
**MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL
RIESGO DE FATALIDAD POR INSUFICIENCIA CARDIACA CON DATOS
CLÍNICOS**

Autores

David Alejandro Gallego Valcárcel

Delly Fabián Lucas Monsalve

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá, Colombia
2021

**MODELOS DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DEL
RIESGO DE FATALIDAD POR INSUFICIENCIA CARDIACA CON DATOS
CLÍNICOS**

Autores

David Alejandro Gallego Valcárcel

Delly Fabián Lucas Monsalve

Trabajo de grado como requisito para optar al título de:

Ingeniero de Sistemas y computación

Director:

Juan Camilo Ramírez

Asesora Metodológica:

Rosalba Cruz Cepeda

Licenciada en Educación

Tipo de Proyecto:

Investigación

Universidad Antonio Nariño
Facultad de Ingeniería de Sistemas
Bogotá, Colombia

2021

INDICE

| | |
|---|--------------|
| INDICE DE ILUSTRACIONES Y TABLAS | |
| RESUMEN..... | |
| INTRODUCCIÓN | 1 |
| 1. PLANTEAMIENTO DEL PROBLEMA | 2 |
| 1.1. DESCRIPCIÓN DEL PROBLEMA | 2 |
| 1.2. FORMULACIÓN DEL PROBLEMA | 2 |
| 1.3. JUSTIFICACIÓN | 2 |
| 1.4. OBJETIVOS..... | 3 |
| 1.4.1. <i>Objetivo general</i> | 3 |
| 1.4.2. <i>Objetivos específicos</i> | 3 |
| 1.5. ALCANCE Y LIMITACIONES | 4 |
| 1.5.1. <i>Alcance</i> | 4 |
| 1.5.2. <i>Limitaciones</i> | 4 |
| 2. ASPECTOS METODOLÓGICOS | 5 |
| 2.1. DEFINICIÓN DEL MODELO DE CLASIFICACIÓN..... | 6 |
| 2.2. ENTRENAMIENTO DEL MODELO | 7 |
| 2.3. EVALUACIÓN DE LOS RESULTADOS | 8 |
| 3. MARCO DE REFERENCIA | 9 |
| 3.1. MARCO TEÓRICO | 9 |
| 3.1.1. <i>Insuficiencia Cardíaca</i> | 9 |
| 3.1.2. <i>Aprendizaje Automático</i> | 10 |
| 3.1.3. <i>Tipos de aprendizaje automático</i> | 11 |
| 3.1.4. <i>Aprendizaje supervisado</i> | 12 |
| 3.1.5. <i>Aprendizaje no supervisado</i> | 12 |
| 3.1.6. <i>Clasificación y Regresión</i> | 13 |
| 3.1.7. <i>Clasificación</i> | 14 |
| 3.1.8. <i>Árboles de decisión</i> | 14 |
| 3.1.9. <i>Bosques aleatorios</i> | 15 |
| 3.1.10. <i>Máquinas de vectores soporte (SVM)</i> | 15 |
| 3.1.11. <i>Redes neuronales</i> | 15 |
| 3.1.12. <i>Python</i> | 16 |
| 3.1.12.1. Scikit-learn | 16 |
| 3.1.13. <i>Análisis componentes principales</i> | 16 |
| 3.1.14. <i>La validación cruzada y los subconjuntos de pruebas</i> | 17 |
| 3.1.15. <i>La varianza explicada</i> | 17 |
| 3.1.16. <i>Optimización de hiperparámetros</i> | 17 |
| 3.1.17. <i>Los clasificadores binarios y los tipos de error</i> | 17 |
| 3.1.18. <i>La matriz de confusión</i> | 18 |
| 3.1.18.1. La métrica de la curva ROC (Receiver Operating Characteristic) | 19 |
| 3.1.18.2. La medida por medio de F1 | 19 |
| 3.1.18.3. Precisión y Exhaustividad (recall) | 19 |
| 3.1.18.4. Precisión y exactitud | 20 |
| 3.1.18.5. Precisión balanceada (sensibilidad y especificidad) | 21 |
| 3.1. ANTECEDENTES..... | 21 |
| 3.3. MARCO LEGAL | 24 |
| 4. DESARROLLO DEL PROYECTO..... | 25 |

| | | |
|-----------|---|-----------|
| 4.1. | DESCRIPCIÓN BREVE DE LOS DATOS A USAR..... | 25 |
| 4.2. | PREPROCESAMIENTO DE DATOS..... | 26 |
| 4.2.1. | <i>Técnicas estadísticas usadas</i> | 26 |
| 4.2.2. | <i>Procedimiento</i> | 26 |
| 5. | RESULTADOS OBTENIDOS..... | 28 |
| 5.1. | MODELO USADO - SUPPORT VECTOR MACHINE (SVM) | 28 |
| 5.1.1. | <i>Resultados de la búsqueda en tabla para SVM</i> | 28 |
| 5.2 | MODELO USADO - RANDOM FOREST CLASSIFIER..... | 28 |
| 5.1.2. | <i>Resultados de la búsqueda en tabla para RF</i> | 29 |
| 5.2. | MODELO USADO - MULTILAYER PERCEPTRON CLASSIFIER (RED NEURONAL) | 30 |
| 5.2.1. | <i>Resultados de la búsqueda entabla para MLP</i> | 30 |
| 5.3. | ITERACIONES CON LOS RESULTADOS OBTENIDOS POR LA BÚSQUEDA EN TABLA | 31 |
| 5.4. | TRANSICIÓN DE LOS RESULTADOS | 32 |
| 5.5. | COMPARATIVA DE RESULTADOS..... | 33 |
| 6. | CONCLUSIONES Y RETROSPECTIVA | 36 |
| | REFERENCIAS BIBLIOGRAFICAS..... | 37 |

INDICE DE ILUSTRACIONES Y TABLAS

| | |
|--|----|
| Ilustración 1 Diagrama BPMN de la metodología aplicada | 5 |
| Ilustración 2. Diagrama Fase 1 BPMN Definición del modelo de clasificación | 6 |
| Ilustración 3. Diagrama Fase 2 BPMN de la metodología aplicada | 7 |
| Ilustración 4. Diagrama Fase 3 BPMN de la metodología aplicada | 8 |
| Ilustración 5. Enfoque y separación de la inteligencia artificial | 11 |
| Ilustración 6. Esquema de funcionamiento de Aprendizaje supervisado | 12 |
| Ilustración 7. Diagrama de flujo del aprendizaje no supervisado | 13 |
| Ilustración 8 Buena exactitud con mala precisión y buena veracidad | 20 |
| Ilustración 9 Buena precisión con mala exactitud y veracidad..... | 20 |
| Ilustración 10. Gráfico de varianza de los datos, con 98% de precisión en $x \sim 10$ | 27 |
| Ilustración 11 Promedio de resultados de las ejecuciones desde su primera iteración hasta la décima y última iteración..... | 32 |
| Ilustración 12. Gráfico de desempeño de modelos utilizados con sus máximos valores | 32 |
| Ilustración 13. Gráfico de desempeño de modelos vs. tope de rendimiento por parte de la investigación [1]..... | 34 |
| Ilustración 14. Variación de los modelos y evaluación del estudio en cuestión..... | 35 |
| | |
| Tabla 1. Clasificación funcional de pacientes con insuficiencia cardiaca | 10 |
| Tabla 2. Resultados de métricas de evaluación para Máquinas de vectores de soporte | 28 |
| Tabla 3. Resultados de métricas de evaluación para Bosques Aleatorios | 29 |
| Tabla 4. Resultados de métricas de evaluación para la red neuronal..... | 31 |
| Tabla 5. Resultados de métricas de evaluación MLP, SVM y RF..... | 31 |
| Tabla 6. Comparativa de resultados por modelo según su promedio | 33 |

RESUMEN

La insuficiencia cardiaca es una enfermedad en la que existe un desequilibrio entre la capacidad del sistema circulatorio y la circulación de sangre en el organismo. Este trabajo de grado busca elaborar una predicción del riesgo de fallecimiento en pacientes con insuficiencia cardiaca, evaluando datos clínicos, por ejemplo: la capacidad de bombeo del corazón, sodio en la sangre, plaquetas, entre otros. Predecir cuándo un paciente morirá por insuficiencia cardiaca es complejo [1], y se puede complicar aún más cuando sólo se cuenta con pocos detalles de lo que padeció durante la enfermedad y hasta su deceso [2].

Este trabajo está enfocado en el campo de ciencias de datos, y consiste en la implementación, configuración y evaluación de modelos utilizando técnicas de aprendizaje automático, para predecir el riesgo de fallecer por insuficiencia cardiaca a partir de datos clínicos de pacientes recopilados [1]. La ejecución de este trabajo se dará en tres fases: en la primera fase se utilizó el análisis estadístico para la medición de errores, entendimiento de los datos, y preprocesado; la segunda fase del proyecto implicó la implementación y configuración de los modelos, finalmente en la tercera se realizó la extracción de resultados y comparación de estos con respecto a la investigación[1].

PALABRAS CLAVES: aprendizaje automático, insuficiencia cardiaca, modelos de aprendizaje, inteligencia artificial.

INTRODUCCIÓN

La Organización Mundial de la Salud (OMS) refleja que la causa de muerte más común en el mundo son las enfermedades cardiacas, para 2015 el fallecimiento se registraba en aproximadamente un 31% de decesos, siendo casi 17.7 millones de personas alrededor del mundo; para Colombia, en 2017 su estimación abarcó entre más de 647.000 afectados [3]. El procedimiento de diagnóstico sobre la insuficiencia cardíaca generalmente consta de la información sintomatológica, cómo lo es si el paciente padece de anemia o diabetes, también el análisis de muestra de sangre proveída por el paciente en consulta, entre otros síntomas. En la búsqueda de la causa del problema, los procedimientos pueden pasar a ser invasivos [4]; generalmente sólo se mantiene un historial médico limitado y de difícil acceso al público. La información que se tiene para este proyecto proviene de una investigación realizada con una muestra de 299 personas proveniente del artículo en el que se apoya este trabajo de grado [1], donde todos poseen insuficiencia cardíaca pertenecientes a las clases III y IV según la NYHA (Asociación del Corazón de Nueva York). A estas personas se les hizo un conjunto de pruebas relacionadas a funciones cardiacas [1]; la información en el conjunto muestra el tiempo desde el diagnóstico de la insuficiencia cardíaca hasta que el paciente mostró mejoría o falleció.

Un precedente que se debe tener en cuenta es el aprendizaje de máquina y su aplicación en el campo de la salud, donde éste se ha ido incrementando aproximadamente en un 30% a 40% año tras año con ayuda del análisis estadístico y probabilístico de otras investigaciones [1] [3]. Aun así, el reto de recopilación y manipulación de datos es complejo, tal cual lo es la predicción de muerte tras padecer alguna enfermedad, y el reto es aún mayor cuando se trata del ser humano y órganos tan complejos como el corazón o el sistema que lo compone [5]. Pero el ser diagnosticado no es un final decisivo para las personas cuando existen varios métodos para mantener un corazón saludable [6]. Por este motivo es necesario mantener una expectativa de vida cuando hay un diagnóstico temprano, y ver cómo los factores clínicos influyen en la muerte de los pacientes [7].

Lo que se busca es utilizar datos clínicos de pacientes que manifestaron insuficiencia cardíaca al poder analizar modelos de clasificación, cuyas selecciones son Maquinas de vectores de soporte, Bosques aleatorios y Redes neuronales; en relación al aprendizaje de máquina e inteligencia artificial (IA), para luego comparar los resultados entre la investigación presentada por los autores del artículo [1] y este trabajo de grado.

1. PLANTEAMIENTO DEL PROBLEMA

1.1. DESCRIPCIÓN DEL PROBLEMA

Se estima que el riesgo de que una persona desarrolle insuficiencia cardíaca a lo largo de su vida está entre el 20% y 46%, tanto para hombres como mujeres [8], y generalmente empeora con el pasar de los días particularmente al no detectarlo a tiempo [3]. En el artículo de referencia [1] se alcanza en su tope aproximadamente 82% de rendimiento en sus análisis estadísticos; siendo esta la motivación para incrementar el valor de este porcentaje con alguna técnica diferente, cómo lo es la inteligencia artificial (de ahora en adelante IA).

La creación de un modelo de IA puede resultar muy útil en el diagnóstico. Aun así, los modelos de aprendizaje automático conocidos hasta ahora que se usan para predecir el diagnóstico de insuficiencia cardíaca logran resultados con una precisión variable, entre el 80% y 90% en sus mejores casos [9]. El conjunto de datos a utilizar ha sido empleado para entrenar modelos con algunas técnicas de aprendizaje de máquina, pero hay técnicas que aún no se han utilizado y que podrían arrojar resultados interesantes, algunos de estos mencionados en los antecedentes.

1.2. FORMULACIÓN DEL PROBLEMA

El reto de recopilar información es algo difícil, es requerido darle sentido a eso que se desea analizar y en algunos casos es necesario contar con grandes volúmenes de datos; las enfermedades son un ejemplo de aquello, de igual manera aplica para los datos con los que se cuentan en este trabajo de grado, donde algunos de los datos de los pacientes se presentan de forma cualitativa, y otros no son ordenables, o incluso información ausente como el peso, estatura, colesterol, entre otros varios [10]. Además, se visualiza un porcentaje de rendimiento estadístico que se consideraría bajo en escalas grandes de cantidades de personas, dígame en el millón o más, siendo un 82% algo muy arriesgado si se trata de controlar el crecimiento de una enfermedad. Aun así, posibilita el usar modelos de aprendizaje con IA optimizando los porcentajes de rendimiento para mejorar la predicción sobre estas enfermedades, por lo que surge la pregunta de ¿Cuáles modelos de aprendizaje automático se pueden usar para predecir el riesgo de fallecer por insuficiencia cardíaca con mayor confiabilidad que los modelos contemplados en el artículo [1]?

