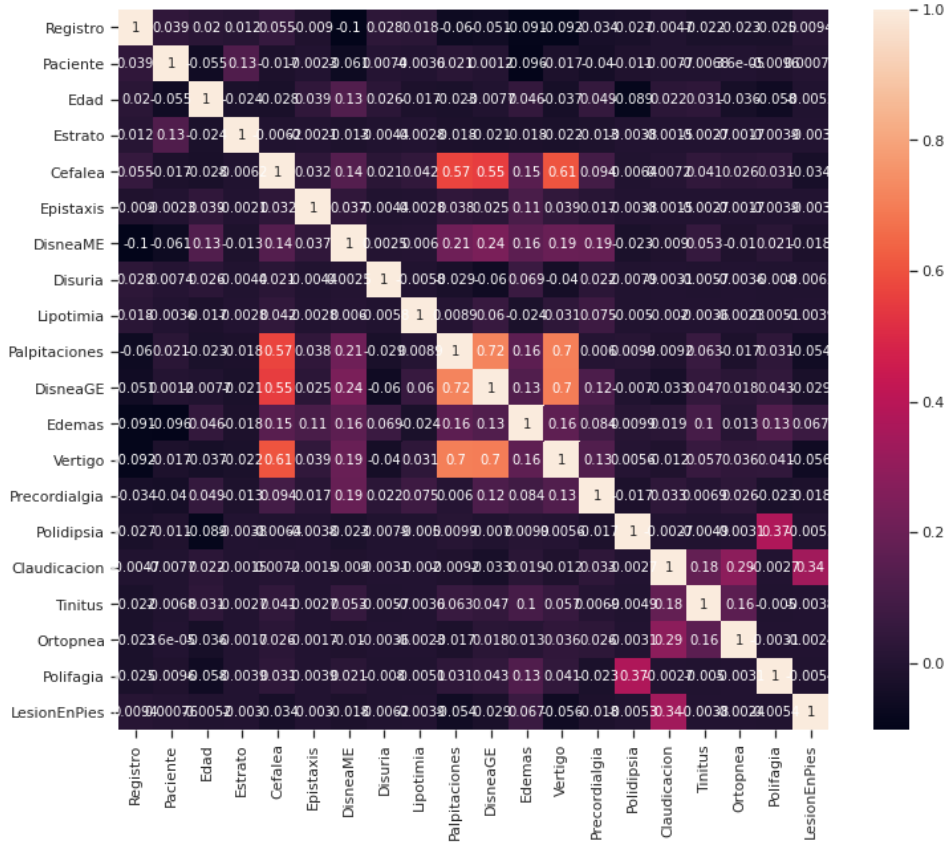



*Ilustración 15 Matriz de correlación, df_1*

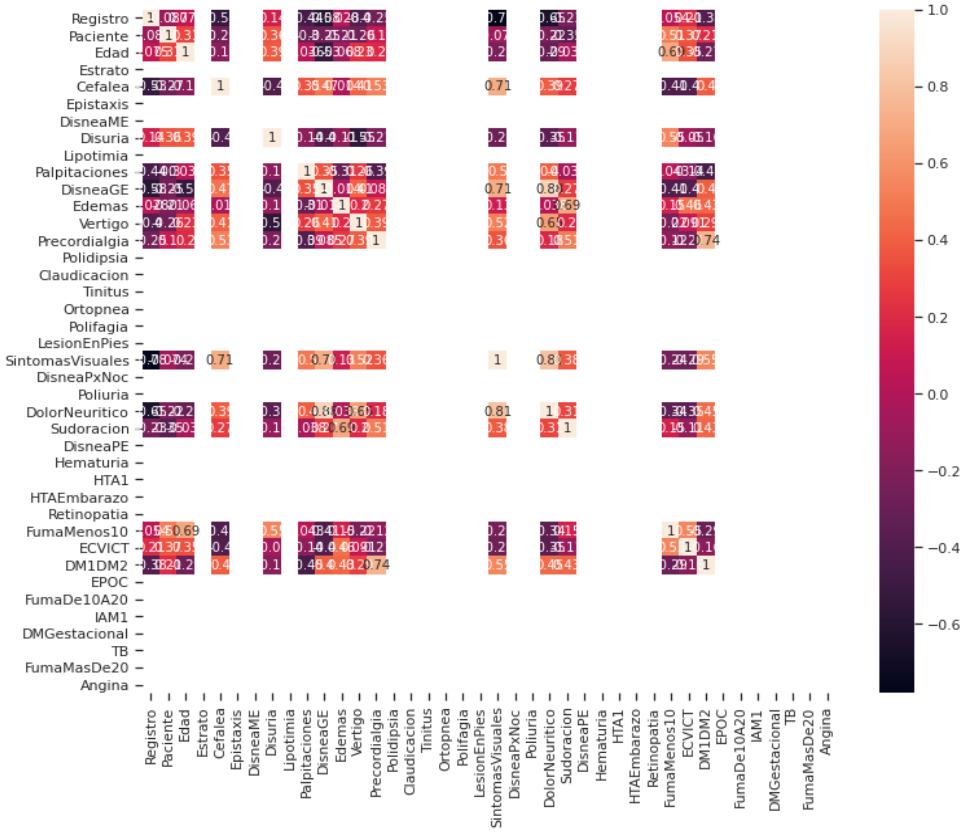



*Ilustración 16 Matriz de correlación, df_2*

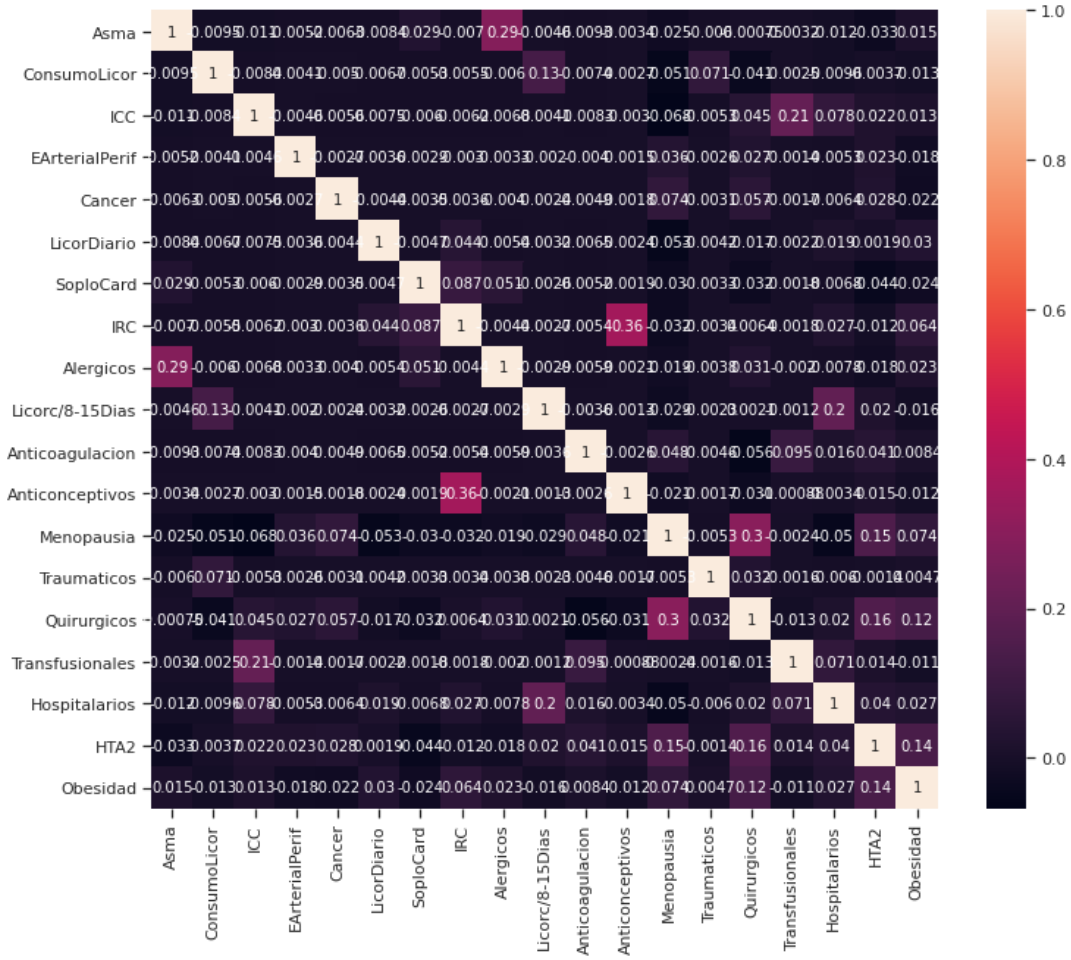



*Ilustración 17 Matriz de correlación, df_3*

### 2.3.2. Preparación de datos

Se implementó *Feature selection* en este proceso para ordena las características por el valor de alguna función de puntuación, que generalmente mide la relevancia de las variables para posteriormente usarse. En este paso se escogen las variables óptimas para usarse en el modelo de predicción. El numero óptimo de características que resultó del *Feature selection* fue de: 40. Con ese dato se pudieron sacar las variables que podían ser de ayuda para la realización del modelo, en la Ilustración 16 se muestran cuales fueron esas variables.

```
'Edad','SignosVitales','Antecedentes_familiares',
'IMC','TrigliceridosResultado','ResultColestTotal',
'Talla','Resultado','ResultHDL.1','DM1DM2','Pulso',
'Nefropatia_diabetica','CausaExterna','Edad.1','ResultLDL',
'Resultado.1','Peso','ResultCreatinina',
'Tension_arterial_sistolica','SMC','FC',
'Disfuncion_renal','ResultHDL','PAS_PAD_T2_2','RsCsRs',
'Dislipidemia.1','Edemas','FR',
'Tension_arterial_diastolica','Pulso.1',
'DietaBajaSal','PAS_PAD_T1_2',
'CodigoDiagnosticoPrincipal_I10X','ECVICT',
'Cefalea','Quirurgicos','ICC','PMI'])
```

*Ilustración 18 Variables optimas*

Posterior a este proceso se procede a eliminar las variables que poseen menor importancia, al tener las variables óptimas se procede a eliminar las demás, las cuales no presentaban mucha relevancia. A continuación, la Ilustración 19 muestras dichas variables.

```
#Eliminando las variables que no son optimas para el modelo
pacientes_df_dummies=pacientes_df_dummies.drop(['Paciente','P Pedio','Alergicos','Dislipidemia','ConsumoLicor','EstadoCivil_Separada',
'ReducPeso','Palpitaciones','DisneaME','EstadoCivil_Viuda','CodigoDiagnosticoPrincipal_E149','Disuria','Deficitpulso','Epistaxis','Claudicacion',
'Tinitus','ConsumeVerdFrutas','Estrato','CodigoDiagnosticoPrincipal_G459','Polifagia','DMGestacional','CodigoDiagnosticoPrincipal_E871','Polidipsia',
'Cancer','EstadoCivil_Soltera','DisneaPxNoc','Licorc/8-15Dias','Soplos','1_indicaciones_sobre_sus_factores_riesgo_predisponente','NoAzucares',
'CodigoDiagnosticoPrincipal_G409','Lipotimia','IRC','LesionEnPies','Obesidad','CodigoDiagnosticoPrincipal_I059','Anticonceptivos','Hematuria','LicorDiario',
