Universidad Católica de Santa María

Facultad de Ciencias e Ingenierías Físicas y Formales

Escuela Profesional de Ingeniería de Sistemas

ANÁLISIS PREDICTIVO DE MUERTE Y SOBREVIDA DE PACIENTES HOSPITALIZADOS MEDIANTE CLASIFICADORES SUPERVISADOS

Tesis presentada por el Bachiller:
Cordova Roque, Edward Gonzalo
para Optar el Título Profesional de Ingeniero de Sistemas.

Asesor: Mgter. Sulla Torres, José

Arequipa - Perú

2017

UNIVERSIDAD CATOLICA DE SANTA MARIA
URB. SAN JOSE S/N - UMACOLLO

FACULTAD DE CIENCIAS E INGENIERIAS FISICAS Y FORMALES

ESCUELA PROFESIONAL DE INGENIERIA DE SISTEMAS

ANTA MARÍA
ES
2
OR

INFORME DICTAMEN DE BORRADOR DE TESIS

VISTO

El Borrador de Tesis titulado:

Análisis Predictivo de Muerte y Sobrevida de Pacientes Hospitalizados Mediante Clasificadores Supervisados

Presentado por (el) (la) (los) Bachiller (es):

Edward Gonzalo Cordova Roque

Nuestro dictamen es:

Aprobado

OBSERVACIONES: Sin observaciones

Arequipa, 06 de Diciembre de 2017

1635

(5154) 382038 (5154) 252542 ucsm@ucsm.edu.pe http://www.ucsm.edu.pe

0000647

Del 000001 al 200000

AGRADECIMIENTOS

Un especial agradecimiento a mi alma mater la Universidad Católica de Santa María y a la Escuela Profesional de Ingeniería de Sistemas, por ser forjadora de mi preparación profesional, por las enseñanzas ofrecidas y por las amistades obtenidas.

A los docentes de la Escuela, por los conocimientos impartidos, por los consejos y la motivación en cada día de mi aprendizaje.

Al Hospital Regional Honorio Delgado, por su apoyo e información brindada.

DEDICATORIA

A mis padres en especial a mi madre Rosa Roque Coila, por el apoyo incondicional siempre y me da su fortaleza mediante su amor que perdura en el tiempo a pesar de los obstáculos.

A mis hermanas Dayana y Shirley por la paciencia y apoyo incondicional

INTRODUCCIÓN

El uso de la tecnología permite, que muchas aplicaciones de análisis de datos puedan ser realizadas en diferentes sectores económicos del país, muchos de los estudios fueron dirigidos por las potenciales aplicaciones en sectores comerciales y empresariales donde el resultado de dichos análisis ayuda a la toma de decisiones a los gerentes y encargados de dichas instituciones.

Por otro lado, el área de salud es una de las prioridades de nuestro país en ese sentido es importante desarrollar aplicaciones que ayuden a tomar decisiones acertadas en el sector para poder aprovechar los recursos de los mismos y de alguna forma disminuir los índices de mortandad de nuestro país.

Es importante que el uso de técnicas de aprendizaje automático, de minería de datos serán usadas para lograr dichos objetivos, siguiendo metodologías adecuadas para poder encontrar una relación de los fallecidos con los datos que se registran en el hospital.

En general los sistemas de información existentes no realizan este tipo de análisis y las herramientas existentes no se adecuan a los sistemas tradicionales, esto es importante ya que el aporte de este trabajo es relevante para demostrar que los usos de técnicas de aprendizaje automático pueden ser usadas en soluciones de problemas regionales lo cual pone a la tecnología al servicio de la población que se necesita más hoy en día, en aplicaciones tecnológicas de impacto en el sector salud.

Los sistemas hospitalarios usan sistemas de información que día a día se registran en grandes cantidades en los diferentes nosocomios de la ciudad pero en la mayoría de los

casos no son usados para la toma de decisiones, es por eso necesario apoyar a los sistemas hospitalarios en el desarrollo de un sistema que sean capaces de tomar decisiones sobre las diferentes hospitalizaciones que se dan en el Hospital Regional Honorio Delgado, el cual es uno de los principales nosocomios de la ciudad de Arequipa.