1.3. JUSTIFICACIÓN

La recolección de información para el estudio [1] alcanzó un porcentaje que se consideraría bajo en un conjunto poblacional grande y omiten varios datos dentro del conjunto que los autores mismos anuncian que usaron, cómo se esclarecen en otros artículos [10-12]. La motivación para efectuar este trabajo de grado es buscar, dentro del aprendizaje automático y con el apoyo de IA, modelos que ajusten dicho porcentaje y poderlo mejorar en cuanto a su rendimiento. Se debe reiterar que la información presentada no es perfecta [11]. En la investigación de referencia [1] se pueden dimensionar algunos datos que faltan; como se afirma en otras investigaciones [12]. La

información faltante como el peso, estatura, ritmo cardiaco, entre otros, no descarta la posibilidad que se pueda limpiar la información, como eliminar datos innecesarios o reemplazarlos para poder usarlos en algún modelo de predicción, así como fueron utilizados previamente usando técnicas de probabilidad [1], [11], [12].

A pesar de que en el artículo [1] se tienen métricas estadísticas y de comprobación, esto se puede optimizar con el uso de aprendizaje automático, ya que tiene útiles herramientas de deducción como los clasificadores y sus respectivas métricas de rendimiento (medida F, área bajo la curva, entre algunas varias), donde se comprueba que el resultado de un modelo opere bajo un margen de error definido; así, se contempla la oportunidad de predecir el riesgo de fallecimiento por una enfermedad crónica, como lo es con la insuficiencia cardiaca.

Con el uso de inteligencia artificial y en específico con el apoyo de aprendizaje automático, es posible predecir el riesgo de muerte por insuficiencia cardiaca con resultados incluso mejores que los otorgados por los modelos estadísticos usados en el caso de estudio [1].

La importancia en este trabajo de grado recae en un análisis y evaluación de modelos basados en *Machine Learning* en el que se evidencia la optimización para la predicción del riesgo de fallecimiento por insuficiencia cardíaca contemplando variables y valores que no se tuvieron en cuenta en la investigación [1].

1.4. OBJETIVOS

1.4.1. Objetivo general

Implementar tres modelos de aprendizaje automático para predecir el riesgo de fallecimiento por insuficiencia cardíaca, a partir de los datos clínicos de pacientes recopilados [1], utilizando técnicas de aprendizaje y parametrizaciones no contempladas aún para identificar el mejor a partir de las métricas comúnmente utilizadas en la literatura científica.

1.4.2. Objetivos específicos

- Elegir tres técnicas de predicción dentro las evaluaciones comúnmente usadas en la literatura investigativa con respecto a los temas de IA, como máquinas de vectores de soporte y bosques aleatorios entre otros.
- Entrenar los modelos contemplando varias técnicas de aprendizaje y de preprocesamiento de datos y buscando, para cada una, la combinación de parámetros que produce el mejor desempeño predictivo.
- Evaluar los modelos implementados utilizando diferentes métricas de desempeño incluyendo la medida F (F1-Score) y el área bajo la curva ROC (AUC), para determinar el mejor.

1.5. ALCANCE Y LIMITACIONES

1.5.1. Alcance

Se realizó un estudio y evaluación de varios modelos de clasificación apoyados en inteligencia artificial, tales como: máquinas de vectores de soporte (SVM), redes neuronales y bosques aleatorios [13]. Seguido de esto se entrenaron los modelos, con el conjunto de datos de estudio aplicando diferentes modelos clasificadores a la información; finalmente, se realizaron las implementaciones y evaluaciones de los modelos del caso de estudio en el lenguaje de programación Python.

Se implementó el aprendizaje de máquina usando varias bibliotecas de Python como Pandas, scikit-learn, entre otras [14], [15]; en la recopilación de información se usaron historiales clínicos ya consolidados por investigadores [1]. La visualización de resultados se presenta después de la ejecución de la aplicación en consola con el uso de los modelos clasificadores.

1.5.2. Limitaciones

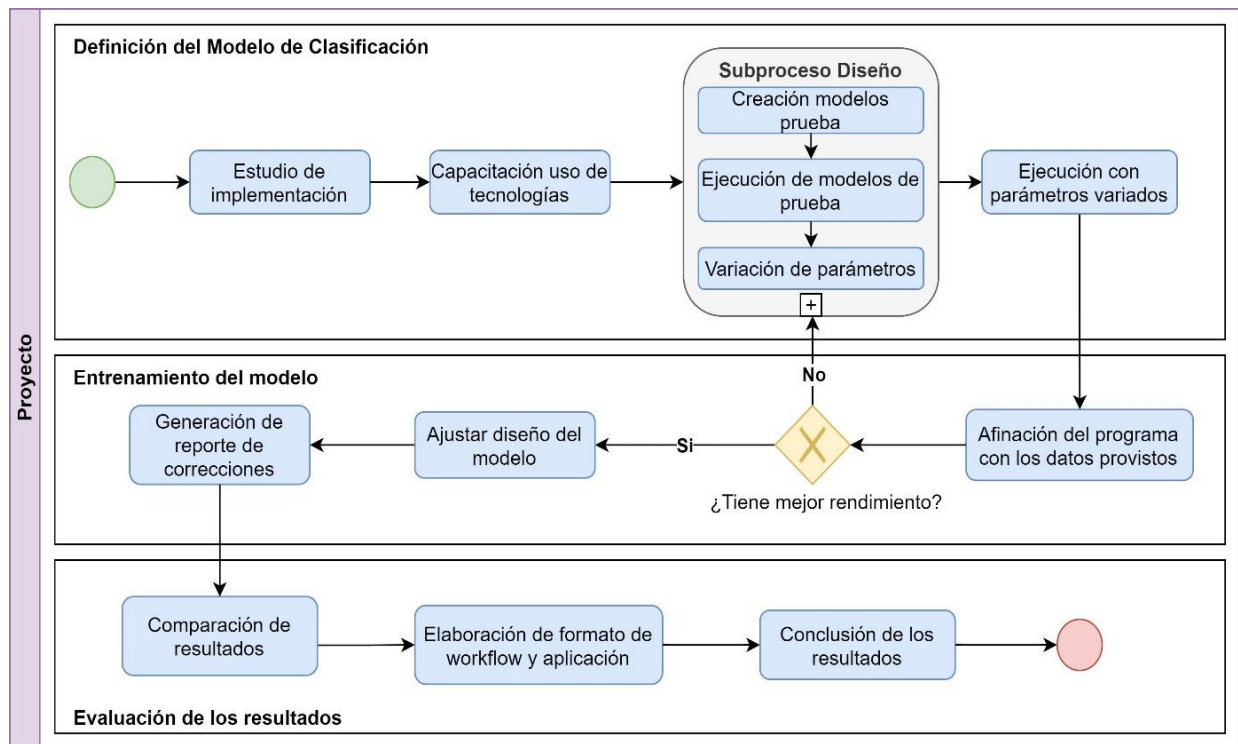
La logística de recolección de datos para este tipo de investigación está por fuera de los recursos de este trabajo de grado, por lo que se usarán datos previamente recopilados por otros investigadores [1].

En cuanto al grupo de información, la manipulación y limpieza de los datos puede llegar a reducir más el volumen de datos con el que ya se cuenta, sin contar el descarte de atributos útiles en el modelo, a pesar de contar con la asesoría del Ingeniero Juan Camilo Ramírez investigador en LACSER (*Laboratory for Advanced Computational Science and Engineering Research*) en el tema de ciencias computacionales.

2. ASPECTOS METODOLÓGICOS

Al ser un tema investigativo, se enfocó en tres actividades generales, cada una con un conjunto de subtarefas que en concreto ayudan a cumplir cada uno de los objetivos planteados. Basándose en el conjunto de datos proporcionado en la investigación [1] se divide el proceso de desarrollo del proyecto en tres fases comprendidas para el entendimiento, ejecución y comparación de lo elaborado en este trabajo de grado. En la Ilustración 1 se puede apreciar el diagrama usando BPMN (Business Process Model Notation) que indica el proceso usado y dentro de las expectativas se considera suficiente para representar de forma general los pasos efectuados en la elaboración de este trabajo de grado.

Ilustración 1 Diagrama BPMN de la metodología aplicada

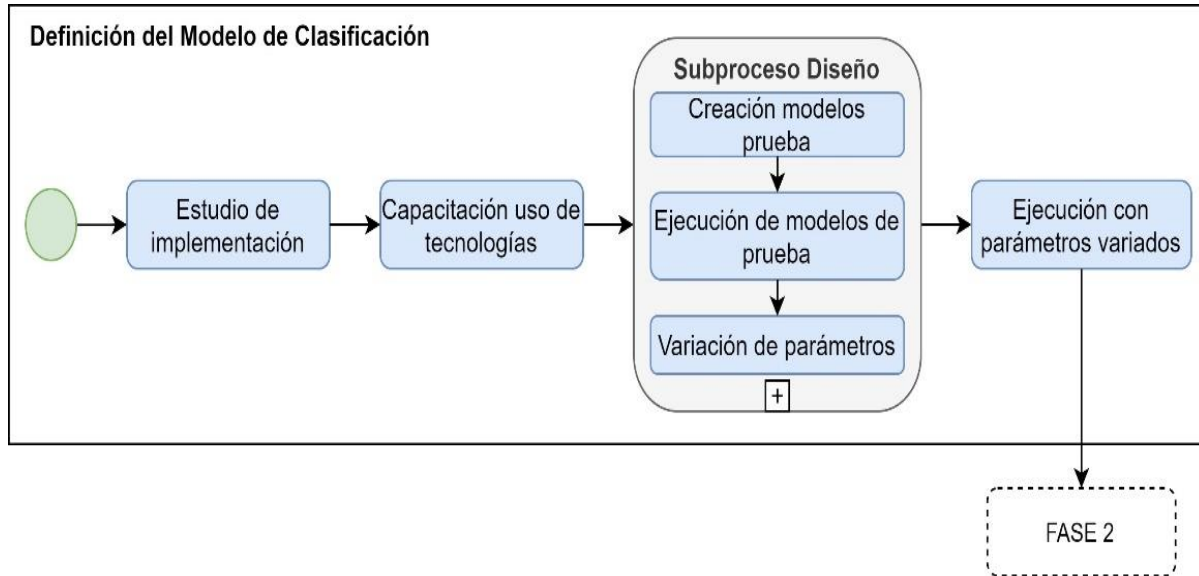


Fuente: Elaboración propia

Para el proyecto se realizó el proceso en tres fases, con sus respectivas actividades y entregas esperadas; se muestra cada una de ellas a continuación:

2.1. Definición del Modelo de Clasificación

Ilustración 2. Diagrama Fase 1 BPMN definición del modelo de clasificación



Fuente: Elaboración propia

Por medio de lectura, y algunas prácticas con las herramientas esperadas para la elaboración del trabajo de grado, se realiza una inmersión introductoria en el lenguaje de programación Python, así mismo el estudio y repaso de los modelos de aprendizaje automático aplicables con base en lo que menciona el artículo, junto a conceptos relacionados a la insuficiencia cardíaca [1] para explorar posibles mejoras en los resultados.

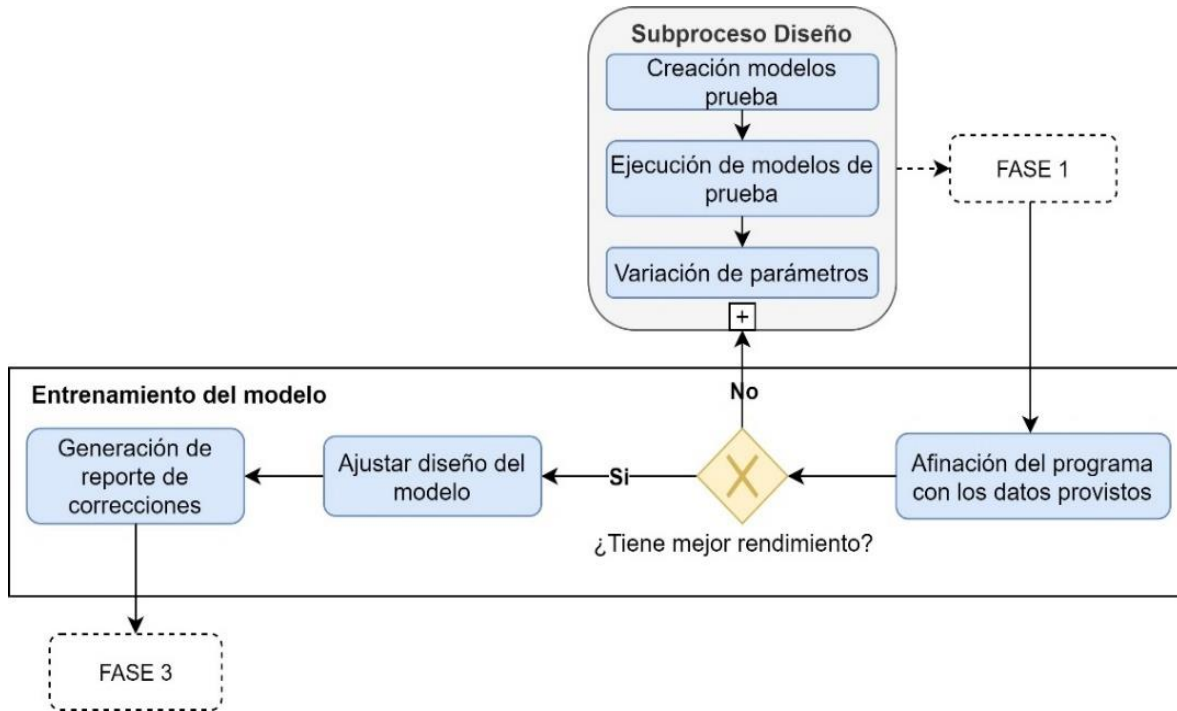
Se contempló el uso de varias técnicas de evaluación de rendimiento de modelos de aprendizaje implementados por nosotros, como las mínimas esperadas que son las que menciona el caso de estudio [1] que son el área bajo la curva (ROC AUC), la medida F (F1-Score) en el caso de este trabajo de grado y adicionales como lo es evaluar por medio de precisión y exhaustividad, todos estos conceptos explicados más adelante en este documento.