'Poliuria','Papiledema','TB','Amputaciones','FumaMenos10','Hipertiroidismo','??lceras','Ortopnea','EPOC','DisneaPE','Hemorragias',
'CodigoDiagnosticoPrincipal_E785','PAS_PAD_P_2','Prevencion_enfermedades_transmisi??n_sexual','Traumaticos','2_indicaciones_control_medicina_general_5_a?±os',
'HTAEmbarazo','Anticoagulacion','PerimetroCint','HTA_ya_diagnosticada','Estilosde_vida_saludable','Transfusionales','Autoestima',
'Prevencion_cancer_cuello_uterino','Resultado.2','Temperatura','Nutrici??n_y_alimentaci??n','Micro_albuminuria','Complicaciones _drogas_psicoactivas',
'El autoexamen de mama','Como_prevenir_diabetes_hipertensi??n_osteoporosis','I_Yugular','Consecuencias_del_consumo_alcohol_cigarrillo',' autocuidado',
'Glicemia_postprandial','IAM2','EnfVascular','4_citologia_cervico_uterina','Nefropatia','Soplocarotideo','CodigoDiagnosticoPrincipal_A078','C Masas',
'Displidemia_(cualquiera)','Retinopatia _avanzada','DM1DM2.1','Nodulotiroideo','Enfermedad_arterial_periferica','Fondo_ojo','Enfermedad_renal',
'CodigoDiagnosticoPrincipal_E121','EArterialPerif','Se_realizo','CodigoDiagnosticoPrincipal_E148','Cardiopulmonar','CodigoDiagnosticoPrincipal_A030',
'CodigoDiagnosticoPrincipal_E119','Creatinina','CodigoDiagnosticoPrincipal_A061','CodigoDiagnosticoPrincipal_E106','CodigoDiagnosticoPrincipal_A219',
'CodigoDiagnosticoPrincipal_E118','CodigoDiagnosticoPrincipal_E109','CodigoDiagnosticoPrincipal_E108','Crueces_AV','CodigoDiagnosticoPrincipal_I10i',
'Abdomen','CodigoDiagnosticoPrincipal_I10x','Masas','CodigoDiagnosticoPrincipal_I119','CodigoDiagnosticoPrincipal_I158','CodigoDiagnosticoPrincipal_I498',
'Megalias','CodigoDiagnosticoPrincipal_L681','CodigoDiagnosticoPrincipal_Z000','Extremidades','Exudado','EstadoCivil_Union libre','Estado_nutricional',
'HTA2','FumaDe10A20','Vertigo','AtividadFisica','NoLicor','IAM1','Angina','Sudoracion','Hospitalarios','FinalidadConsulta','Sensibilidad',
'PredominioIngestaDeGrasa','Retinopatia','HTA1','UsarEdulc','Dolor','C_ Soplos','Diabetes_mellitus_(DM2)_1','Hipotiroidismo','UtilizaSalero','Menopausia',
'DisneaGE','PAS_PAD_T1_1','Diabetes_mellitus_(DM2)_2','DolorNeuritico','Asma','Reflejos','EstadoCivil_Casada','InterconsultaNutricionista','NoFumar',
'DisminuGrasa','HacerEjer','PAS_PAD_T2_1','EnfCoronaria','SoploCard','Precordialgia','FumaMasDe20','SintomasVisuales','Registro'
],axis=1)
```