Este nosocomio tiene un sistema de información que registra día a día información de hospitalizaciones, datos de los pacientes así como diferentes diagnósticos que los médicos registran, servicios que solicitan y operaciones que se realizan a los pacientes, es usado más para registrar, sin ningún tipo de análisis posterior, es en ese sentido que el objetivo de esta tesis es clasificar los datos de los pacientes para determinar la predicción de muerte o sobrevida de sus hospitalizaciones.

En el capítulo uno se presenta el planteamiento del problema y se presentan el objetivo general y objetivos específicos, así como la solución planteada. En el capítulo dos se presenta el estado del arte de los diferentes trabajos relacionados, así como las bases teóricas de las técnicas empleadas para la solución planteada. En el capítulo tres se describe el marco metodológico usado en la presente tesis, así como los alcances, limitaciones y la propuesta planteada de la tesis para la solución del problema. En el capítulo cuatro los resultados son presentados donde se realiza el análisis de los mismos de las diferentes técnicas utilizadas en la solución del problema planteado. Finalmente se presentan las conclusiones, recomendaciones y trabajos futuros.

RESUMEN

En la actualidad el área de aprendizaje automático viene siendo utilizada en la extracción del conocimiento e información en grandes volúmenes de datos, posibilitando muchas aplicaciones en diferentes sectores productivos, un área de vital importancia en nuestro país es el sector salud, en ese sentido, desarrollar aplicaciones en salud son de suma importancia para prever situaciones que puedan ocurrir, como por ejemplo, en un hospital se registran ingreso de pacientes día a día y dicha información puede ser usada para apoyar a la toma de decisiones.

En ese sentido, este trabajo propone el análisis predictivo de muerte y sobrevida de pacientes hospitalizados mediante clasificadores supervisados, el cual desarrolla las etapas de minería de datos y KDD para el tratamiento de la información de los datos registrados de un sistema de información del Hospital Regional Honorio Delgado, para tal objetivo se utilizan tres técnicas: redes neuronales de tipo *Backpropagation*, clasificador bayesiano y máquinas de vectores de soporte, a pesar que los registros de las personas que fallecen es mínima comparada con la gente que normalmente se trata, el modelo consiguió llegar a un 99.00 % de acierto siendo este un nivel de exactitud y confiabilidad adecuado para que pueda ser útil en la prevención de recursos del hospital entre otras cosas que puedan hacerse.

Palabras claves: Máquinas de vectores de soporte, redes neuronales, clasificador bayesiano, predicción de muerte y sobre vida.

ABSTRACT

At present, the machine learning area has been used in the extraction of knowledge and information in large volumes of data, making many applications in different productive sectors, an area of vital importance in our country is the health, in that sense develop applications in health are of importance to predict situations that may occur, such as in a hospital are registered patients' admission day by day and such information can be used to support decision making.

In this sense, this work proposes the predictive analysis of death and survival of hospitalized patients using supervised classifiers, which develops data mining and KDD stages for the treatment of the information of the registered data of a Hospital Regional Honorio Delgado. information system, for such objective is to use three techniques: backpropagation neural networks, naive Bayesian classifier and support vector machines, although the records of people who die are minimal compared to the people who are normally treated, the model to reach 99.00% of correct accuracy and reliability, so that it can be useful in the prevention of hospital resources among other things that can be done.

Keywords: Support Vector Machine, Naïve Bayes, Neural Networks, Prediction of death over life.

CONTENIDO

INDICES DE FIGURAS

Acrónimos

SVM	*Support Vector Machine*
HRHD	Hospital Regional Honorio Delgado
KDD	*Knowledge Discovery in Databases*
RNA	Redes Neuronales Artificiales
UCI	Unidad de Cuidados Intensivos
SSE	*Sum de Squared Error*
ROC	*Receiver Operating Characteristic*
EHR	*Electronic Healt Register*
MLP	Multicapas Perceptron
ART	*Adaptive Resonance Theory*
RN	Redes Neuronales
MSE	*Mean Square Error*
TIC	Tecnologías de la Información y Comunicación

CAPÍTULO 1: PLANTEAMIENTO DE LA INVESTIGACIÓN.