Para la variación de parámetros fue necesario ver qué resultados del estudio [1] en concreto fueron obtenidos en sus evaluaciones de rendimiento. Es acá donde se usó el método de búsqueda de tabla para realizar las ejecuciones con los parámetros que facilita la biblioteca de scikit-learn.

Con los resultados obtenidos en la búsqueda de tabla posterior a la variación de parámetros se realiza una ejecución adicional para comprobar fiabilidad en los consiguientes resultados y tiempos de ejecución del modelo.

2.2. Entrenamiento del Modelo

Ilustración 3. Diagrama Fase 2 BPMN de la metodología aplicada



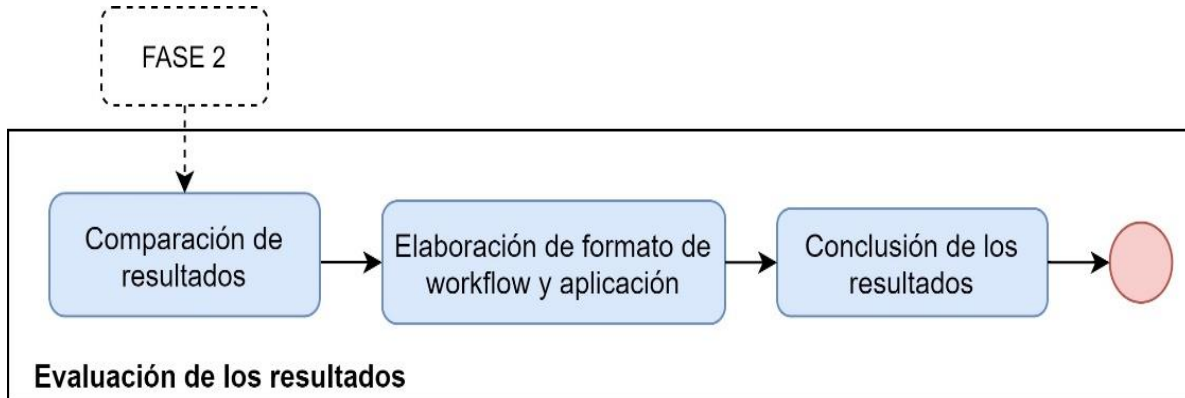
Fuente: Elaboración propia

Ya que la ejecución de los modelos es un proceso secuencial e iterativo es necesario agrupar los resultados conforme se van generando en sus ejecuciones. En la afinación del programa con los datos provistos se efectuó el preprocesado, volviendo a repetir el ciclo hasta una iteración en la que no haya un cambio drástico entre los porcentajes de las métricas. Esta fase incluye la creación del programa con uso de IA particularmente del uso de aprendizaje automático y contemplando los modelos en los siguientes pasos:

- Ejecución de la prueba de entrada con enfoque de validación cruzada, para el modelo obtenido con los datos preprocesados.
- Grabar los registros de resultados en el archivo plano para comparar los efectos de los cambios de parámetros posteriormente.
- Introducción a la fase 3, donde se realiza la evaluación y comparación de resultados.

2.3. Evaluación de los resultados

Ilustración 4. Diagrama Fase 3 BPMN de la metodología aplicada



Fuente: Elaboración propia

En esta última fase del desarrollo del proyecto, se efectúa formalmente la comparación de los resultados obtenidos [1] con cada uno de los tres modelos esperados y definidos en la primera fase siendo estos: Maquinas de vectores de soporte, Bosques aleatorios, y Perceptrón de multicapa (redes neuronales). Esta fase abarca los siguientes ítems:

- Generar reporte de resultados del modelo implementado y corregido.
- Realizar la comparación entre los resultados obtenidos y los descritos en el artículo.
- Crear el formato de flujo de trabajo que se desempeñó durante las 3 fases del proyecto y finalización del programa.
- Entregar resultados de actividades y documento del trabajo de investigación junto a la conclusión encontrada.

3. MARCO DE REFERENCIA

3.1. MARCO TEÓRICO

3.1.1. Insuficiencia Cardíaca

La insuficiencia cardíaca es una enfermedad donde el sistema circulatorio no es capaz de transportar la sangre correctamente a todo el organismo, desarrollándose por anomalías en algunas condiciones en los seres humanos, tales como nivel de azúcar, fuerza de bombeo del corazón, entre otras varias [5], [16], [17], presentándose principalmente en el corazón.

Para conocer los aspectos en que se basaron para realizar el estudio [1] es necesario mencionar de qué tratan algunos padecimientos adicionales que fueron tomados en cuenta en el conjunto de datos del cual relata el artículo [1]. De igual forma, es crucial ver cuáles son sus estados normales o en qué rango de valores se puede considerar saludable; según las investigaciones [2], [11], [12] las cuales hacen profundidad en lo que se muestra en el estudio original [1], hay más afecciones y medidas que influyen en la insuficiencia cardíaca (desde ahora atributos) que son:

- La diabetes: es un padecimiento que influye en la insuficiencia cardíaca, ya de por sí establece un problema del manejo de azúcar. Un nivel de glucosa sanguínea en ayunas por debajo de 100 miligramos por decilitro (mg/dl) (5,6 milimoles por litro [mmol/l]) se considera normal. Un nivel de glucosa sanguínea en ayunas entre 100 y 125 mg/dl (5,6 a 7 mmol/l) se considera prediabetes [18].
- El índice de fracción de eyección: es la tasa de compresión del corazón con base en su ventrículo izquierdo para el bombeo de sangre del corazón hacia el resto del cuerpo, siendo su estado saludable cuando se puede comprimir a más del 50% de su tamaño [11].
- El número de plaquetas: el número promedio que posee el cuerpo humano, siendo lo normal de estas en la sangre es de 150,000 a 400,000 por microlitro (mcL) o 150 a $400 \times 10^9/L$. Más allá de eso es un problema hormonal, y menos de eso es un problema de deficiencia en el hígado [17].
- Nivel de sodio en la sangre: el sodio es un electrolito y ayuda a la regulación de la cantidad de agua que hay dentro y alrededor de las células, donde lo normal de sodio en la sangre oscila entre 135 y 145 miliequivalentes por litro (mEq/L) [19].
- La creatinfosfoquinasa: es un tipo de enzima en el cuerpo que principalmente predomina en el corazón, el cerebro y el músculo esquelético (tendones). Esta enzima es la que le da “poder” o fuerza a los músculos y tendones [17]. Sus valores normales oscilan entre 32 y 294 U/L en hombres y 33 a 211 U/L en mujeres.
- La creatinina: es un producto de desechos del metabolismo humano. Su secreción es dada por parte de los riñones, este desecho es filtrado desde la sangre luego de que hay reacciones químicas por esfuerzo muscular [13]. Los valores normales en la sangre pueden

ser de 0,84 a 1,21 miligramos por decilitro (74,3 a 107 micromoles por litro), donde estas métricas varían según el tipo de medición, entre hombres y mujeres, y según la edad [20].

Existe una clasificación relevante en este estudio determinada por la New York Heart Association (NYHA), que es la asociación del corazón de Nueva York; es utilizada como método para la clasificación funcional de pacientes con insuficiencia cardíaca. Fue propuesta en 1928 y revisada en varias ocasiones, la última en 1994. Designa 4 clases (I, II, III y IV), basándose en las limitaciones en la actividad física del paciente ocasionadas por los síntomas cardíacos [21], evidenciados en la Tabla 1.

Tabla 1. Clasificación funcional de pacientes con insuficiencia cardiaca

| CLASE | CLASIFICACIÓN FUNCIONAL NYHA |
|-------|---|
| I | No se limita la actividad física. La actividad ordinaria no ocasiona excesiva fatiga, palpitaciones o disnea (Ahogo o dificultad en la respiración) |
| II | Ligera limitación de la actividad física. La actividad ordinaria ocasiona fatiga, palpitaciones o disnea. |
| III | Marcada limitación de la actividad física. Actividad física menor que la ordinaria ocasiona fatiga, palpitaciones o disnea. |
| IV | Incapacidad para llevar a cabo cualquier actividad física sin confort. Los síntomas de insuficiencia cardíaca o de síndrome ansioso pueden estar presentes incluso en reposo. Si se realiza cualquier actividad física, la inconformidad aumenta. |

Fuente: New York Heart Association (NYHA)

Las enfermedades cardiovasculares poseen un alto índice de ser provenientes por antecedentes familiares, siendo 2 de cada 27 personas las afectadas por alguna de estas en su núcleo familiar [5]. Lo anterior es particularmente visible en adultos mayores de 35 años [1].

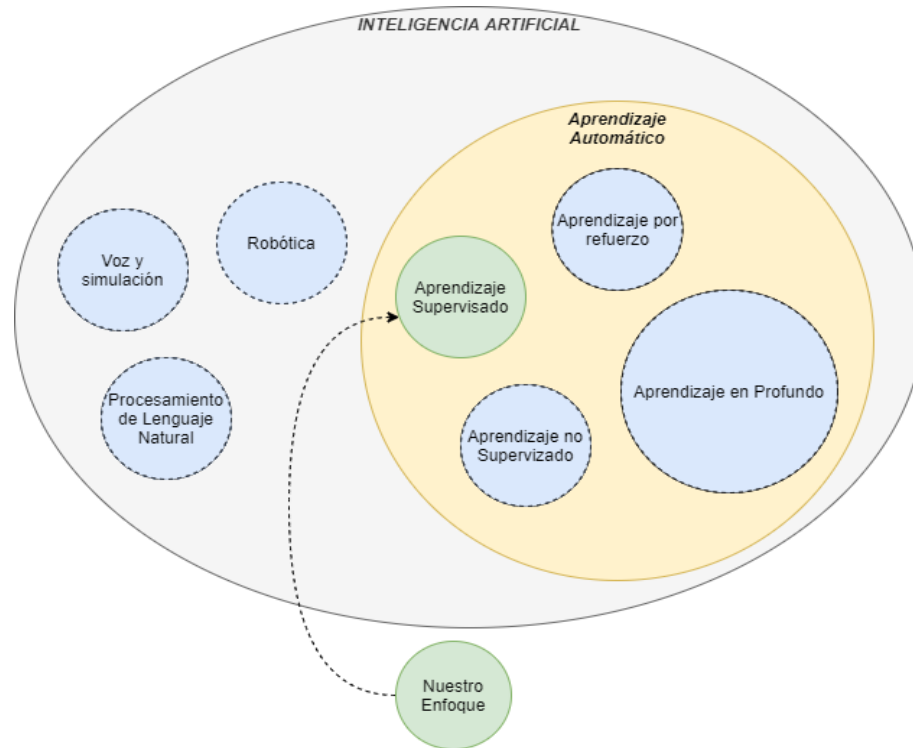
3.1.2. Aprendizaje Automático

El aprendizaje automático o del inglés, Machine Learning, es un conjunto de técnicas de estadística y probabilidad aplicada con el uso de tecnologías basadas en ciencias de datos para desarrollar software predictivo sobre **agentes inteligentes** siendo estos principalmente actores que operan bajo una acción dentro de un entorno bajo unas condiciones particulares y con respuestas finitas [22].

En sus usos principales y prácticos, recae la operación en Big Data, el reconocimiento sobre patrones y descifrado. El Machine Learning, desde ahora ML, es una subrama de la inteligencia artificial, dentro de las ciencias de computación, donde se subdivide en tres tipos de aprendizaje

automático. En la Ilustración 5, se puede observar el supra conjunto que abarca el campo de inteligencia artificial (IA).

Ilustración 5. Enfoque y separación de la inteligencia artificial



Elaboración: Fuente propia

3.1.3. Tipos de aprendizaje automático

La característica principal del aprendizaje automático es emplear conjuntos de algoritmos que aprenden y resuelven problemas a partir de la experiencia. Hay diversos tipos de problemas que se abordan con técnicas de aprendizaje automático. Entre ellos se encuentran los problemas de clasificación del que se basa este proyecto de grado, donde se requiere predecir un atributo del conjunto de datos denominado **clase** donde este es el factor clave a predecir. En este contexto es el **fallecimiento** de un paciente por insuficiencia cardiaca [23].

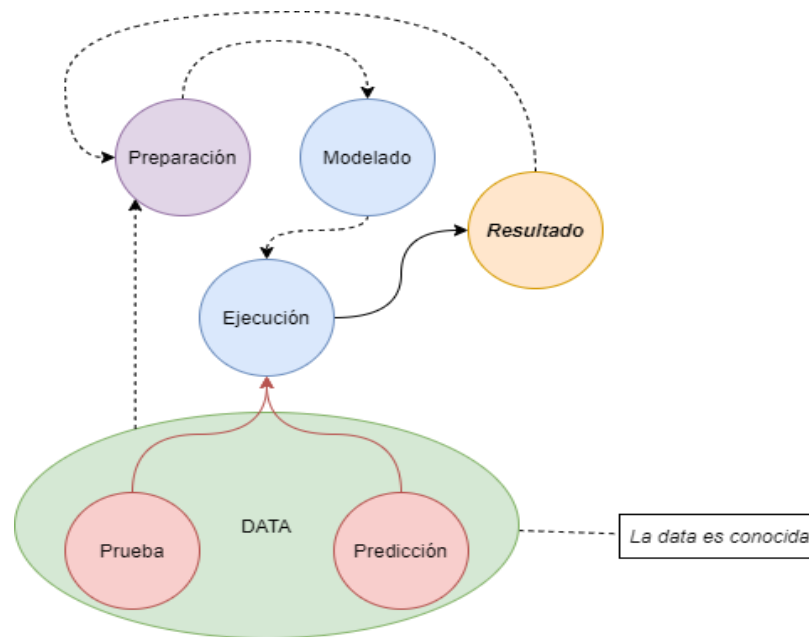
Los algoritmos de aprendizaje automático pueden aprender de formas distintas (véase ilustración 1), pero sobresalen dos: el aprendizaje supervisado que es un método técnico para estimar un patrón de resultados a partir de datos de entrenamiento; o con aprendizaje no supervisado referente a que no hay un ajuste de datos si no del modelo con base en sus observaciones. Esto difiere del aprendizaje supervisado en el que no hay experiencia o información previa para tratar su resultado [23].