*Ilustración 19 Variables con poca relevancia*

**Balance del Dataframe**

Al finalizar el paso anterior, se obtiene un conjunto de datos con 8396 filas y 39 columnas. A continuación, la cantidad de registros de cada categoría.

| Nombre | Cantidad |
|---|---|
| Riesgo 1 | 1082 |
| Riesgo 2 | 4951 |
| Riesgo 3 | 2363 |

**Balanceo de Clases**

Teniendo en cuenta de que los datos estaban desequilibrados, se realiza el procedimiento para corregir este problema, para ello se utiliza el método SMOTE. Este método funciona de la siguiente manera: selecciona primero una instancia de clase minoritaria a al azar y encuentra sus vecinos de clase minoritaria más cercanos. A continuación, la instancia sintética se crea escogiendo uno de los vecinos k más cercanos b al azar y conectando a y b para conformar un segmento de línea en el espacio de entidades. Las instancias sintéticas se crean como una mezcla convexa de ambas instancias elegidas a y b.[20]

De esta manera, al hacer el balanceo de los datos los registros quedaron de esta manera:

| Nombre | Cantidad |
|---|---|
| Riesgo 1 | 4951 |
| Riesgo 2 | 4951 |
| Riesgo 3 | 4951 |

**Normalización de datos**

Teniendo los resultados del balance de los datos se procede a hacer la normalización del Dataframe, para esto se utiliza el método *StandardScaler*, este método permite estandarizar los datos, lo cual útil para lo que son negativos.
Se organizan los datos en una distribución estándar. Para hacer la normalización se dividió el Dataframe en dos partes dos variables diferentes). La variable x como la variable independiente contiene una sección de los datos a normalizar. Mientras que la variable Y como la variable dependiente contienen la columna de respuesta (ClasificacionFinalRiesgoCardiovascular). Luego de aplicar el método de estandarización en la variable independiente (x) se puede observar en la

ilustración 20 algunas de las columnas cuyos datos se le aplico la normalización.

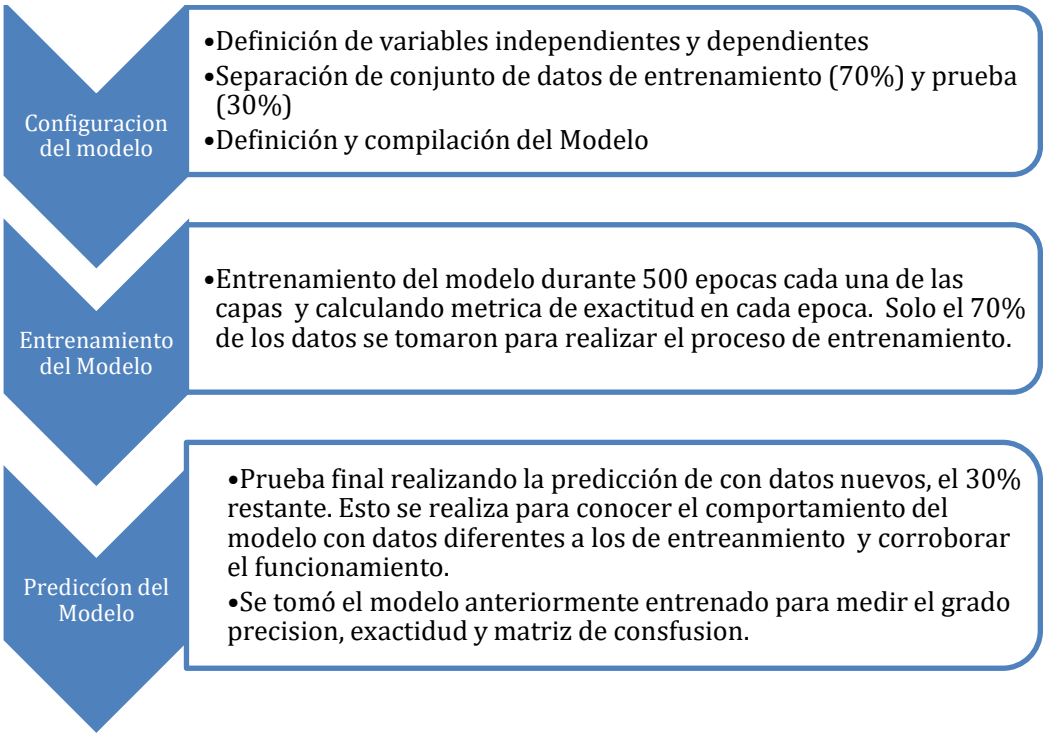
| | Edad | SignosVitales | Antecedentes_familiares | IMC | TrigliceridosResultado | ResultColestTotal |
|---|---|---|---|---|---|---|
| **0** | 0.216020 | 0.128713 | 0.440305 | -0.031218 | -0.069531 | -0.021812 |
| **1** | 0.190701 | 0.129540 | 0.443134 | -0.031418 | -0.069978 | -0.021952 |
| **2** | -0.009944 | 0.134162 | -0.067714 | -0.032539 | -0.072475 | -0.022735 |
| **3** | 0.170895 | -0.260553 | -0.074172 | -0.035643 | -0.079387 | -0.024903 |
| **4** | 0.111372 | 0.130467 | -0.065850 | -0.031643 | -0.070479 | -0.022109 |