1.1. Planteamiento del problema.

Los clasificadores supervisados son capaces de construir modelos que optimicen un criterio de rendimiento, utilizando datos históricos o experiencia previa. En ausencia de la experiencia humana, para resolver una disyuntiva que requiere explicación precisa, los sistemas implementados por modelos clasificadores han sido parte importante en la toma de decisiones. A parte, cuando este problema requiere prontitud por su naturaleza, los clasificadores transforman los datos en conocimiento y aportan aplicaciones exitosas muchos de ellos son capaces de hacer sistemas de predicción lo cual es útil para ciertas instituciones como por ejemplo bancos, colegios, hospitales.

En el caso de Hospitales, es importante realizar el estudio de factores de muerte y sobrevida de pacientes del H.R.H.D (Hospital Regional Honorio Delgado), existen estudios estadísticos y aplicaciones salubristas para determinarla, mas no integrados simultáneamente como parte de una probabilidad clasificatoria como modelo. Por ello, este estudio determinará, el uso de un clasificador supervisado más eficiente en tiempo y resultado que permita predecir las clases entre pacientes, así apoyar al personal de salud a tomar la decisión más óptima y a prevenir futuras alzas en el índice de mortalidad de su comunidad. Los problemas que generan el alza de este indicador ya son conocidos y puestos en valor en esta investigación: "Hoy en día existe suficiente evidencia que demuestra que las principales causas de la muerte son debido a los diferentes diagnósticos médicos que presentan los pacientes. Asimismo, sabemos cuáles son las medidas más eficaces y seguras para tratar estas emergencias.

1.2. Objetivo General

- Clasificar los datos de los pacientes para determinar la predicción de muerte o sobrevida de sus hospitalizaciones.

1.3. Objetivos Específicos

- Recolectar y preparar los datos para ser utilizados por los algoritmos de clasificación.
- Determinar y seleccionar las características más relevantes para la predicción.
- Analizar y comparar los algoritmos de clasificación supervisada.
- Elegir el mejor algoritmo de clasificación para casos de hospitalización.

1.4. Pregunta de investigación

¿Se podrá predecir los casos de muerte o sobrevida con los datos de los pacientes hospitalizados mediante algoritmos de clasificación supervisada?

1.5. Línea y sub-línea de investigación

Línea : Inteligencia Artificial

Sub-línea : Aprendizaje automático

1.6. Palabras Clave

Mortalidad hospitalaria, aprendizaje automático, clasificadores supervisados.

1.7. Solución propuesta

1.7.1. Justificación e importancia

El estudio es importante por la necesidad médica de determinar la probabilidad que tiene un paciente hospitalizado para salir con vida o no de una intervención, esto puede permitir tomar mejores decisiones sobre los recursos que la institución utiliza con una propuesta de este tipo se podrá optimizar las hospitalizaciones ya que si se sabe que un paciente no tiene riesgo de fallecer la cantidad de días que se encuentre en el nosocomio puede ser disminuida por otro lado si el riesgo es mayor por la predicción del sistema el monitoreo es constante y las hospitalizaciones de una mayor cantidad de días ayudara a los pacientes a tener una mejor calidad de atención y disminuir en algo el riesgo de fallecer.

Por otro lado la aplicación de los algoritmos clasificadores supervisados como, redes neuronales (perceptrón), clasificador bayesiano y máquinas de vectores de

soporte, aplicadas a casos reales en herramientas computacionales consolidan el campo de inteligencia artificial como área promisoria para futuras aplicaciones médicas esto consolida las investigaciones como útiles en aplicaciones reales dejando de lado la conjeturas de que las tesis solo son investigativas mas no practicas por eso probar con datos reales hace que dicha herramienta sea promisoria para el área de predicción de muerte o sobre vida.