3.1.4. Aprendizaje supervisado

Este tipo de aprendizaje parte de un conocimiento ya dado y del que es posible dar un predicado sea falso o verdadero, es esto lo que se quiere saber en este tipo de aprendizaje. Particularmente se requiere de un conjunto de datos, estos pueden ser documentos o imágenes que estén ya **etiquetadas y clasificadas**, donde las mismas etiquetas y clasificaciones serán usadas en el entrenamiento para deducir una función que haga lo mejor posible en el mapeo entre unas entradas definidas por un usuario y una salida esperada [24].

Los datos de entrenamiento constan de duplas (X, Y), siendo X un conjunto de variables que predicen una determinada salida Y. La variable a predecir Y puede ser una variable cuantitativa (como en el caso de problemas de regresión) o cualitativa (como en el caso de problemas de clasificación) [25]. Los algoritmos supervisados más populares son bosques aleatorios, redes neuronales y máquina de vectores de soporte (SVM). En la ilustración 6 se puede visualizar un diagrama de flujo del método supervisado.

Ilustración 6. Esquema de funcionamiento de Aprendizaje supervisado



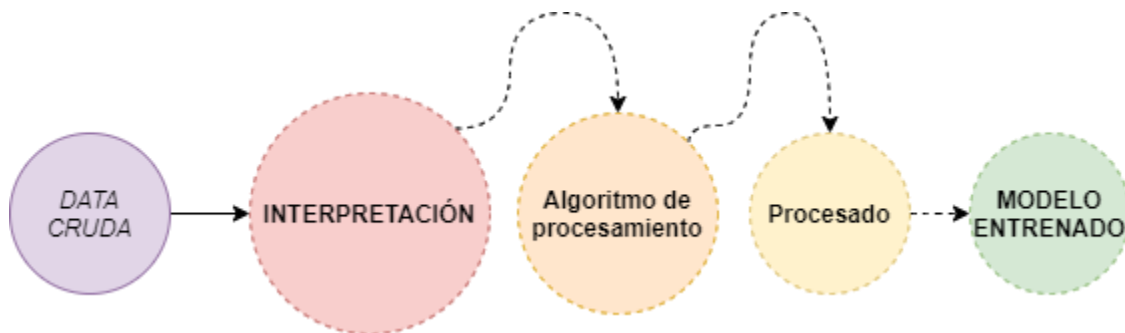
Fuente: Elaboración propia

3.1.5. Aprendizaje no supervisado

El aprendizaje no supervisado es un proceso de aprendizaje independiente; ocurre sin la ayuda de un supervisor. En algunos casos no se conocen los estados de la salida de información y los datos por lo general no están clasificados ni etiquetados. A su vez, los modelos que caben en este tipo de aprendizaje suelen tener un resultado descriptivo más no definen un estado futuro, pues es desconocido en su salida [24].

El objetivo es agrupar los documentos sin conocimientos o intervenciones adicionales de modo que los documentos dentro de un grupo son más similares que los documentos entre grupos. Dentro de este tipo de algoritmos, el agrupamiento o *clustering* en inglés, es el más utilizado, ya que se reparten los datos en grupos que poseen características similares entre sí. Una aplicación de estos métodos es la compresión de imágenes. Entre los principales algoritmos de tipo no supervisado destacan mezclas gaussianas, agrupamiento jerárquico y K-medias [25]. En la ilustración 7 se puede visualizar un diagrama de flujo del método no supervisado. La interpretación se refiere a el conjunto de datos con sentido que luego es pasado por un algoritmo que ordena y adapta la información, para que finalmente en el procesado pase la información adaptada que se evalúa para generar un modelo entrenado.

Ilustración 7. Diagrama de flujo del aprendizaje no supervisado.



Fuente: Elaboración propia

3.1.6. Clasificación y Regresión

En probabilidad y estadística las formas de evaluar un comportamiento o tendencia sobre algún evento pueden ser medidas usando métodos especiales, como lo es el cálculo de probabilidad dentro de un conjunto de eventos y características.

El uso de probabilidad ayuda cuando es necesario resaltar algunos grupos de datos sobre otros. Es aquí donde el aprendizaje automático es fructífero, ya que es una especialización de probabilidad y estadística y al usar ciencias computacionales permite clasificar y distinguir conjuntos de datos [37].

En aprendizaje de máquina la clasificación toma el concepto de clase (o atributo) dentro de un conjunto de características dónde su objetivo es diferenciar atributos, por ejemplo, la predicción del clima; los factores pueden ser varios (pero finitos) como la altura de las nubes, temperatura, nivel sobre el mar, entre otros. Pero sólo uno de ellos puede ser crucial para su aprendizaje, por lo que al usar clasificación en un **modelo con un conjunto de atributos dados**, (por ejemplo Bosques Aleatorios que necesariamente requiere alguna entrada en su modelo), el resultado es un **atributo en particular** entre dicho conjunto para ver qué entidades relacionadas a ese atributo pueden ser asociados en alguna categoría previamente definida [32]. Cuando se habla

de regresión es muy similar a clasificación; sin embargo, su resultado ya no es una cualidad sino un **número** que define un valor límite de algo [37]; puede definir si se tiene mucho o poco de algo, si es o no altamente probable, y así sucesivamente.

3.1.7. Clasificación

La clasificación referencia al problema de identificar a cuál conjunto o subconjunto de categorías pertenece un evento u objeto; por ejemplo, la parte de identificar si un correo es spam o no [38]. Los datos son divididos en grupos que poseen una cualidad (datos cualitativos) y un cierto grado o cantidad de otra característica que es medible (datos cuantitativos), los datos cualitativos indican de qué se compone o **qué características destacables** presentan los objetos de un análisis en una observación [38]–[40], mientras que los cuantitativos definen **cuánto hay de algo**. Viene a ser la parte numérica de la razón en un análisis; un atributo cuantitativo es buscado más para dar una medición de alguna característica contable en un experimento [32].

En este trabajo de grado es necesario usar un modelo de clasificación con base en el fallecimiento de algún paciente (el atributo de clase), pero para efectuar dicha operación es necesario ver con qué otros atributos acompañan el evento de que un paciente fallezca. Se puede observar el hecho que los datos [1] indican si el paciente tuvo o no diabetes, fue anémico, tenía un nivel reducido de plaquetas, aplicaba si tenía un índice de fracción de eyección, medio, alto o bajo. entre otros estados. Pero esto no se detiene ahí. Los datos de estudio son “todos números”. Se indica si fue mujer o no con 0 o 1, se dice si fue anémico o no con 0 o 1, lo cual no implica que estos puedan ser atribuidos al conjunto de fallecidos. Entrando en más detalles con respecto a lo anterior, es posible encontrar datos que parecen ser numéricos, pero que no lo son o mejor dicho, tienen que ser transformados [38], [41].

El identificar estos datos supone ver el sentido de estos, en dónde un sí o no es equivalente a decir si es 1 o 0, pero no es posible **ordenar** los datos de tal manera que estos puedan ser **priorizados** o tengan un **peso** mayor sobre los otros que lo acompañan. El que un **dato tenga orden** implica que pueda distinguirse uno de otro y que éste pueda ser ubicado de cierta manera que se pueda dar preferencia sobre otro conjunto de información cuando es necesario realizar una medición [44], contrario a los valores categóricos donde éstos sólo indican si pertenecen a un conjunto u otro.

3.1.8. Árboles de decisión

Los árboles de decisión son una estructura matemática que con ayuda de probabilidad infunden criterios de aceptación para la toma de decisiones, un contexto de favorabilidad en un evento dado [26]. Permiten al investigador o analista de un estudio definir los mejores caminos u optar por el criterio que mejor se le adapte a su contexto del problema, siendo tema bastante amplio y de constante uso en conceptos de economía e incluso médicos[26].

3.1.9. Bosques aleatorios

Es un algoritmo de machine learning de tipo de aprendizaje supervisado, donde se usa como herramienta de modelado predictivo y no como herramienta descriptiva [27]. Es buena elección para entrenar modelos en corto periodo de tiempo, produce incluso sin ajustar parámetros un buen resultado, se usan para tareas de clasificación como de regresión. Este algoritmo crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable, donde al crecer los árboles se busca la mejor característica entre un subconjunto aleatorio de características [28]–[30].

El algoritmo que emplea selecciona al azar las observaciones y características de un conjunto de datos para construir varios árboles de decisión y luego promedia los resultados para dar una respuesta a una entrada; la desventaja de este algoritmo al usar gran cantidad de árboles puede hacer que sea lento e ineficiente para las predicciones en tiempo real [27].

3.1.10. Máquinas de vectores soporte (SVM)

Las máquinas de vectores soporte son un conjunto de técnicas y algoritmos de clasificación [28], [31], generalmente usados en problemas de selección binaria, pero a lo largo del surgimiento de nuevas necesidades, estas se han usado para otro tipo de problemas, como agrupamiento, regresión y problemas de multi clasificación, reconocimiento de caracteres, categorización de texto o hipertexto entre otras utilidades; dentro de su tarea de clasificación, una SVM, entra en su categoría lineal, ya que su separación o afinidad en los datos consta de eliminar el ruido (información no útil) con el fin de obtener datos transformados y entendibles aplicados a un conjunto de datos, o algún tipo de información agrupada [30].

3.1.11. Redes neuronales

Las redes neuronales emulan la estructura y el comportamiento del cerebro, utilizando los procesos de aprendizaje para buscar una solución a diferentes problemas; son un conjunto de algoritmos matemáticos que encuentran las relaciones no lineales entre conjuntos de datos [30], [32]; suelen ser utilizadas como herramientas para la predicción de tendencias y como clasificadoras de conjuntos de datos. Las redes neuronales artificiales son eficientes en tareas, tales como el reconocimiento de patrones, problemas de optimización o clasificación, y se pueden integrar en un sistema de ayuda a la toma de decisiones, pero no son una panacea capaz de resolver todos los problemas: todo lo contrario, son modelos muy especializados que pueden aplicarse en dominios muy concretos. Se denominan neuronales porque están basadas en el funcionamiento de una neurona biológica cuando procesa información.

Las redes neuronales supervisadas son técnicas para extraer datos a partir de las relaciones de entrada-salida y para almacenar tales relaciones en ecuaciones matemáticas que pueden utilizarse en actividades de pronóstico o en la toma de decisiones [32], [33]. Este tipo de red

requiere que el usuario especifique la salida deseada y, por tanto, la red aprende a detectar la relación entre las entradas y las salidas suministradas, mediante un proceso iterativo y adaptativo; una vez que la red se ha entrenado, se puede utilizar con datos que nunca haya visto o puede ser incluida en un programa para el apoyo a las decisiones [34].

3.1.12. Python

Es un lenguaje de programación creado en 1991, su diseñador (Guido van Rossum) enfocó su utilización para ser un lenguaje de propósito general y de tipado dinámico [14] cuyo enfoque para 2020 tiende ser usado en Inteligencia Artificial y para script en máquinas virtuales [14], [35]. Este es usado como lenguaje multi-paradigma como (imperativo, funcional y orientado a objetos), en el cual se suele realizar programas para emplear inteligencia artificial, y últimamente bajo plantillas como Django se crean páginas web. Se considera un lenguaje interpretado, donde no se necesita de un compilador de código fuente para poder su inicialización; algunas de sus ventajas son la rapidez de aprendizaje, así como su desarrollo de funcionalidades en proyectos de baja escalabilidad [14]. En este trabajo de grado, Python va a ser el lenguaje para utilizar por su robustez y curva de aprendizaje relativamente corta en aplicaciones correspondientes a ciencia de datos.

3.1.12.1. Scikit-learn

Es una de las muchas bibliotecas para el uso de aprendizaje automático de software libre bajo el lenguaje de programación Python [14][15]. En esta se incluye varios algoritmos para poder implementar predicciones a partir de datos, entre ellos están: algoritmos de clasificación, regresión y uso de modelos como máquinas de vectores de soporte, bosques aleatorios, *Gradient boosting*, *K-vecinos cercanos* y agrupamiento espacial basado en densidad de aplicaciones con ruido (DBSCAN). Está diseñada para interoperar con las bibliotecas numéricas y científicas [36].

Scikit-learn es un proyecto el cual empezó con el nombre scikits.learn, un proyecto de *Google Summer of Code* por David Cournapeau. El código de base original fue escrito más de una vez por diversos desarrolladores. En 2010 Fabián Pedregosa, Gael Varoquaux, Alexandre Gramfort y Vincent Michel tomaron el liderazgo del proyecto e hicieron la primera distribución pública en febrero del 2010 [36].

3.1.13. Análisis componentes principales

PCA (Principal Component Analysis) es en pocas palabras, una técnica utilizada para describir un grupo de información con base en nuevas variables que salen de otros métodos estadísticos utilizados para medir la veracidad de la información en cuanto a precisión [39] esta técnica de descripción de datos es comúnmente usada en el análisis estadístico junto a la validación cruzada (Cross Validation)[40], [42].

3.1.14. La validación cruzada y los subconjuntos de pruebas

Es una técnica matemática aplicada en estadística que consiste en evaluar los resultados de un análisis estadístico [40] consiste en la agrupación de pequeños conjuntos de pruebas para probar la independencia de los datos (ver que estos no sigan un patrón o similitudes) [40], para el caso de este proyecto, se usó su estimación en grupos de **k conjuntos** siendo sus valores en 5 y 10 por el número de variables a procesar, el mínimo valor para hacer subconjuntos entre 2 variables es 2, si contamos los 9 que se tienen saldrían aproximadamente **$9!$** de conjuntos, y luego del procesado de datos, pueden salir incluso más, algo que está fuera del alcance de este proyecto de grado, ya que un aprendizaje de este tipo pasaría a tomar semanas sólo en la experimentación de una sola iteración de un modelo [42].

3.1.15. La varianza explicada

Manifiesta que dentro de un conjunto de datos cualquiera y que esté normalizado (dentro de una escala y un rango finito), la varianza es la probabilidad de que tan erráticos están los datos en cuanto a su dispersión estadística con respecto a la media de sus datos [29]. Si la desviación indica que tanto se asemejan entre sí los valores más cercanos a su media determinando el rango mayor y menor, la varianza cumple el objetivo de indicar qué tan dispersos están los datos [29], [31]. En el análisis con python, particularmente con scikit-learn, al usar PCA este muestra un gráfico de pantalla (grado de varianza explicada) donde el usuario puede seleccionar de forma interactiva el número de componentes principales, se usó esta facultad para explicar la dispersión de los datos iniciales [36].