*Ilustración 20 Estandarización de datos*

## 3. Construcción del modelo

Con el objetivo de tener un primer acercamiento a la fase de definición y construcción del modelo de predicción, se implementó un primer modelo de nivel menos avanzado y complejo, *clasificación con árbol de decisión.*

Posteriormente y basándose en los buenos resultados que arrojo este primer modelo se inició y puso en marcha el desarrollo, parametrización, entrenamiento y evaluación del modelo principal, el cual viene dado por una *Red Neuronal,* utilizando la librería de *Keras* bajo *Tensor Flow* de *Google.*




### 3.1. Modelo de Árbol de decisión utilizando Sklearn

El modelo de árbol de decisión permitió conocer como era el comportamiento de los datos preparados cuando se aplicaba una técnica de para la predicción del Riesgo Cardiovascular. A continuación, se detalla los parámetros para la construcción de este:

### Definición de variables independientes y dependientes

La variable *X* representa el conjunto de atributos sin tener en cuenta la variable de respuesta. Mientas la variable *y* contiene solo los datos de la variable de respuesta.

```
X = pacientes_df_normalizado.drop("ClasificacionFinalRiesgoCardiovascular",1)
y = pacientes_df_normalizado["ClasificacionFinalRiesgoCardiovascular"]
```

### Separación de conjunto de datos de entrenamiento y prueba

Se muestra la separación del conjunto de datos para el entrenamiento (70%) y el conjunto de datos de prueba (30%).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

### Definición de algoritmo de árbol de decisión

El método para generar el algoritmo de árbol de decisión de denomina DecisionTreeClassifier el cual recibe el parámetro de criterion y establecido en entropy, que se define como la función que permite conocer la calidad de una división.

```
algoritmo = DecisionTreeClassifier(criterion = 'entropy')
```

### Entrenando el modelo

Se ingresa la función *fit* que recibe dos parámetros: variable independiente y la variable dependiente destinadas al proceso de entrenamiento inicial del modelo de clasificación.

```
algoritmo.fit(X_train, y_train)
```

**Se realiza la predicción**

Posterior a través del método *predict* se le envíe el conjunto de datos independientes de X, cuyo resultado es guardado en otra variable y_pred.

```
y_pred = algoritmo.predict(X_test)
```

**Evaluación del modelo**

Por último, se calcula una matriz de confusión con la función *confusion_matrix*, que recibe el conjunto de respuestas esperadas vs las predichas por el modelo, a la precisión que está arrojando el modelo entrenado, en la ilustración 21 se muestra los valores que resultaron.

```
#Matriz de confusion
matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión:')
print(matriz)

#Calculo la precisión del modelo
precision = precision_score(y_test, y_pred, average='micro')
print('Precisión del modelo:')
print(precision)
```

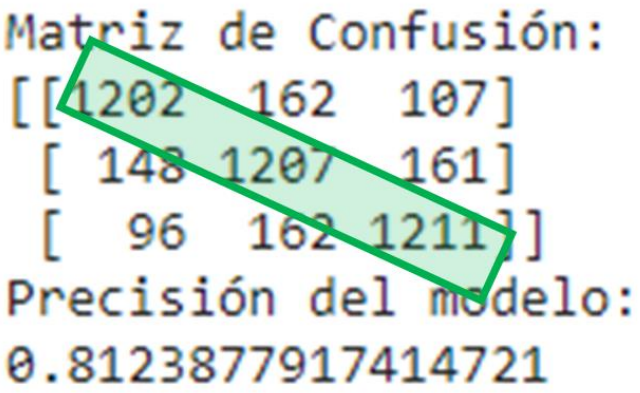



*Ilustración 21 Resultados Matriz de confusión y precisión del modelo de árbol de decisión*

Los datos arrojados como resaltados e marcados en verde en la ilustración 21, en la diagonal principal, indican que fueron predichos correctamente***: 3.620.*** Mientras aquellos que están por fuera de esta sección verde representan predicción incorrecta: ***836*** y finalmente la precisión fue del ***81,23%*** aproximadamente.

## 3.2. Red Neuronal en Keras de Tensor Flow

La librería de Keras bajo Tensor Flow proporciona la posibilidad de diseñar, crear, compilar, entrenar y evaluar una red neuronal con un altísimo nivel de estabilidad y versatilidad, siendo muy útil a la hora de resolver problemas de clasificación.

En lo que requiere al modelo se definió como el núcleo principal de este la red neuronal. Para determinar los mejores parámetros que satisfagan los objetivos de la predicción, se utilizó una técnica que permite encontrar esos mejores parámetros dentro de modelo de este tipo.

### Definición de variables independientes y dependientes

La variable X representa el conjunto de atributos sin tener en cuenta la variable de respuesta. Mientas la variable y contiene solo los datos de la variable de respuesta.

```
X = pacientes_df_normalizado.drop("ClasificacionFinalRiesgoCardiovascular",1)
y = pacientes_df_normalizado["ClasificacionFinalRiesgoCardiovascular"]
```

### Separación de conjunto de datos de entrenamiento y prueba

Se muestra la separación del conjunto de datos para el entrenamiento (70%) y el conjunto de datos de prueba (30%).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

### Definición del modelo

Las redes neuronales en la unión de varias neuronas entre sí agrupadas por capas. El modelo fue construido con tres capas, donde la primera representa los datos de entradas, conformada por 60 neuronas y una

función de activación igual a *relu.* La siguiente parte viene siendo la capa oculta quien cuenta con 30 neuronas y función de activación similar a la anterior capa, *relu*, y por último la de salida posee tres neuronas multiconectadas como las anteriores, tiene las tres respectivas neuronas y los datos pasan por una función sigmoidea para el cálculo pertinente.