En la actualidad la problemática de muerte y sobrevida en Hospital Regional Honorio Delgado, se debe al indicador determinante de los diagnósticos que presenta cada paciente, el hospital carece de sistemas capaces de apoyar a la toma de decisiones eso debido a la desconfianza en dichos aplicativos y también a la mala organización y almacenamiento de la información que dicho nosocomio maneja, dificultando el uso adecuado y correcto de lo mismo para que se puedan construir aplicativos útiles que puedan ayudar a la toma de decisiones.

1.7.2. Descripción de la solución

Se propone entonces que evaluando cada uno de los indicadores más importantes en la construcción del clasificador, se descarten aquellos que no cumplen las siguientes características:

- especificidad > 90%
- clasificación correcta > 90%
- clasificación incorrecta < 10%
- sensibilidad >90%
- mean Absolute error < 0.1 ideal
- kappa statistic >0.79, >0.9 ideal

- root mean squared error < 0.3, <1 ideal
- relative Absolute error <25%, <1 ideal
- root relative squared <50%, 0 ideal

El clasificador que cumpla con estas especificaciones, se considera como óptimo para la integración en un sistema que evaluará la base de datos que contengan los datos de las pacientes a comparar con un registro nuevo de entrada.

1.8. Hipótesis

Es posible usar aprendizaje automático para la predicción de muerte o sobrevida de sus hospitalizaciones.

1.9. Variables

1.9.1. Variable independiente

Técnicas de aprendizaje automático.

1.9.2. Variable dependiente

Predicción de muerte o sobrevida de sus hospitalizaciones.

CAPÍTULO 2: FUNDAMENTOS TEÓRICOS.

2.1. Estado del arte

En el estado del arte existen trabajos sobre predicción de mortalidad hospitalaria, la mayoría se orienta al uso de la sala de cuidados intensivos y también al uso específico de ciertas pruebas que depende de los diferentes casos clínicos que los pacientes sufren como cáncer o enfermedades coronarias, entre otras.

2.1.1. Consideraciones Iniciales

Es importante notar, que en la presente tesis se intenta predecir la mortalidad de los pacientes hospitalizados por eso es importante ver que modelos de predicción vienen siendo utilizados para los diferentes casos y así poder elegir el más adecuado para resolver este problema. A continuación, describiremos los principales trabajos que han venido desarrollando en diferentes lugares del mundo

- (Rahmanian, et al., 2010) Predicción de la mortalidad hospitalaria y el análisis de la supervivencia a largo plazo después de las principales complicaciones no

cardiacas en pacientes con cirugía cardíaca. Este estudio fue diseñado para investigar la incidencia y los resultados a corto y medio plazo después de las principales complicaciones en los pacientes de cirugía cardíaca. Se determinaron los predictores independientes de mortalidad operativa para crear un modelo para la predicción del resultado. Un foco particular fue el destino de los pacientes después de la aparición de estas complicaciones. Se utilizaron técnicas de regresión lineal para demostrar sus predicciones.

- (Cai, 2016) Predicción en tiempo real de la mortalidad, la readmisión y la duración de la estadía utilizando datos de historiales médicos electrónicos. El objetivo de este trabajo es desarrollar un modelo predictivo de las predicciones en tiempo real de la duración de la estancia, la mortalidad y la readmisión para los pacientes hospitalizados que utilizan registros de salud electrónicos (EHR). Se construyó un modelo de Red Bayesiana para estimar la probabilidad de que un paciente hospitalizado pueda morir en los próximos 7 días. La red utiliza datos administrativos y de laboratorio específicos del paciente y se actualiza cada vez que se dispone de un nuevo resultado de la prueba de patología. Se utilizaron registros médicos electrónicos de 32 634 pacientes ingresados en un hospital metropolitano de Sydney a través del departamento de emergencias entre julio de 2008 y diciembre de 2011. El modelo fue probado en datos de 2011 y entrenado en los datos de años anteriores. Los resultados mostraron que el modelo alcanzó una precisión media diaria de 80%.