3.1.16. Optimización de hiperparámetros

La optimización y el ajuste de parámetros de un problema consta en la elección de las entradas óptimas para un algoritmo en aprendizaje automático, un hiperparámetro es un parámetro más, no obstante, su valor sirve para controlar el proceso de aprendizaje de un modelo [31]. Uno de los enfoques principales para la selección de los parámetros para medir su efectividad con respecto a métricas de rendimiento (véase F1 Score y ROC AUC) [43] es la búsqueda en tabla (o búsqueda en la grilla) (***Grid Search***), es una técnica que consiste, en pocas palabras, iterar sobre los datos de prueba y luego con los datos de aprendizaje para hallar la mejor combinación de parámetros con respecto al rendimiento dado de cada una de las combinaciones [40], [41], esta técnica va probando combinaciones de posibles parámetros para un modelo dado, sea SVM, Bosques aleatorios, etc.

3.1.17. Los clasificadores binarios y los tipos de error

La forma de discretizar los datos, es decir, transformar variables y funciones en un universo continuo a una forma contable (discreta)[44], esto es usado en clasificación binaria ya que consiste en ordenar un conjunto de datos en dos y sólo dos grupos [45] generalmente apoyándose en

predicados y estructuras evaluativas que resultan en un sencillo Si o No, o de forma booleana en un Falso o Verdadero.

Dentro de la clasificación binaria se pueden encontrar 4 posibles (tabla de verdad) grupos de resultados en sus evaluaciones, donde dos de ellos son errores dentro de la hipótesis en una evaluación, estos clasificadores binarios son ejecuciones que dan una expectativa de un resultado y cuando se ejecuta su evaluación puede resultar en ser o no cierta [37], son expresados en Falsos Negativos, Falsos positivos, Verdaderos positivos, Verdaderos negativos, estos son explicados más adelante en este documento.

- El error de Tipo I: es cuando el investigador niega/rechaza la hipótesis cuando esta es cierta en una población Falso (rechazada) – Positivo (es en realidad verdadera).
- El error de Tipo II: Cuando el investigador **NO** rechaza la hipótesis, y esta es falsa en la población, tómesese como acepta que es falsa (Falso) y es falsa en la población (negativo).

3.1.18. La matriz de confusión

En aprendizaje automático, detalladamente en clasificación estadística, la matriz de confusión es un conjunto de combinaciones de predicción que se utiliza para hacer mediciones de exactitud probabilística con respecto a un conjunto de datos estudiado, permitiendo así, evaluar la precisión del procedimiento de un algoritmo [52]. Es una tabla matriz que mide el error de lo que se *espera versus lo actual* [43].

La matriz usa dos estados de clasificación, siendo falso y verdadero (el estado esperado) contra positivo y negativo (el estado actual). Cuando se ubica en su ejecución se encuentra con las siguientes especializaciones:

- Falsos Negativos (False Negatives - FN): Un falso negativo indica que el resultado esperado es falso (en un predicado) cuando se evalúa, es en sí, el no rechazo de una hipótesis mientras que esta es falsa en una población.
- Verdaderos Positivos (True Positives - TP): Implica que el resultado esperado de una evaluación lógica es verdadero y que en su valor evaluado fue acertado, indica los éxitos de la evaluación.
- Falsos Positivos (False Positives - FP): Un falso positivo indica que en la evaluación de un predicado su resultado predecible es falso, pero en su evaluación este es verdadero, es por consiguiente una falsa alarma, según el contexto este puede ser el peor o mejor resultado en una evaluación probabilística.
- Verdaderos Negativos (True Negatives - TN): Indican que el resultado de un predicado es verdadero en su resultado, pero en su evaluación es falsa.

3.1.18.1. La métrica de la curva ROC (Receiver Operating Characteristic)

De acuerdo a la TDS (teoría de detección de señales) la curva ROC (en español, Característica Operativa del Receptor) es una evaluación de enfoque especializado gráfico que valora en algún sistema de clasificación de dos respuestas, estas respuestas varían según un espacio de descarte [46] este es sencillamente un gráfico que indica una relación de medidas con respecto a los porcentajes de medidas de probabilidades que se esperen sean verdaderas y estas sean ciertas (True Positives) contra las medidas que se esperan que no sean esperadas y sean a su vez ciertas [43].

Para el ejemplo, si se desea saber a qué relaciona una probabilidad de muerte por falla cardiaca, el peor de los diagnósticos es que al paciente se le menciona que **no** va a morir, pero este en verdad muera un “Falso positivo” (FP), si por el contrario se le indica la paciente que **no** va a morir y este no muera, implica que fue un “verdadero positivo” (TP), en concreto la relación gráfica es mostrada de las *cantidades* de casos de TP, con respecto a las cantidades de casos de FP.

3.1.18.2. La medida por medio de F1

Similar al área bajo la curva de ROC, el F-Score, o la medida de F, es una técnica de estadística que consiste en la razón entre la precisión y la exhaustividad de una medida probabilística [47], consiste en la evaluación de la precisión que es el número de TP sobre la cantidad de TP junto a FP.

La exhaustividad hace referencia a TP sobre la cantidad de FN más TP, en pocas palabras emplea la media armónica dentro de dos conjuntos [48].

Modelo Matemático de la evaluación de la medida-F

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$

Fuente: Fawcett, Tom. An introduction to ROC analysis

3.1.18.3. Precisión y Exhaustividad (recall)

La métrica de precisión, en reconocimiento de patrones y clasificación es la fracción de la matriz de confusión que evalúa con base en los valores positivos entre todos sus aciertos positivos y negativos, es decir predice el valor positivo de una función dentro de la matriz de confusión [43]

Modelo evaluativo de la clasificación de precisión en clasificación binaria

$$PRECISION = \frac{TP}{(TP + FP)}$$

Fuente: Fawcett, Tom. An introduction to ROC analysis

La métrica de precisión responde a la pregunta “¿Qué tantos elementos seleccionados son relevantes?”

Recall es la métrica de la matriz de confusión cuya evaluación consiste en los positivos en fracción de los elementos relevantes, que son los falsos negativos y verdaderos positivos, ya que son los aciertos comprendidos dentro de una clasificación [46],[43].

Modelo evaluativo de la clasificación de exhaustividad en clasificación binaria

$$RECALL = \frac{TP}{(FN + TP)}$$

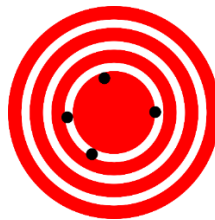
Fuente: Fawcett, Tom. An introduction to ROC analysis

La métrica de exhaustividad responde a la pregunta “¿Qué tantos elementos relevantes han sido seleccionados?”

3.1.18.4. Precisión y exactitud

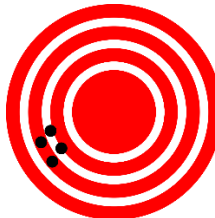
La precisión y exactitud miden la veracidad de una evaluación en una clasificación dentro de un análisis estadístico [49], comprende los positivos verdaderos y positivos negativos contra los falsos positivos y falsos negativos (la verdad versus las negaciones), de acuerdo a la ISO 5725, la precisión y exactitud se comprenden de la siguiente forma:

Ilustración 8 Buena exactitud con mala precisión y buena veracidad



Fuente: BIPM e ISO 5725-2 2019

Ilustración 9 Buena precisión con mala exactitud y veracidad



Fuente: BIPM e ISO 5725-2 2019

Modelo evaluativo de la clasificación de exactitud en clasificación binaria

$$exactitud = \frac{TP + TN}{(TP + FP + FN + TN)}$$

Fuente: Fawcett, Tom. An introduction to ROC analysis

3.1.18.5. Precisión balanceada (sensibilidad y especificidad)

La precisión balanceada mide el grado de verdad y de falsedad de una evaluación en una clasificación binaria, representa la sensibilidad que es lo de verdad positivo y la especificidad que son los casos que son realmente negativos [50], se representa de la siguiente forma:

Modelo evaluativo de la clasificación de sensibilidad y especificidad en clasificación binaria

$$sensibilidad = \frac{TP}{(TP + FN)}$$

$$especificidad = \frac{TN}{(TN + FP)}$$

Fuente: Fawcett, Tom. Basic Principles of ROC Analysis, E. Matz Charles

La precisión balanceada responde a las siguientes preguntas, de acuerdo a los principios básicos en el análisis bajo la curva [49].

- ¿Cuántos objetos de importancia se seleccionaron?
- ¿Cuántos componentes negativos se comportan como negativos?

3.1. ANTECEDENTES

El uso de la inteligencia artificial permite avances significativos en cuanto a la generación de modelos predictivos sobre una población, por lo cual, se da la oportunidad de hacer uso del aprendizaje automático y algoritmos de inteligencia artificial. Los datos de estudio proveídos por los anexos referenciados en el artículo base, indican que fueron tomados por “bootstrapping”, siendo una técnica de muestreo cuyo objetivo es dar estimaciones en una distribución estadística [41], estas estimaciones pueden llegar a ser más exactas con el uso de modelos de clasificación. Los pacientes que participaron en este estudio luego de pasar el periodo de prueba (entre 4 y 8 semanas) mostraron una decaída por insuficiencia cardiaca, al no poderse predecir debido al tiempo y a la sintomatología presentada por el perfil clínico del paciente [1], evidenciando una clara necesidad del uso de modelos clasificadores y gran parte del campo de la inteligencia artificial.

Se han realizado varios estudios que comprenden temas similares al que se relaciona en este documento, tales como en el artículo “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone” [12], está basado en la demostración de la supervivencia y factores más importantes para la detección de insuficiencia cardiaca con el uso de aprendizaje automático para demostrar que la creatinina sérica y la fracción de eyección son las características más relevantes, dando como resultado la aceptación de la

hipótesis sobre estas dos variables y el importante uso que tendría estas herramientas para el apoyo a los médicos a la hora de predecir si un paciente con insuficiencia cardíaca pueda sobrevivir. Esta investigación [12] implementó algunos de los métodos de clasificación que se emplean en este proyecto de grado como lo son bosques aleatorios, redes neuronales y máquinas de vectores de soporte, por ende es una referencia vital para poder demostrar el uso de más de dos características se puede generar una mejor precisión utilizando el mismo conjunto de datos.

Otro ejemplo de la aplicación de ciencias de datos y computacionales como el aprendizaje automático aplicado, es el artículo de la Dra. Andrea Miyazawa [51] el cual es una investigación sobre la aplicación predictiva del uso de aprendizaje automático en padecimientos cardiovasculares. Esta investigación usa la aplicación de redes neuronales y aprendizaje en profundidad para asistir a cardiólogos e ingenieros biomédicos en su diagnóstico proveniente de los datos de dispositivos de asistencia médica como el electrocardiograma, ecocardiograma, pruebas de laboratorio de biopsias entre otros; demostrando una base del 80% de precisión sobre enfermedades cardíacas, particularmente en la irrigación sanguínea en derrames cerebrales.

Pero no sólo han realizado investigaciones directas con ciencias computacionales; fuera de este campo, en el estudio [33], se hace un análisis de componentes inteligentes con una red adaptativa basada en el sistema de inferencia difusa (ANFIS de sus siglas en inglés); este sistema inteligente se ocupa de la combinación de la extracción de características y la clasificación de las formas de onda de la señal Doppler (medidas en la válvula cardíaca), en el que operan los datos relacionados a pacientes con fallas en la válvula tricúspide que da paso al flujo sanguíneo. La precisión de clasificación de este sistema inteligente PCA-ANFIS fue del 96% para sujetos normales y del 93,1% para sujetos anormales, personas sin y con afecciones cardíacas.

Para los estudios relacionados directamente con el conjunto de datos [2], [11], también se han aplicado técnicas de predicción, llegando a conclusiones curiosas, y algo inesperadas, por ejemplo en la investigación [11] donde se demuestra que los atributos más aclaratorios fueron: Suero de Creatinina, la fracción de eyección, si el paciente fumaba, y la edad del mismo; de igual manera, se evidencia que el DataSet presenta un desbalance propio en los datos, por lo que su rigurosidad en su preprocesado la data reducida fue de al menos un 34%. Así mismo, en otro artículo [16] se realiza un estudio con el mismo conjunto de datos, basado en el género del paciente; donde se da prioridad a la relación del tiempo diagnosticado, con respecto a la muerte influido por el género del paciente; al igual que en un estudio previo [11] encuentran inconsistencia en el escalado de la data en los atributos de la edad y el escalado de las plaquetas. Y, si se ve para la investigación [12] se categoriza con el algoritmo de regresión logística por característica, demostrando validez sobre el suero de creatinina y el porcentaje de fracción de eyección del corazón como los atributos más destacados para su estudio, concluyendo que la muestra era pequeña, sólo 299 pacientes con una gran desproporción, así como la falta de atributos claves (como la estatura, peso, masa corporal, estado del colesterol etc).

Para todo lo anterior se remarca la diferencia en esta investigación en el que se usarán los datos ya generados por un estudio previo, para entrenar software de aprendizaje automático, y poder dar una predicción sobre una particularidad de síntomas ya dados; en adición dar predicciones más confiables en términos de las métricas de desempeño, como el uso de F1 y el ROC AUC entre otras más contempladas, también se relaciona el uso del lenguaje Python como herramienta estadística y con posible intervención usando Java para hacer ejecuciones paralelas, teniendo en cuenta que en las investigaciones [2], [11], [12] usaron fundamentalmente el lenguaje R.