Luego se realiza la compilación en Tensor Flow, recibiendo como parámetros una función de costo de *categorical_crossentropy*, un optimizador de *Adam* y por último la métrica *accuracy.*

```
def c_model():
    model = Sequential()
    model.add(Dense(60, activation='relu'
    model.add(Dense(30, activation='relu'))
    model.add(Dense(3, activation='sigmoid'
    model.compile(loss='categorical_crossentropy', opti
mizer='Adam', metrics=['accuracy'])
    return model
```

Configuración del modelo y control de épocas

A continuación, se asignan el modelo parametrizado anteriormente y el número de épocas que va a ejecutarse este. Estas épocas permiten en aplicar la propagación hacia atrás la cual consiste en encontrar las mejores ponderaciones y sesgos de entrada para obtener un resultado más preciso o mermar la pérdida.

```
model = KerasClassifier(build_fn=c_model, epochs=500)
```

**Entrenamiento de la red neuronal**

Gracias a la función *fit* que recibe dos parámetros: variable independiente y la variable dependiente destinadas al proceso de entrenamiento inicial de la red neuronal.

```
model.fit(X_train, y_train)
```

Al finalizar las 500 del ciclo de épocas el modelo en la fase de entrenamiento llega al umbral sostenido de precisión superior al 94% y pérdida del 0.0619.

### Realización de Predicción

El siguiente paso es la aplicación de la predicción, utilizando la función *predict* la cual recibe el conjunto de variables independientes definidas anteriormente para pruebas.

```
y_pred = model.predict(X_test)
predictions = [float(round(X_test)) for X_test in y_pred]
```

### Evaluación del modelo

En este punto se aplican técnicas para conocer el grado de afinidad y coherencia del modelo con las expectativas reales. A través de una matriz de confusión con la función *confusion_matrix,* que recibe el conjunto de respuestas esperadas vs las predichas por el modelo, la precisión que está arrojando el modelo entrenado, en la ilustración 21 se muestra los valores que resultaron.

```
#Matriz de confusion
matriz = confusion_matrix(y_test, y_pred)
print('Matriz de Confusión:')
print(matriz)

#Calculo la precisión del modelo
precision = precision_score(y_test, y_pred, average='micro')
print('Precisión del modelo:')
print(precision)
```

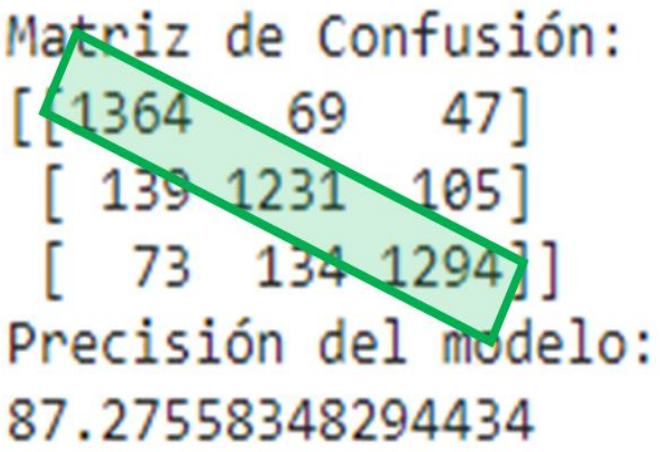



*Ilustración 22 Resultados Matriz de confusión y precisión del modelo redes neuronales*

Los datos arrojados como resaltados e marcados en verde en la ilustración 22, en la diagonal principal, indican que fueron predichos correctamente***: 3.889.*** Mientras aquellos que están por fuera de esta sección verde representan predicción incorrecta: ***567*** y finalmente la precisión fue del ***87.27%*** aproximadamente.

Ademas, se le aplicó al algoritmo otras tecnicas de evaluacion, la primera es la precisión, que con este se puede medir la calidad del modelo. En los términos más simples, Precisión es la relación entre los Positivos Verdaderos y todos los Positivos. Para la declaración de problemas, esa sería la medida de los pacientes que identificamos correctamente tener una enfermedad de todos los pacientes que realmente la tienen. La fórmula matemática es la siguiente:

$$Precision = \frac{True\ Positive(TP)}{True\ positive(TP) + False\ Positive(FP)}$$

Con lo anterior, el modelo arroja una precisión en la clasificación de los riesgos. Se pueden ver los resultados obtenidos en la ilustración 23.

| **Precisión** | |
|---|---|
| 1 | 87% |
| 2 | 86% |
| 3 | 89% |

Así mismo, se implementó la técnica de *Recall,* esta es la medida del modelo que identifica correctamente los True Positives. Por lo tanto, para todos los pacientes que realmente tienen la enfermedad, el *Recall* muestra

cuántos identificamos correctamente que pertenecen a su clasificación. La fórmula matemática para el *Recall* es la siguiente:

$$Recall = \frac{True\ Positive(TP)}{True\ positive(TP) + False\ Negative(FN)}$$

De esta manera, el modelo arroja el *Recall* en la clasificación de los riesgos. Se pueden ver los resultados obtenidos en la ilustración 23.

| **Recall** | |
|---|---|
| 1 | 92% |
| 2 | 83% |
| 3 | 86% |

Por otra parte, esta *F1-Score* el cual combina las medidas de la precisión y *Recall* para así devolver una medida de calidad más general del modelo, la fórmula matemática para calcular el *F1-Score* es la siguiente:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Con esto, los resultados del *F1-Score* de la clasificación de los riesgos se pueden ver en la ilustración 23.

| **F1-Score** | |
|---|---|
| 1 | 89% |
| 2 | 85% |
| 3 | 88% |

Por último, *Accuracy,* el cual mide el porcentaje de casos que el modelo ha acertado correctamente. La fórmula empelada para esta técnica es la siguiente:

$$Accuracy = \frac{True\ Positive(TP) + True\ Negative(TN)}{True\ positive(TP) + False\ Negative(FN) + False\ Positive(FP) + False\ Negative(FN)}$$

Como resultado, la exactitud del modelo (*Accuracy)* es de 87%. Este resultado se muestra en la ilustración 23.

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 1 | 0.87 | 0.92 | 0.89 | 1480 |
| 2 | 0.86 | 0.83 | 0.85 | 1475 |
| 3 | 0.89 | 0.86 | 0.88 | 1501 |
| accuracy | | | 0.87 | 4456 |
| macro avg | 0.87 | 0.87 | 0.87 | 4456 |
| weighted avg | 0.87 | 0.87 | 0.87 | 4456 |

*Ilustración 23 técnicas de evaluación del modelo*

# 4. RESULTADOS Y DISCUSIONES

.

El modelo predicción para clasificar el riesgo cardiovascular en los pacientes del hospital municipal de Arjona, paso por todas las etapas planteadas inicialmente. Con los datos proporcionados se realizó una limpieza rigurosa que incluyó; conocimiento de los datos, rellenados de datos faltantes, balanceo de clases, análisis estadístico y normalización general de estos.