- (Gomes, et al. 2010) Desarrolló un modelo de predicción de la mortalidad hospitalaria, basado en datos del Sistema de Información Hospitalaria del Sistema Nacional de Salud. Se realizó un estudio transversal utilizando datos de 453 515 autorizaciones de ingreso hospitalario correspondientes a 332 hospitales en Rio Grande do Sul, Sur de Brasil, en el año 2005. A partir de la relación entre muertes observadas y esperadas, Y esto se comparó con el ranking de la tasa de mortalidad. La técnica utilizada fue regresión logística se utilizó para desarrollar el modelo predictivo usando algunas variables como sexo, edad, diagnóstico y uso de una unidad de cuidados intensivos. Se obtuvieron intervalos de confianza (95%) en los 206 hospitales con más de 365 ingresos por año.

- (Savastano & Cremaschi, 2009) Analiza la mortalidad en la Unidad de Cuidados Intensivos (UCI) del Hospital Central de Mendoza y evaluar el valor predictivo de la escala APACHE II (Evaluación Fisiológica Aguda y de Salud Crónica). Se realizó un estudio retrospectivo y observacional de los pacientes ingresados a la Unidad de Cuidados Intensivos del Hospital Central de Mendoza, desde el 01/11/06 hasta el 31/03/08. Se calculó la distribución de sexos y de edades de la muestra, la estadía promedio, principales motivos de ingreso a la UCI y la puntuación APACHE II en las primeras 24 horas de internación. Se calculó la mortalidad esperada y la mortalidad obtenida global y se analizó el coeficiente entre ambas mortalidades La mortalidad obtenida fue 72% mayor a la mortalidad esperable según la puntuación APACHE II, demostrando esta Escala un bajo valor predictivo en nuestra UCI. La diferencia entre mortalidades podría parcialmente explicarse por la alta prevalencia de entidades con mortalidades subvaloradas por este modelo pronóstico, como pacientes poli traumatizados y

neurocríticos

- (Ticona & Huanca, 2005) Se analizó los factores de riesgo de la mortalidad perinatal hospitalaria en el Perú y determinar su valor predictivo utilizando información del Sistema Informático Perinatal de 9 hospitales del Ministerio de Salud del año 2000. Se incluyó madres con productos ≥1000 g. Para el análisis las tasas de expresaron por mil nacidos vivos (nv), con intervalo de confianza al 95%, regresión logística y curvas ROC. Siendo este uno de los primeros trabajos de predicción de riesgo de mortalidad en nuestro país por es necesario incluirla en el estado del arte.

- (Archibald et al, 2012) Desarrollaron un sistema de puntaje predictivo para identificar pacientes con mayor riesgo de mortalidad intrahospitalaria. Basado en análisis de los datos clínicos de los pacientes de las exacerbaciones de la enfermedad pulmonar obstructiva gestionada en el Reino Unido, recogidos en 11 hospitales siendo un total de 1031 pacientes fueron incluidos en la cohorte de validación. La tasa de mortalidad intrahospitalaria fue del 5,2%. Se identificaron predictores independientes de mortalidad y se derivó un nuevo sistema de puntuación para la predicción de la mortalidad intrahospitalaria. La puntuación incorporó 6 variables clínicas fácilmente obtenibles: acidosis, albúmina, urea, presencia de confusión, puntuación de disnea de MRC y edad. La puntuación mostró una fuerte discriminación, con un área bajo la curva característica de funcionamiento del receptor (ROC) de 0,84.

- (Shams et al., 2012) Analizó la readmisión hospitalaria como una métrica crítica de calidad y costo de la atención médica. Aunque en los últimos años se han llevado a cabo varias intervenciones como la gestión de la transición y la reingeniería, la eficacia y la sostenibilidad dependen de cuán bien puedan identificar y orientar a los pacientes con alto riesgo de Re hospitalización. Basándose en la literatura, la mayoría de los modelos actuales de predicción de riesgo no alcanzan un nivel de precisión aceptable generalmente no consideran la historia de readmisión del paciente y los impactos de los cambios del paciente a lo largo del tiempo a menudo no discriminan entre reingresos planificados e innecesarios. Ellos usaron máquinas de vectores de soporte para su predictor alcanzando un 83.6% de precisión.