3.3. MARCO LEGAL

A continuación, se especificarán algunas leyes que aplican al desarrollo de este proyecto:

- Ley 23 de 1982: es la ley de derechos de autor que contempla los derechos morales y patrimoniales del software (Ley 23, 1982).
- Decreto 1360 de 1989: equipara el software a una creación literaria, aunque el software va más allá del código en sí mismo, su apariencia puede ser objeto de protección igualmente por el derecho de autor (Decreto 1360, 1989).
- Licencia uso Python: Python Software Foundation License o PSFL [52], es una licencia de software libre permisiva, es compatible bajo la licencia GPL de GNU (Licencia Pública General de GNU, a partir de la versión 2.1.1), hace referencia a que mientras no exista un acuerdo económico entre las partes por el producto se puede tomar la licencia como gratuita por estar como código abierto.
- Licencia uso scikit-learn. Licencia BSD [15]. Dicta que la redistribución y el uso en las formas de código fuente, con o sin modificaciones, están permitidos siempre que se cumplan las condiciones.

Este trabajo de grado respetará lo indicado por las anteriores leyes.

4. DESARROLLO DEL PROYECTO

4.1. Descripción breve de los datos a usar

Los siguientes son los datos que se utilizaron en la investigación y detalle en cada uno de ellos, teniendo como base 299 registros iniciales:

Datos categóricos:

- Evento (Event): representa si hubo o no deceso en un paciente que fue diagnosticado con una enfermedad cardiaca (0 No, 1 Sí).
- Género (Gender): especifica el género de la persona, (0 como femenino, 1 como masculino).
- Fumador (Smoking): indica que el paciente tenía antecedentes por fumar (0 No, 1 Sí).
- Diabetes: representa el estado si el paciente mostró síntomas relacionados a diabetes. (0 No, 1 Sí)
- Presión sanguínea (BP): representa si el paciente tenía presión sanguínea alta, sobre sus valores normales (120/80 mmHg). (0 No, 1 Sí)
- Anemia: indica si el paciente manifiesta síntomas anémicos. El rango normal de hemoglobina se define generalmente como 13,2 a 16,6 gramos (g) de hemoglobina por decilitro (dL) de sangre para los hombres y 11,6 a 15 g/dL para las mujeres (0 No, 1 Sí)

Datos Numéricos:

- Edad (Age): representa la edad del paciente, todos los registros son mayores de 40 años.
- Tiempo (Time): indica los días en los que el paciente fue diagnosticado, fueron en promedio 130 días siguiendo el comportamiento de 4 a 285 días.
- Fracción de eyección (EF): indica el porcentaje de contracción en el ventrículo izquierdo, representándose de la siguiente forma acorde los valores [4] yacen en:
 - Bajo riesgo: fracción de eyección mayor o igual a 50%
 - Riesgo moderado: fracción de eyección entre 36-49%
 - Alto riesgo: fracción de eyección menor o igual a 15%
- Sodio (Sodium): representa el nivel de sodio en la sangre (electrolitos) valores normales 136 a 145 mEq/L.
- Creatinina Sérica: refleja los valores de funcionamiento de los riñones con resultado normales de 0.7 a 1.3 mg/dL (de 61.9 a 114.9 μ mol/L) para los hombres y de 0.6 a 1.1 mg/dL (de 53 a 97.2 μ mol/L) para las mujeres.
- Plaquetas: indican el número de plaquetas aproximado que se tiene en la sangre por paciente; la cantidad normal de plaquetas en la sangre es de 150,000 a 400,000 por microlitro (mcL) o 150 a $400 \times 10^9/L$

-
- Creatinfosfoquinasa (CPK): valores normales de CPK total de 10 a 120 microgramos por litro (mcg/L).

4.2. Preprocesamiento de datos

Al inicio la metodología planteada, se analizó la información deseada en el conjunto de datos sobre la cual se basa este proyecto, donde se encuentra que hay un flujo de datos consistentes en categorías. Para la selección de datos se usó validación cruzada representado a K como el número de subconjuntos en las categorías disponibles para tener una cobertura equilibrada de los datos prueba y datos de entrenamiento [40], apoyándose de igual forma en técnicas como reducción y eliminación de variables, como *PCA* o *RFE* [53], [33]

VARIABLES como CPK, que se encuentra en una escala muy dispereja es transformada, a una escala manipulable como se evidencia en [1], [11], ya que no son variables significativas para el estudio sin su normalización. De acuerdo a la información que da en conclusión el artículo en el que se basa este proyecto, y los estudios [2], [11], [12], se hace énfasis en qué medidas las Plaquetas, el Sodio y el CPK en regresión son datos que no aportan cambios significativos en sus estudios, y en el artículo base, se da en conclusión de que atributos como, el *género*, si *fuman* o no, así como la *diabetes* manejan un comportamiento de tendencia de cambio en los resultados de una forma mínima; sin embargo, todas estas variables fueron utilizadas usando escalados y transformaciones de PCA, recordando que es una técnica de reducción de variables.

4.2.1. Técnicas estadísticas usadas

Para los atributos *Fracción de eyección (EF)*, se operan dos conjuntos indicando 1 si está arriba de 50 y 0 en caso contrario; para el nivel de *sodio* los registros que se encuentren entre el intervalo de 0.62 a 0.91 se aplicó normalización por técnica estadística de escalado (máximo y mínimo) al igual que la *creatinina*, donde se especifica que los valores encontrados dentro del rango 0.15 a 0.45 se tomarán como 1 y en caso contrario como 0, la clase de *plaquetas* son normalizadas por medio del conjunto 0.15 y 0.454 usando escalado min-max, al igual que con el sodio y la creatinina, donde los que se encuentren en este intervalo serán 1, y su contrario será 0.

4.2.2. Procedimiento

Debido a que hay atributos de tipo nominal, y no son entradas válidas para los algoritmos, particularmente en el uso de redes neuronales, por lo que se usan en un procedimiento de clasificación para transformarlos en datos numéricos, por ejemplo, si el paciente manifestaba alta presión sanguínea o no, se visualiza como (0, 1) si tenía o no, sin embargo carece de orden, es decir, no hay forma de priorizar los datos en un factor numérico para saber cuál es primero, por lo que en su preparación se visualizarán de la siguiente forma:

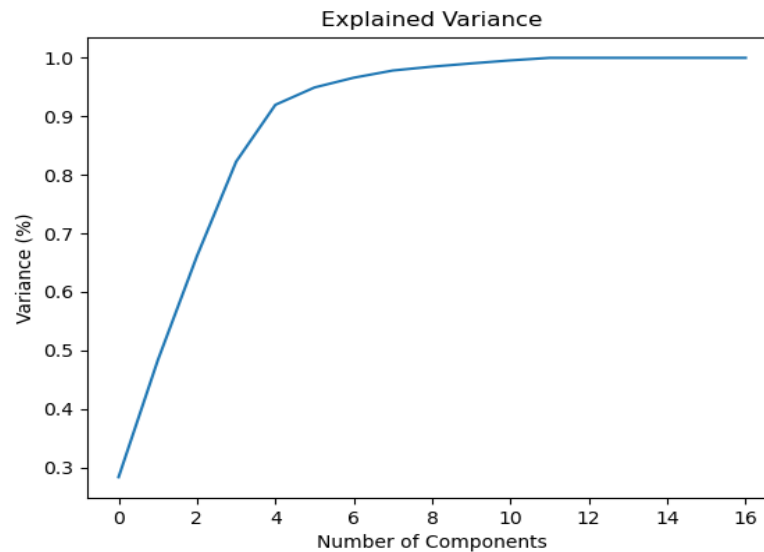
- $BP_0 = (1, 0)$, donde 1 es si NO hubo presión arterial alta, y 0 si hubo, pero no entraba en este ordenamiento.

- BP_1 = (1, 0), donde 0 es sí SÍ hubo presión arterial alta, y 0 cuando no aplicaba en su ordenamiento.

Dando como resultado luego de llevar el anterior proceso una duplicidad en algunas columnas para pasar de tener 13 a 17 atributos sin contar el atributo del evento.

Primer preprocesado de datos:

Ilustración 10. Gráfico de varianza de los datos, con 98% de precisión en $x \sim 10$



Fuente: Elaboración propia

En la ilustración 10, se visualiza en el eje X que los atributos (componentes) más destacables son cerca de 10 de los 17 atributos totales pertenecientes a todo el conjunto de datos. A su vez se usaron dos técnicas para seleccionar según el gráfico de varianza 10 componentes, las cuales son **PCA** (Principal Component Analysis) siendo una técnica utilizada para dar un análisis de los componentes que entran en el análisis del modelo y **RFE** que es otra técnica de una selección hacia atrás de los predictores (modelos) [29], [31], [39], [40]. Esta técnica empieza por construir un modelo sobre un conjunto completo de predictores y computa un puntaje importante para cada predictor. Los predictores menos importantes son entonces eliminados, el modelo es reconstruido y los puntajes importantes vuelven a ser computados. En la práctica, el analista especifica el número de subconjuntos predictores para evaluar, así como el tamaño de cada subconjunto. Por lo tanto, el tamaño del subconjunto es un parámetro de ajuste en la RFE.

Al iniciar con el aprendizaje con máquinas de vectores de soporte, se empleó usando 6 parámetros más importantes con el anteriormente mencionado PCA sobre la data preprocesada, a su vez un número de subconjuntos de $K = 5$ y $K = 10$; usando un vector de soporte cruzado (SVC).

5. RESULTADOS OBTENIDOS

5.1. Modelo usado - Support Vector Machine (SVM)

Para la búsqueda de la mejor configuración del modelo se emplearon los valores en ejecución y valores de los parámetros en la búsqueda de tabla:

- C (constante arbitraria): 0.1, 1, 10, 100.
- Relación de Gamma (Escala del error): 1, 0.1, 0.01, 0.001, 0.0001.
- Clasificador Gamma: scale, auto.
- Peso de clase: dict, balanced, none
- Tipo de kernel: todos los disponibles en la biblioteca (linear, poly, rbf, sigmoid).

5.1.1. Resultados de la búsqueda en tabla para SVM

SVM (mejores posibles valores):

- Constante arbitraria: 1.
- Parámetro Gamma: auto (automático)
- Peso de clase: balanced (Balanceado)
- Kernel: Punto de distribución de los resultados, para este fue rbf.

Lo anterior hace referencia a las posibles combinaciones entre todas las clasificaciones según los parámetros anteriores, en cuanto a sus medidas de clasificación y performance se obtienen los siguientes resultados, teniendo en cuenta que K es el número de subconjuntos seleccionados para hacer la validación cruzada sobre el modelo seleccionado:

Tabla 2. Resultados de métricas de evaluación para Máquinas de vectores de soporte

| K | ROC AUC | F1 | Precisión | Recall | Accuracy | Balanced accuracy | Average precision |
|--------|----------|----------|-----------|----------|----------|-------------------|-------------------|
| k = 10 | 0.877619 | 0.747175 | 0.686486 | 0.828889 | 0.822989 | 0.825873 | 0.777277 |
| k = 5 | 0.874087 | 0.734151 | 0.682080 | 0.804211 | 0.816158 | 0.813508 | 0.760287 |

Fuente: Elaboración propia

5.2 Modelo usado - Random Forest Classifier

Al iniciar la técnica de aprendizaje se usaron los siguientes parámetros para el modelo de clasificación, siendo el número de estimadores las particiones específicas para dar la predicción, se varían en los tres múltiplos de 10 ya que es el promedio del número de variables expresadas entre el atributo de clase y las variables independientes para su predicción más uno, particularmente se usa este estimador en Random Forest, en el artículo [30] se usa una predicción

con hasta 500 árboles, sin embargo es exagerado ya que no hay tal profundidad en nuestro “bosque” ya que son 299 datos, donde varían entre 1 a 17 datos. Los parámetros usados en el modelo fueron:

- n-estimators: 10, 100 y 1000.
- criterio: gini y entropy.
- Máxima profundidad por árbol: 17 a 34 nodos (del número de variables hasta el doble de estas.
- Mínimo de muestras a separar: el mínimo de caminos entre nodos posibles, empezando con una rama binaria (2) hasta 5 hojas, gracias a la documentación de scikit-learn donde especifica que si los valores están en un intervalo estos deben ir en el máximo número de intervalos posibles.
- Máximas características: automática, raíz cuadrada, logaritmo en base 2. Este parámetro especifica el número de características al momento de variar los parámetros de la grilla.
- Máximo número de hojas: 17, ya que no tiene sentido tener más hojas que las columnas (atributos) independientes existentes, según especifica [30].
- Warm Start (inicio rápido): Por temas de rendimiento y recursos de infraestructura se ubica en falso para evitar constantes llamados y más ciclos de reprocesos (call backs) con registros ya usados.

5.1.2. Resultados de la búsqueda en tabla para RF

Random Forest (mejores posibles resultados):

- Criterion: Gini.
- Máxima profundidad del árbol: 17 nodos.
- Máximas características en su distribución: logaritmo en base 2.
- Máximos números de nodos hojas (o finales): 17.
- Mínima separación entre las muestras (Split samples): 2
- número esperado de estimadores: 10

Los resultados de esta grilla para Random Forest son:

Tabla 3. Resultados de métricas de evaluación para Bosques Aleatorios

| K | ROC AUC | F1 | Precisión | Recall | Accuracy | Balanced accuracy | Average precision |
|--------|----------|----------|-----------|----------|----------|-------------------|-------------------|
| k = 10 | 0.835352 | 0.585715 | 0.795000 | 0.534444 | 0.80977 | 0.722937 | 0.724024 |
| k = 5 | 0.831295 | 0.611073 | 0.675685 | 0.648947 | 0.812881 | 0.721723 | 0.699745 |

Fuente: Elaboración propia

5.2. Modelo usado - MultiLayer Perceptron Classifier (Red neuronal)

Para este modelo de redes neuronales se usa el enfoque de 1 capa, ya que no aplica el aprendizaje en profundo, sin embargo por el número de neuronas para esa capa, es necesario hacer varias iteraciones, aproximadamente 5 veces del total de datos ya que como se tenían en un principio datos no ordinales, al pasarlos a ordinales estos fueron incrementados, por lo que el valor por defecto de 100 iteraciones del modelo ofrecido por scikit-learn no es suficiente para la ejecución, por tal razón es necesario habilitar la ejecución en paralelo y manejar el llamado que se hace devuelta al flujo inicial de los algoritmos en cada iteración tal cual especifica la documentación de scikit-learn [36] ya que hay un incremento considerable de resultados en la búsqueda de grilla.