Posteriormente se aplicó un modelo de clasificación con el algoritmo de árboles de decisión, librería S. Learn, el cual dio un valor de precisión superior al 82%. Para el desarrollo del modelo de predicción final se utilizó la implementación de redes neuronales con la ayuda de Keras – Tensor Flow.

Comparando el resultado obtenidos luego de la construcción del modelo de predicción para clasificación de riesgo cardiovascular, tenemos que este finalmente proporcionó un grado de satisfacción correcto frente a los resultados que obtuvieron los investigadores e instituciones reseñadas en el estado del arte del presente proyecto.

En comparación con otros estudios, el proyecto realizado por la facultad de farmacia, por la universidad Complutense en Madrid. Realizaron el proyecto de la hipertensión arterial: importancia de su prevención. Se especifica el tratamiento farmacológico para el tratamiento de la HTA, esto con el fin de controlar la presión arterial del paciente y más a largo plazo reducir la morbimortalidad, fundamentalmente de las enfermedades cardiovasculares, cerebrovasculares y renales asociadas a la HTA. El método de estudio que se implementa para la obtención del resultado fue por medio de encuestas que se les hacían a los pacientes, las variables de estudio para esta investigación son las siguientes:

- herencia genética
- sexo
- edad
- raza
- obesidad
- sedentarismo
- dieta
- alcohol
- tabaco
- estimulantes

- estrés

Teniendo en cuenta lo anterior, la investigación dio como resultado la clasificación de los pacientes entre rangos de edades, entre los mas propensos están los pacientes en el rango de 66 a 85 años. Además, hace la clasificación de los efectos de riesgo sobre las edades de los pacientes, la ilustración 24 muestra los resultados de la investigación.

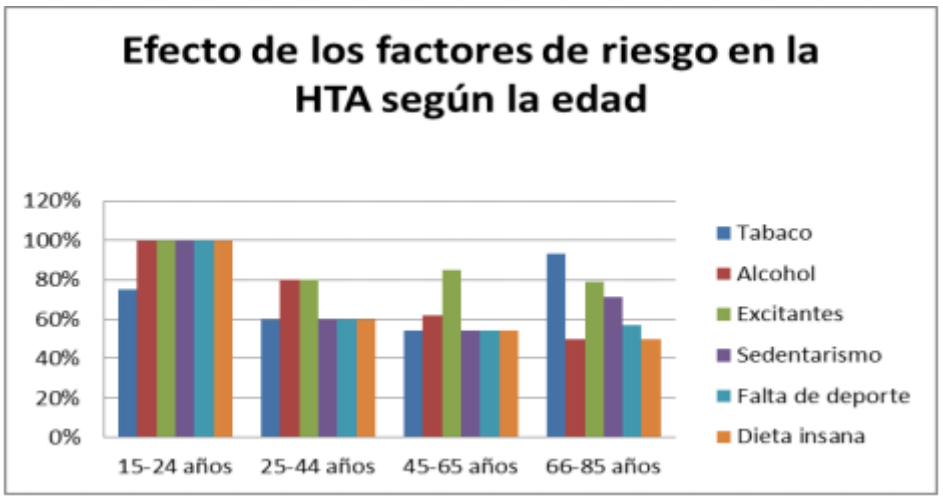



*Ilustración 24 Resultados de la investigación realizada por la facultad de farmacia por la universidad Complutense*

Sabiendo esto, se puede decir que la diferencia que existe entre las dos investigaciones es que, los resultados obtenidos de la investigación de la universidad de Complutense se basa en clasificar los riesgos de la enfermedad dividiendo los datos en rangos de edades y así mostrar cuanto es el porcentaje de las variables influyentes con respecto a la edad del paciente, mientras que este proyecto se basa en hacer la clasificación de los pacientes teniendo en cuenta de igual medida variables influyentes en la enfermedad para así clasificarlos entre los tres tipos de riesgos existentes. Alto, moderado y bajo, como antes se mencionó.

Además, el proyecto de Árbol para predecir el desarrollo de la cardiopatía hipertensiva, elaborado por el Hospital General Universitario "Carlos Manuel de Céspedes". Bayamo, Cuba. En el cual se realizó primeramente la adquisición de los datos por medio de la minería de datos (Data Mining-DM), en pacientes hipertensos. Establecieron para dicha investigación los siguientes criterios:

- Pacientes hipertensos de 20 años o más de edad.
- Pacientes con antecedente patológico personal de HTA esencial de 10 años o más, tiempo en que es más frecuente el daño orgánico.

En la ilustración 25 se puede ver las variables de estudio de la investigación.

Caracterización de la muestra. Variables cualitativas

| Variables | Categoría | No. | (%) |
|---|---|---|---|
| Edad (dicotómica) | ≤ 65 años | 407 | 33,92 |
| | > 65 años | 793 | 66,08 |
| Sexo | Masculino | 609 | 50,75 |
| | Femenino | 591 | 49,25 |
| Estadio de la HTA | Estadio 1 | 724 | 60,33 |
| | Estadio 2 | 476 | 39,67 |
| Control de la HTA | Controlado | 720 | 60,00 |
| | No controlado | 480 | 40,00 |
| Microalbuminuria | Sí | 421 | 35,08 |
| | No | 779 | 64,92 |
| Hábito de fumar | Sí | 511 | 42,58 |
| | No | 689 | 57,42 |
| Alcoholismo | Sí | 371 | 30,92 |
| | No | 829 | 69,08 |
| Obesidad | Sí | 445 | 37,08 |
| | No | 755 | 62,92 |
| Sedentarismo | Sí | 573 | 47,75 |
| | No | 627 | 52,25 |
| Exceso de sal | Sí | 467 | 38,92 |
| | No | 733 | 61,08 |

*Ilustración 25 Variables categóricas o cualitativas de la investigación realizada por el Hospital General Universitario "Carlos Manuel de Céspedes". Bayamo, Cuba.*

Además de los criterios que se descartaron, son los siguientes:

- Adolescentes con HTA esencial (por la poca frecuencia de daño a órganos diana).
- Paciente con HTA esencial de menos de 10 años de evolución.
- Pacientes con cardiopatía isquémica, a pesar de su elevada frecuencia en el hipertenso, donde el rol de la HTA, aunque evidente, no sería el único factor influyente en su aparición (en la presente investigación se evaluaron los efectos directos de la HTA, por lo que la inclusión de esta forma clínica puede inducir a sesgos de selección y clasificación).
- Enfermos con trastornos de la conducción interventricular y auriculoventricular.