Para tal grilla se usan los siguientes parámetros:

- Una capa oculta por ser un aprendizaje corto con poco volumen, iniciando con un valor de 9 neuronas ya que es el promedio de entradas (17) y salidas (1) neuronas; este valor se iteró y probó hasta llegar a una mejor precisión con 30 neuronas.
- Su parámetro de activación, se usan los proveídos por scikit-learn, logística, tangente hiperbólico y función rectilínea unitaria.
- El solucionador en los modelos SGD y ADAM, ya que no se LBFGS ya que se busca una clasificación estocástica.
- El alpha o valor de regularización (máximo de error) es solo dos valores, muy exacto o 99.9999 de exactitud en el modelo o hasta un máximo de 0.95 en su exactitud.
- El ritmo de aprendizaje es constante y adaptativo, no se usa “invscaling” (o escalado inverso) por el tema de que no es un conjunto de valores que incrementan en el tiempo.
- Las iteraciones deben subir hasta 1500 por temas de convergencia, los datos no empezaron a converger sino hasta las 1400 iteraciones aproximadamente.
- Se habilita el inicio en caliente ya que se requiere todo el procesado de la máquina en donde se usa el modelo, por tal razón el parámetro en la grilla del número de jobs se ubica en “-1” para poder ejecutar en forma paralela y con callback (manejo de registros) las predicciones.

5.2.1. Resultados de la búsqueda entabla para MLP

MultiLayer Perceptron (MLP):

- Secuencia de activación: tangente hiperbólico
- Alpha: 0.001
- Tamaño de la capa de neuronas: 30
- Ritmo de aprendizaje (learning rate): Constante
- Ritmo de aprendizaje inicial: 0.001
- Máximas iteraciones (con paralelismo incluido): 1800

- Solver que dio mejor rendimiento: SGD

Al finalizar, luego de aproximadamente 11 minutos de ejecutar el modelo se obtienen las siguientes métricas para el clasificador de percepción multicapa (1 sola capa).

Tabla 4. Resultados de métricas de evaluación para la red neuronal

| K | ROC AUC | F1 | Precisión | Recall | Accuracy | Balanced accuracy | Average precision |
|--------|----------|----------|-----------|----------|----------|-------------------|-------------------|
| k = 10 | 0,887362 | 0,695747 | 0,744192 | 0,684444 | 0,826322 | 0,782063 | 0,792407 |
| k = 5 | 0.868851 | 0.684662 | 0.732493 | 0.637368 | 0.822881 | 0.772442 | 0.776299 |

Fuente: Elaboración propia

El modelo que ha dado su mejor medida entre las medidas del área bajo la curva (ROC AUC) y el score F1 (nuestras métricas primordiales), son:

Tabla 6: Tabla de resultados de métricas de evaluación MLP, SVM y RF.

Tabla 5. Resultados de métricas de evaluación MLP, SVM y RF

| Métrica/Modelo | MLP | SVM | RF |
|----------------|--------|--------|--------|
| ROC AUC | ~0.887 | ~0.877 | ~0.869 |
| F1-Score | ~0.695 | ~0.747 | ~0.683 |

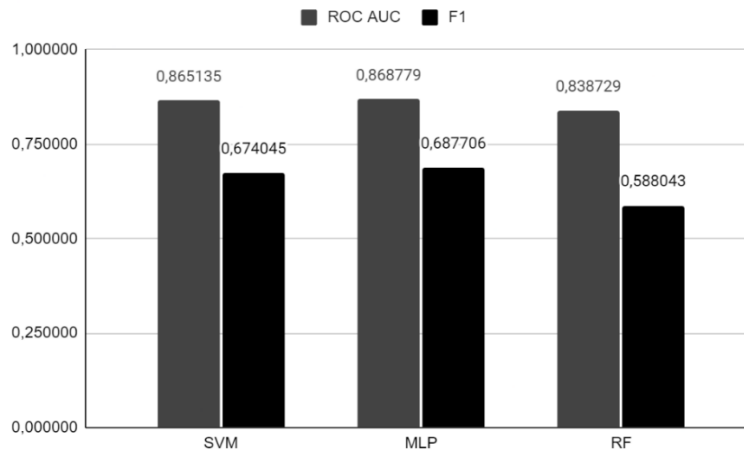
Fuente: Elaboración propia

Con el número de subconjuntos $k = 10$ se observa que con todos los parámetros el valor más acertado recae en el área bajo la curva entre MLP y RF, por una muy mínima diferencia, sin embargo, en el F1-Score, métrica que favorece más a la precisión que la exactitud, los valores más acertados son para SVM y RF.

5.3. Iteraciones con los resultados obtenidos por la búsqueda en tabla

Con los parámetros obtenidos de la búsqueda en tabla se realiza una iteración más con cada uno de los 3 modelos añadiendo un parámetro de forma aleatoria a cada modelo para verificar que la búsqueda de grilla coincida con las métricas de rendimiento que mostró con respecto a los primeros resultados, el recuento general de las iteraciones sobre cada modelo incluido los resultados se evidencian en la siguiente tabla, incluyendo las últimas ejecuciones:

Ilustración 11 Promedio de resultados de las ejecuciones desde su primera iteración hasta la décima y última iteración



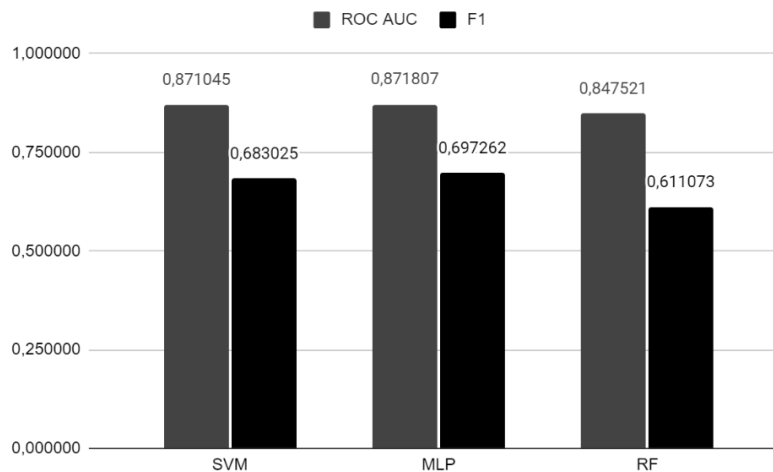
Fuente: Elaboración Propia

5.4. Transición de los resultados

Los modelos tienden a comportarse de forma similar en cuanto a su métrica ROC, con la última iteración en SVM sólo se ha incrementado en un 1% en cuanto a su rendimiento cuando son 10 dimensiones, recordar que k representa el número de subconjuntos, para la validación cruzada entre cada iteración que se ejecute dentro del modelo [42].

Para la anterior tabulación de resultados se encuentra el siguiente gráfico comparando los resultados de su promedio efectuado y su máximo de entre todas sus iteraciones:

Ilustración 12. Gráfico de desempeño de modelos utilizados con sus máximos valores



Fuente: Elaboración propia

5.5. Comparativa de resultados

De acuerdo a lo que se remarca en el estudio de referencia y en el cual se basa este proyecto de grado, la eficiencia del análisis de supervivencia indica un referente de un 81% en su precisión con respecto a la medida de ROC AUC, también en el artículo se usa calibración con respecto a la pendiente de su análisis [1], la cual no es relevante para nuestro estudio ya que no usamos evaluaciones con respecto a señales o curvas de una entrada y salida (por la misma razón que no podemos replicar nuevos datos).

En cuanto a las mediciones realizadas se detalla un promedio para el clasificador MLP de 87.38% en cuanto a su rendimiento, 86.51% para SVM y 83.87% para RF, cabe destacar que hay una desviación estándar de un 1.83% por lo que dichos promedios variaron de la siguiente forma:

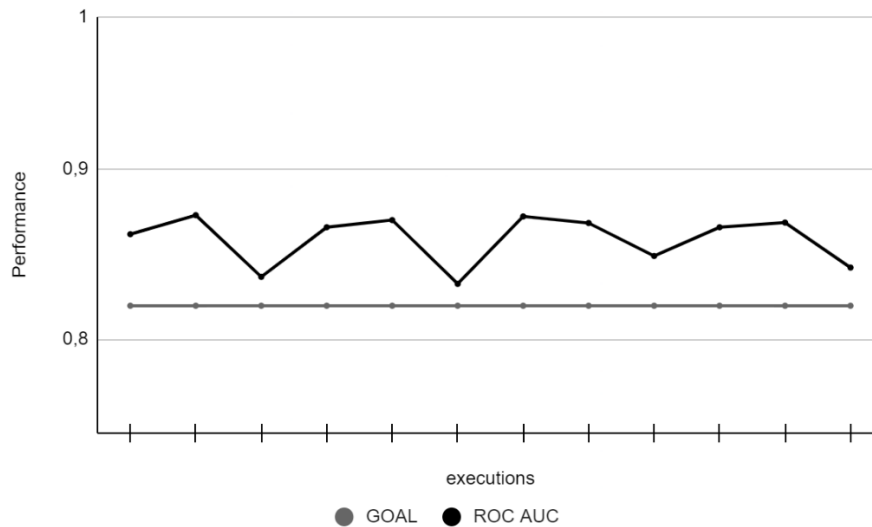
Tabla 6. Comparativa de resultados por modelo según su promedio

| Modelo | Media | desvío +1,83% | desvío -1,83% |
|--------|--------|---------------|---------------|
| SVM | 86,51% | 88,34% | 84,69% |
| MLP | 87,38% | 89,21% | 85,55% |
| RF | 83,87% | 85,70% | 82,04% |

Fuente: Elaboración propia

Comparando los resultados se observa que hay una gran variación con respecto a la dispersión de los datos, a pesar de que se muestra una tendencia lineal, que oscila entre 0.82 y 0.89 en sus topes mínimo y máximo, el modelo presenta un fuerte crecimiento desde 82% del rendimiento hasta un 87%, de ahí en adelante sigue una tendencia constante no mayor al 87.4% en su rendimiento, una buena oscilación entre los porcentajes no debería ser superior al 2.5% entre cambios ya que no nos serviría que su tope mínimo de la media sea menor al resultado obtenido en el artículo basado, por lo cual se realizaron más iteraciones para hacer comprobaciones como se muestra en la tabla 7; para la ilustración 13. se muestra la forma de crecimiento de la curva de ROC y su remarcación en el mínimo ideal de 0.81 con respecto a la investigación [1] siendo el eje X el promedio de la ejecución de los tres modelos usados y el eje Y el porcentaje de rendimiento del modelo con base en la métrica de ROC

Ilustración 13. Gráfico de desempeño de modelos vs. tope de rendimiento por parte de la investigación [1].

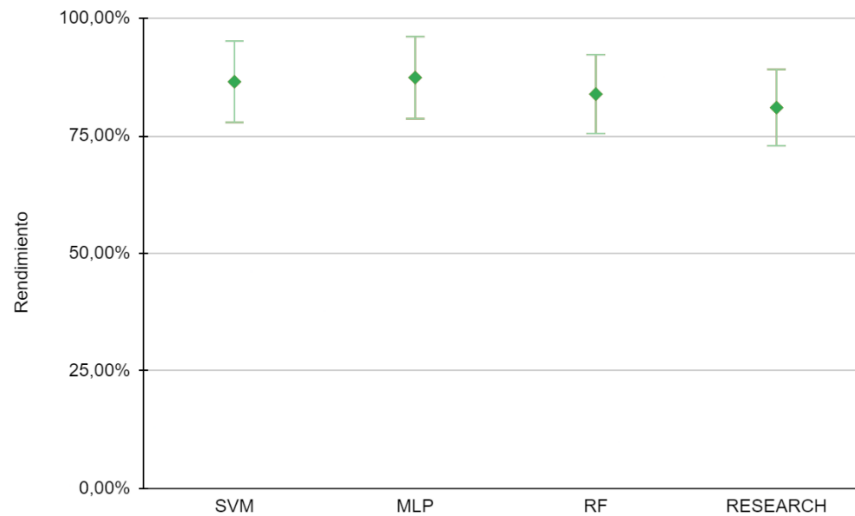


Fuente: Elaboración propia

En la ilustración 14, el resultado tope, 0.81 en comparación con el artículo base [1] es pasado desde la segunda iteración, para cerciorarnos en la precisión del modelo se realizaron esas ejecuciones varias veces 10 iteraciones con cada modelo, terminando en una tendencia en caída ya que el modelo no mostraba mejor rendimiento en cualquiera de sus clasificadores y la curva se iba aplanando, por lo que no mostraba ninguna tendencia luego de cierto número de ejecuciones.

En cuanto al error y la relación con respecto a la investigación se detalla que fue efectivo al aprendizaje usando un clasificador por medio de redes neuronales, con una capa de 30 neuronas, siendo poco menos del doble de los atributos normalizados y preprocesados, como no se buscaba usar aprendizaje en profundo, no fue necesario el uso del desarrollo de otro modelo con más capas de redes neuronales.

Ilustración 14. Variación de los modelos y evaluación del estudio en cuestión



Fuente: Elaboración propia

Por lo anterior, se hace demostración del uso del aprendizaje de máquina en la predicción de supervivencia por padecimiento de insuficiencia cardiaca en pacientes diagnosticados con efectividad del 82% a 85% usando un criterio pesimista con respecto al 81% que menciona el estudio en la métrica de ROC AUC siendo aproximadamente 4% más efectivo con la metodología usada en este trabajo investigativo, adicionalmente, la medida en F1 es lo más cercana a los que se muestra en otras investigaciones [2], [11], [12] que recalcan el trabajo predictivo con menor número de variables, sin embargo con conclusiones particulares sobre el origen y manipulación del conjunto de variables y de datos.

En el siguiente capítulo se muestran las conclusiones, y aprendizaje en retrospectiva al desarrollo de este proyecto de grado, como un introductorio al mundo de la investigación en ciencias computacionales.