- Pacientes que no padecieran otros estados mórbidos que pudieran provocar la cardiopatía.

Los resultados obtenidos de esta investigación, usando la técnica de árbol de decisión, este predijo el riesgo a desarrollar la enfermedad es de 82.598% de los pacientes, además arrojó en la curva de ROC de 86% y la tasa de verdaderos positivos en un 73.3% y de 92.1% para las clases 1 y 2.

En vista de los datos obtenidos, además de su coincidencia en algunas variables de estudio con la anterior investigación, posterior a la limpieza y el rellenado de datos para el modelo de predicción para clasificación de riesgo cardiovascular. Sé implementaron dos técnicas para hacer las predicciones, la primera, árbol de decisión que tuvo una la precisión fue del 81,23% aproximadamente al clasificar los tres tipos de riesgos. Alto, moderado y bajo, los cuales tomaron los valores de: 1, 2 y 3 respectivamente. La segunda, red neuronal utilizando Keras – Tensor Flow, el cual arrojó una precisión superior al 86%, aparte de la implementación de las métricas de evaluación. Precisión, Recall, F1 Score y accuracy, las cuales arrojaron resultados que superan los 80% en la clasificación de los riesgos.

Por último, se hace la comparación del resultado del proyecto de modelo de predicción para clasificación de riesgo cardiovascular, con la investigación Red neuronal para el diagnóstico de hipertensión arterial realizada por la escuela de ingeniería de la universidad de la Rioja (UNIR). Utilizando herramientas como: Node.js, Visual Studio Code Insider, se hizo la extracción de los datos. Herramientas como Weka, RapidMiner, se utilizaron para hacer la clasificación, imputación de datos faltantes y el balanceo de estos. Por ultimo las herramientas utilizadas para hacer el modelo son NetBeans, Launch4j-3.12 y GitHub. En los resultados que se obtuvieron, el mejor resultado que se arrojó fue de 73.8% de datos bien clasificados y 24.2% de datos mal clasificados, en la ilustración 26 se puede ver los resultados obtenidos por el modelo con más detalle.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      156530               73.8447 %
Incorrectly Classified Instances     55442               26.1553 %
Kappa statistic                          0.4767
Mean absolute error                      0.3604
Root mean squared error                  0.424
Relative absolute error                 72.0896 %
Root relative squared error             84.7937 %
Total Number of Instances           211972

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,812    0,336    0,708      0,812    0,757      0,482    0,796     0,797     1
                 0,664    0,188    0,779      0,664    0,717      0,482    0,796     0,784     2
Weighted Avg.    0,738    0,262    0,744      0,738    0,737      0,482    0,796     0,791