6. CONCLUSIONES Y RETROSPECTIVA

Dado los resultados obtenidos utilizando las tres técnicas clasificatorias de aprendizaje automático (SVM, redes neuronales y bosques aleatorios), se puede determinar varias conclusiones; con el artículo [12], se evidencia una mejora en cuanto a precisión de los modelos utilizados ya que en sus resultados más óptimos en cuanto a métricas ROC AUC tuvieron una precisión de 82.2% y en F1 Score de 71.9% utilizando un modelo de regresión logística donde se debe aclarar que solo se tuvieron en cuenta dos características creatinina sérica y la fracción de eyección es por esto que la utilización de más de dos características logrará una mejor precisión del modelo.

En general los resultados obtenidos entre esta investigación y las presentadas en otras [2], [12], muestran una similitud que varía entre un 1.3% a 1.4%. se observa tal cual como se hace mención en algunas investigaciones posteriores [12], y aunque es claro que el mejor modelo resultante es el operado bajo red neuronal en este caso, el *perceptrón de multi capa (MLP)* con su clasificador y que los datos son productivos en cuanto a los resultados que estos prometen, sin embargo, no muy coherentes desde una perspectiva clínica ya que tal cual se hace mención en otros artículos [2], [12] el corto tamaño del data set sumado a que no se esclarece muy bien por qué usar datos clínicos de insuficiencia cardiaca, sin tener en cuenta variables críticas que mejorarían la precisión y rendimiento de los resultados de las métricas presentadas, estos datos, tales como la altura, el peso, nivel de colesterol, grados de afección por falta de sueño o cansancio, entre otros.

En cuanto a la efectividad del proyecto y desarrollo de este, se abre un campo de aplicación bastante grande, y a modo introductorio de lo que es el aprendizaje automático, acá, donde se usaron solo 3 modelos de muchos, pero que fueron considerados con objetivo de comparar y replicar los resultados de instancia de una investigación previa [1] donde nuestro objetivo primordial general fue el de implementar modelos para evaluar y determinar lo más óptimos de un pequeño conjunto seleccionado, de tal forma y que pueda ser posible dar un mejor resultado de lo que se obtuvo en la investigación la cual nos basamos.

REFERENCIAS BIBLIOGRAFICAS

- [1] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," 2017, doi: 10.1371/journal.pone.0181001.
- [2] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: A case study in Pakistan," *PLoS One*, vol. 14, no. 2, p. e0210602, Feb. 2019, doi: 10.1371/journal.pone.0210602.
- [3] "Insuficiencia cardíaca: Colombia cuenta con un novedoso medicamento para tratar la insuficiencia cardíaca | Sanamente | Caracol Radio."
https://caracol.com.co/programa/2017/07/10/sanamente/1499721979_703710.html (accessed Apr. 11, 2021).
- [4] Silvana Estrada Escobar, Pablo Álvarez Rochaital, and Lucía Florio Legnani, "Causas de muerte de pacientes con insuficiencia cardíaca con fracción de eyección reducida en seguimiento por un programa multidisciplinario."
<http://www.scielo.edu.uy/pdf/rumi/v4n3/2393-6797-rumi-4-03-17.pdf> (accessed Apr. 11, 2021).
- [5] organización mundial de la salud OMS, "Enfermedades Cardiovasculares," 2017.
[https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/es/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)).
- [6] "Dieta saludable para el corazón: 8 pasos para prevenir enfermedades del corazón - Mayo Clinic." <https://www.mayoclinic.org/es-es/diseases-conditions/heart-disease/in-depth/heart-healthy-diet/art-20047702> (accessed Apr. 11, 2021).
- [7] "Insuficiencia cardíaca tratamientos, síntomas e información en CuidatePlus."
<https://cuidateplus.marca.com/enfermedades/enfermedades-vasculares-y-del-corazon/insuficiencia-cardiaca.html> (accessed Apr. 11, 2021).
- [8] J. V. Abascal, I. Mayra, G. Mosqueda, L. Vega, and A. Iii, "Riesgo cardiovascular, una herramienta útil para la prevención de las enfermedades cardiovasculares Cardiovascular risk, a useful tool for prevention of cardiovascular diseases," 2011. Accessed: Apr. 11, 2021. [Online]. Available: <http://scielo.sld.cu>.
- [9] "Técnicas de ml en medicina cardiovascular."
https://www.researchgate.net/publication/279850557_Tecnicas_de_ml_en_medicina_cardiovascular (accessed Apr. 11, 2021).
- [10] "Preguntas frecuentes sobre los registros | Los Institutos Nacionales de Salud."
<https://salud.nih.gov/investigacion-clinica/preguntas-frecuentes-sobre-los-registros/> (accessed Apr. 11, 2021).
- [11] O. O. Oladimeji and O. Oladimeji, "Predicting Survival of Heart Failure Patients Using Classification Algorithms," *JITCE (Journal Inf. Technol. Comput. Eng.)*, vol. 4, no. 02, pp. 90–94, Sep. 2020, doi: 10.25077/jitce.4.02.90-94.2020.
- [12] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, Feb. 2020, doi: 10.1186/s12911-020-1023-5.
- [13] Stuart J. Russell y Peter Norvig. *Inteligencia Artificial. Un Enfoque Moderno*. Segunda Edición. Prentice Hall. Madrid, 2004. pp. 482-488.
- [14] "What's New In Python 3.9 — Python 3.9.4 documentation."
<https://docs.python.org/3/whatsnew/3.9.html> (accessed Apr. 11, 2021).
- [15] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python

- Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. Accessed: Apr. 11, 2021. [Online]. Available: <http://scikit-learn.sourceforge.net>.
- [16] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, “Gender based survival prediction models for heart failure patients: A case study in Pakistan,” *PLoS One*, vol. 14, no. 2, p. e0210602, Feb. 2019, doi: 10.1371/journal.pone.0210602.
- [17] Michael A. Chen, “Hipertensión arterial - adultos: MedlinePlus enciclopedia médica.” <https://medlineplus.gov/spanish/ency/article/000468.htm> (accessed Apr. 11, 2021).
- [18] C. García García, “Diabetes mellitus gestacional RESUMEN Artículo de revisión.” Accessed: Apr. 11, 2021. [Online]. Available: www.revistasmedicasmexicanas.com.mx.
- [19] M. Michael A. Chen, “Qué es la enfermedad cardiovascular: MedlinePlus enciclopedia médica.” <https://medlineplus.gov/spanish/ency/patientinstructions/000759.htm> (accessed Apr. 11, 2021).
- [20] A. Castro Beiras and E. Barge Caballero, “Characteristics and prognosis of heart failure with a normal ejection fraction,” *Rev. Esp. Cardiol. Supl.*, vol. 6, no. F, pp. 9F-14F, Jan. 2006, doi: 10.1016/S1131-3587(06)74830-0.
- [21] R. V. Bello, “Escala NYHA (New York Heart Association).”
- [22] J. Carlos de Jesús Montero Rodríguez, R. Roshan Biswal, and E. Sánchez de la Cruz, “Algoritmos de aprendizaje automático de vanguardia para el diagnóstico de enfermedades Advanced Machine Learning Algorithms for Disease Diagnosis.”
- [23] A. Gonzalo, “Tipos de aprendizaje automático.” <https://machinelearningparatodos.com/tipos-de-aprendizaje-automatico/> (accessed Apr. 11, 2021).
- [24] A. A. Arzucan”ozgür, “SUPERVISED AND UNSUPERVISED MACHINE LEARNING TECHNIQUES FOR TEXT DOCUMENT CATEGORIZATION,” 2002.
- [25] J. Luna Gonzalez, “Tipos de aprendizaje automático. La Inteligencia Artificial (IA) está en... | by Javier Luna Gonzalez | SoldAI | Medium.” <https://medium.com/soldai/tipos-de-aprendizaje-automatico-6413e3c615e2> (accessed Apr. 11, 2021).
- [26] R. Barrientos *et al.*, “Árboles De Decisión Como Herramienta En El Diagnóstico Médico,” *Artic. Orig.*, pp. 20–24, 2009, [Online]. Available: https://www.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf.
- [27] M. R. Segal, “Machine Learning Benchmarks and Random Forest Regression,” *Biostatistics*, no. May 2003, pp. 1–14, 2004, [Online]. Available: <http://escholarship.org/uc/item/35x3v9t4.pdf>.
- [28] B. J. Mortazavi *et al.*, “Analysis of Machine Learning Techniques for Heart Failure Readmissions,” *Circ. Cardiovasc. Qual. Outcomes*, vol. 9, no. 6, pp. 629–640, Nov. 2016, doi: 10.1161/CIRCOUTCOMES.116.003039.
- [29] O. Sangrador, M. Arias M, and O. Páez, “INFERENCIA ESTADÍSTICA Inferencia estadística: probabilidad, variables aleatorias y distribuciones de probabilidad.” Accessed: Apr. 11, 2021. [Online]. Available: <https://www.sergas.es/Saude-pu->.
- [30] R. Caruana, N. Karampatziakis, and A. Yessenalina, “An Empirical Evaluation of Supervised Learning in High Dimensions.” Accessed: Apr. 11, 2021. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>.
- [31] C. B. Do, C.-S. Foo, and A. Y. Ng, “Efficient multiple hyperparameter learning for log-

-
- linear models.”
- [32] A. Pitarque, J. F. Roy, and J. C. Ruiz, “Redes neurales vs modelos estadísticos: Simulaciones sobre tareas de predicción y clasificación,” 1998.
- [33] E. Avci and I. Turkoglu, “An intelligent diagnosis system based on principle component analysis and ANFIS for the heart valve diseases,” *Expert Syst. Appl.*, vol. 36, no. 2 PART 2, pp. 2873–2878, Mar. 2009, doi: 10.1016/j.eswa.2008.01.030.
- [34] F. Ocaris Pérez Ramírez and H. Fernández Castaño, “LAS REDES NEURONALES Y LA EVALUACIÓN DEL RIESGO DE CRÉDITO*.”
- [35] P. Alonso Gaona García, C. Enrique Montenegro Marín, and E. Eduardo Gaona Garcia, “MODELO INFORMÁTICO PARA AUTENTICIDAD DE CONTENIDOS MEDIANTE EL CONCEPTO DE WEB OF TRUST SOBRE PLATAFORMAS VIRTUALES LCMS.”
- [36] “scikit-learn: machine learning in Python — scikit-learn 0.24.1 documentation.” <https://scikit-learn.org/stable/> (accessed Apr. 11, 2021).
- [37] J. Piccini, “Arboles de Clasificación y Regresión basados en atributos funcionales y su utilización en el contexto de procesos epidémicos,” *Udelar*, pp. 15–46, 2009.
- [38] R. Hernandez Sampieri, “Análisis de datos cuantitativos.” [https://ciberinnova.edu.co:10004/archivos/plantilla-ovas1-slide/documents-UCN-Canvas/proyecto-integrador-II/lecturas unidad 2/TEMA 1/Capitulo 10-Sampierionlisis de datos cuantitativos 270-335.pdf](https://ciberinnova.edu.co:10004/archivos/plantilla-ovas1-slide/documents-UCN-Canvas/proyecto-integrador-II/lecturas%20unidad%20TEMA%201/Capitulo%2010-Sampierionlisis%20de%20datos%20cuantitativos%20270-335.pdf) (accessed Apr. 11, 2021).
- [39] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, “How many principal components? stopping rules for determining the number of non-trivial axes revisited,” *Comput. Stat. Data Anal.*, vol. 49, no. 4, pp. 974–997, Jun. 2005, doi: 10.1016/j.csda.2004.06.015.
- [40] A. W. Moore, “Cross-validation for detecting and preventing overfitting.” Accessed: Apr. 11, 2021. [Online]. Available: www.cs.cmu.edu/~awm.
- [41] R. Ledesma, “Introducción al Bootstrap: Desarrollo de un ejemplo acompañado de software de aplicación,” *Tutor. Quant. Methods Psychol.*, vol. 4, no. 2, pp. 51–60, 2008, doi: 10.20982/tqmp.04.2.p051.
- [42] “An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike’s Criterion on JSTOR.” <https://www.jstor.org/stable/2984877?seq=1> (accessed Apr. 11, 2021).
- [43] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
- [44] Y. Lu and C. Rasmussen, *Simplified Markov Random Fields for Efficient Semantic Labeling of 3D Point Clouds*. .
- [45] “Linear systems : a state variable approach with numerical implementation : DeCarlo, Raymond A., 1950- : Free Download, Borrow, and Streaming : Internet Archive.” <https://archive.org/details/linearsystemssta0000deca> (accessed Apr. 17, 2021).
- [46] P. a Flach, “The many faces of ROC analysis in machine learning,” *ICML Tutor.*, vol. 20, no. 2, pp. 538–46, 2004, [Online]. Available: <http://www.cs.bris.ac.uk/~flach/ICML04tutorial/ROCTutorialPartI.pdf>.
- [47] Y. Sasaki and R. Fellow, “The truth of the F-measure,” 2007.
- [48] “Using the Price-to-Earnings Harmonic Mean to Improve Firm Valuation Estimates on JSTOR.” <https://www.jstor.org/stable/41948650?seq=1> (accessed Apr. 11, 2021).
- [49] C. E. Metz, “Basic Principles of ROC Analysis.”

-
- [50] “Evaluación de Modelos Predictivos: Ejemplo Python - Machine Learning Blog.” <https://machinelearningenespanol.com/2021/01/12/evaluacion-de-modelos-predictivos/#recall> (accessed Apr. 17, 2021).
- [51] A. A. Miyazawa, “Artificial intelligence: the future for cardiology,” *Heart*, vol. 105, no. 15, pp. 1214 LP – 1214, Aug. 2019, doi: 10.1136/heartjnl-2018-314464.
- [52] “The 3-Clause BSD License | Open Source Initiative.” <https://opensource.org/licenses/BSD-3-Clause> (accessed Apr. 11, 2021).
- [53] K. B. Duan, J. C. Rajapakse, H. Wang, and F. Azuaje, “Multiple SVM-RFE for gene selection in cancer classification with expression data,” *IEEE Trans. Nanobioscience*, vol. 4, no. 3, pp. 228–233, Sep. 2005, doi: 10.1109/TNB.2005.853657.