=== Confusion Matrix ===

     a     b   <-- classified as
 86278 19942 |     a = 1
 35500 70252 |     b = 2
```

*Ilustración 26 Red neuronal para el diagnóstico de hipertensión arterial realizado por la escuela de ingeniería de la universidad de la Rioja (UNIR). Resultados obtenidos*

Haciendo la comparativa de resultados obtenidos con la investigación anterior, se puede decir que, en términos de resultados, el modelo de predicción para clasificación de riesgo cardiovascular arroja mejores resultados, precisión superior al 86%, viendo los porcentajes en las métricas de evaluación. Precisión, Recall, F1 Score y accuracy, las cuales arrojaron resultados que superan los 80% en la clasificación de los riesgos.

# 5. CONCLUSIONES Y RECOMENDACIONES

Las enfermedades cardiovasculares, como bien se pudo observar, presentan un reto muy importante para las organizaciones de salud en cada rincón del mundo. Y sin duda alguna en América Latina y particularmente aquí en Colombia las condiciones socioeconómicas, el ámbito cultural, el tipo de clima y en fin los malos hábitos de vida en general apalancan en gran medida los riesgos de sufrir de este tipo de patología, sin dejar de lado el factor hereditario que significa uno de los más determinantes.

Con el propósito de contribuir a contar con un mecanismo que permita identificar oportunamente pacientes con riesgos bajos, medios y más concretamente aquellos con alto nivel de sufrir de Hipertensión Arterial, se desarrolló el siguiente proyecto el cual en la etapa final demostró con cifras confiables poder realizar una predicción en términos muy precisos y exactos, utilizando modelo de inteligencia artificial de aprendizaje profundo. A continuación, se especifica algunas conclusiones que dan cuenta de buenos resultados finales, siempre teniendo en cuenta la metodología de trabajo implementada:

- Fue determinante contar con una base preliminar de términos y conceptos utilizados en esta área de la salud que pudiera aclarar y conocer con mayor veracidad los datos obtenidos del hospital municipal de Arjona, Bolívar.

- Gracias al análisis inicial de los datos estadísticos y descriptivos proporcionados, fue crucial determinar de forma acertada el conjunto de atributos que contaban con una alta correlación con respecto a la variable final de respuesta en el conjunto de la base de datos. Garantizando así unos datos bien preparados y normalizados para llevar a cabo la siguiente fase el proyecto.

- En el siguiente paso se logró tener un entrenamiento suficiente para que el modelo de inteligencia artificial realizara el proceso de aprendizaje basado en datos previos de cada uno de los pacientes.

- Cuando se analiza los resultados finales luego de este modelo entrenado se le aplicara la predicción con nuevos datos, fue posible

confirmar que el modelo se comportó con una precisión y exactitud sobre el 87%.

Finalmente es posible concluir que aplicar modelos de inteligencia artificial con aprendizaje profundo en el campo de la salud para ayudar a establecer riesgos como en este caso de sufrir de Hipertensión, es excelente alternativa de aprovechamiento de los avances tecnológicos para un fin común, de gran impacto en la sociedad contemporánea. Es pertinente resaltar que es necesario impulsar la formación en el área de IA en los actuales y futuros profesionales, y a su vez también capacitar al grupo de especialistas de salud, para que en su día a día estos modelos de predicción constituyan aliados estratégicos en el desarrollo de sus actividades. Por último, una recomendación general para las organizaciones mundiales y locales de salud, fomentar y promocionar campañas de prevención, que mitiguen precisamente los riegos de contraer este tipo de patología que año a año, cobran millones de vidas alrededor del mundo.

En el desarrollo del proyecto se logró abarcar y desarrollar los puntos definidos en el alcance. Como recomendación general y para afinar la estructura, diseño, arquitectura y resultados finales del proyecto, se sugiere hacer primeramente un acercamiento con un profesional de la salud, el cual utilice durante un periodo de tiempo prudente. Y luego realizar un sondeo para ver el comportamiento de este en un ambiente real. Es pertinente evaluar confiablidad, disponibilidad, integridad de datos, rendimiento del modelo, cabe resaltar que antes es necesario construir una interfaz o dashboard que permita fácilmente interactuar con los datos de entrada y salida del modelo, idealmente que sea desarrollado en una plataforma tipo web, soportada por una infraestructura de servidor muy potentes, orientados al alto procesamiento de datos, lo sugerido es que sea en la nube, ya que garantizar un ambiente apto, capaz de procesar y calcular el modelo, requiere de inversiones alta en hardware.

## Bibliografía


[1] «OMS | Preguntas y respuestas sobre la hipertensión», *WHO*. http://www.who.int/features/qa/82/es/ (accedido oct. 27, 2020).

[2] «dia-mundial-hipertension-2017.pdf». Accedido: oct. 27, 2020. [En línea]. Disponible en: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ENT/dia-mundial-hipertension-2017.pdf.

[3] «¿Qué es la hipertensión arterial? | CuídatePlus», *CuidatePlus*, mar. 26, 2009. https://cuidateplus.marca.com/enfermedades/enfermedades-vasculares-y-del-corazon/hipertension-arterial.html (accedido oct. 27, 2020).

[4] «Causas y factores de riesgo de la Hipertensión Arterial | Hospital Clínic Barcelona», *Clínic Barcelona*. https://www.clinicbarcelona.org/asistencia/enfermedades/hipertension-arterial/causas-y-factores-de-riesgo (accedido oct. 28, 2020).

[5] A. Álvarez Aliaga, J. C. González Aguilera, L. del R. Maceo Gómez, A. Frómeta Guerra, S. Bárzaga Morell, y A. E. Cervantes Infante, «Árbol para predecir el desarrollo de la cardiopatía hipertensiva», *Rev. Cuba. Med.*, vol. 53, n.º 3, pp. 266-281, sep. 2014.

[6] «maria del carmen avila lillo.pdf». Accedido: oct. 28, 2020. [En línea]. Disponible en: http://147.96.70.122/web/tfg/tfg/memoria/maria%20del%20carmen%20avila%20lillo.pdf.

[7] J. E. P. Rodriguez, L. Boada-Morales, D. P. Florez, y M. del P. Castellanos-Duarte, «Predicción del riesgo cardiovascular e hipertensión arterial según Framingham en pacientes de atención primaria en salud. Estudio FRICC», *Rev. Colomb. Med. Física Rehabil.*, vol. 26, n.º 2, Art. n.º 2, dic. 2016, doi: 10.28957/rcmfr.v26n2a3.

[8] «cim172c.pdf». Accedido: ene. 17, 2021. [En línea]. Disponible en: https://www.medigraphic.com/pdfs/revcubinfmed/cim-2017/cim172c.pdf.

[9] «GARCIA MONTERO, YOLANDA.pdf». Accedido: ene. 17, 2021. [En línea]. Disponible en: https://reunir.unir.net/bitstream/handle/123456789/6937/GARCIA%20MONTERO%2c%20YOLANDA.pdf?sequence=1&isAllowed=y.

[10] A. Leha *et al.*, «A machine learning approach for the prediction of pulmonary hypertension», *PLoS ONE*, vol. 14, n.º 10, oct. 2019, doi: 10.1371/journal.pone.0224453.

[11] A. Álvarez Aliaga, J. C. González Aguilera, L. del R. Maceo Gómez, A. Frómeta Guerra, S. Bárzaga Morell, y A. E. Cervantes Infante, «Árbol para predecir el desarrollo de la cardiopatía hipertensiva», *Rev. Cuba. Med.*, vol. 53, n.º 3, pp. 266-281, sep. 2014.

[12] J. E. P. Rodriguez, L. Boada-Morales, D. P. Florez, y M. del P. Castellanos-Duarte, «Predicción del riesgo cardiovascular e hipertensión arterial según Framingham en pacientes de atención primaria en salud. Estudio FRICC», *Rev. Colomb. Med. Física Rehabil.*, vol. 26, n.º 2, Art. n.º 2, dic. 2016, doi: 10.28957/rcmfr.v26n2a3.

[13] Y. Morillo, «Deep learning o aprendizaje profundo | Qué es y cómo funciona», *Futuro Electrico*, ago. 27, 2020. https://futuroelectrico.com/deep-learning-aprendizaje-profundo/ (accedido oct. 28, 2020).

[14] «El origen de Deep Learning», *Máster en Deep Learning : Universidad de Alcalá - Madrid*, mar. 04, 2019. https://master-deeplearning.com/origen-deep-learning/ (accedido oct. 28, 2020).

[15] V. Roman, «Aprendizaje Supervisado: Introducción a la Clasificación y Principales Algoritmos», *Medium*, abr. 01, 2019. https://medium.com/datos-y-ciencia/aprendizaje-supervisado-introducci%C3%B3n-a-la-clasificaci%C3%B3n-y-principales-algoritmos-dadee99c9407 (accedido oct. 28, 2020).

[16] «Regresión Lineal | Aprende Machine Learning». https://www.aprendemachinelearning.com/tag/regresion-lineal/ (accedido oct. 28, 2020).

[17] N. Dandekar, «Intuitive explanation of Learning to Rank (and RankNet, LambdaRank and LambdaMART)», *Medium*, ene. 14, 2016. https://medium.com/@nikhilbd/intuitive-explanation-of-learning-to-rank-and-ranknet-lambdarank-and-lambdamart-fe1e17fac418 (accedido oct. 28, 2020).

[18] N. X. Vinh, J. Epps, y J. Bailey, «Information theoretic measures for clusterings comparison: is a correction for chance necessary?», en *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, Montreal, Quebec, Canada, 2009, pp. 1-8, doi: 10.1145/1553374.1553511.
[19] «Hipertensión arterial - adultos: MedlinePlus enciclopedia médica». https://medlineplus.gov/spanish/ency/article/000468.htm (accedido oct. 28, 2020).
[20] «Definición de variable — Definicion.de», *Definición.de*. https://definicion.de/variable/ (accedido oct. 28, 2020).
[21] J. Brownlee, «SMOTE for Imbalanced Classification with Python», *Machine Learning Mastery*, ene. 16, 2020. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (accedido dic. 01, 2020).
[22] «MARIA DEL CARMEN AVILA LILLO.pdf». Accedido: ene. 17, 2021. [En línea]. Disponible en: http://147.96.70.122/Web/TFG/TFG/Memoria/MARIA%20DEL%20CARMEN%20AVILA%20LILLO.pdf.