- (Francia & Casademont, 2013) Propusieron un modelo de Predicción de la mortalidad intrahospitalaria en medicina interna, analizando los modelos probabilísticos determinando si la edad es un factor de importancia en la determinación de la exactitud de sus predicciones, este trabajo demostró que los modelos fisiológicos muestran una mejor predicción siendo la edad un factor importante en la predicción.

- (Ruiz & Benito, 2016) En su tesis doctoral desarrollaron un predictivo de mortalidad a corto plazo en ancianos ingresados por patología médica, usando la puntuación de *Acute Psicologic Score*, el número de síndromes geriátricos y la concentración de hemoglobina plasmática, siendo variables complejas pero determinantes en el modelo predictivo para personas ancianas, pero también son difíciles de ajustar en modelos lineales proponen que se exploren técnicas más

dinámicas para mejorar la predicción.

- (Cardona, 2012) Por otro vemos que trabajos como este muestran la aplicación de árboles de decisión en la salud pública, demostrando el gran potencial de las técnicas supervisadas en aplicaciones de salud, dado que estas técnicas permiten hacer análisis multicausales de cualquier evento y permiten considerar condiciones demográficas, clínicas, sociales, de accesibilidad a los servicios de salud, resultados de laboratorio o imagenológicos, de las personas con riesgo de desarrollar el evento en salud objeto de interés. Los desenlaces que frecuentemente se predicen a partir de estas técnicas son: presencia de infecciones o enfermedades, asignación de tratamiento, grado de severidad y muerte. Esto permite tomar como una alternativa a los árboles de decisión como alternativa para poder elaborar un modelo predictivo.

- (Serna et al, 2015) Este trabajo muestra la comparación de técnicas de aprendizaje automático, en pacientes de la Unidad de Cuidados Intensivos (UCI) para la detección temprana de los factores de riesgo asociados a las readmisiones, mortalidad, e infecciones en UCI, esto puede incrementar la calidad de la atención y reducir los costos en el largo plazo. Los datos usados provienen de la UCI de uno de los hospitales de alta complejidad en Colombia. En lo que respecta a este artículo, es la primera vez que el aprendizaje automático, se usa en el campo de la salud en este país. Los resultados muestran que las patologías de la aorta, cáncer, enfermedades neurológicas, y enfermedades respiratorias, así como procedimientos invasivos como la diálisis, traqueostomías y broncoscopias, se correlacionan positivamente con la probabilidad de ser readmitido, de morir, y de

adquirir una infección de catéter en la UCI.

- (Sierra Araujo, 2006) El objetivo en el uso de clasificadores supervisados, es construir modelos que optimicen un criterio de rendimiento, utilizando datos o experiencia previa. En ausencia de la experiencia humana, para resolver una disyuntiva que requiere explicación precisa, los sistemas implementados por modelos clasificadores han sido parte importante en la toma de decisiones.

- (Paliouras, 2003) Las enfermedades cardiovasculares han aumentado significativamente en la última década, ocupando el segundo lugar en crecimiento después del cáncer que sigue ocupando el primer lugar de mortalidad a nivel mundial. Este crecimiento ha dado lugar a un significativo aumento de estudios científicos que analizan las señales electrocardiográficas en morfología, amplitud y/o duración desarrollando técnicas automáticas de clasificación de arritmias cardiacas que soporten objetivamente el diagnóstico del especialista Los registros electrocardiográficos obtenidos de equipos Holter son analizados e interpretados por programas propietarios, que luego el médico especialista interpreta para emitir el diagnóstico al paciente, siendo algunas veces demorado y agotador debido al creciente número de registros. Sin embargo, la aplicación de nuevas técnicas de clasificación basadas en el aprendizaje automático, aportan investigaciones sobre métodos de clasificación supervisada y su evolución en los meta-clasificadores, los cuales mejoran el poder predictivo de los sistemas de clasificadores convencionales.

- (Rodríguez Porrero, 2016) Las ciudades inteligentes tienen como objetivo mejorar la calidad de vida de los ciudadanos. Para ello se utilizan las tecnologías de la información y la comunicación (TICs) como herramientas para transformar y mejorar los procesos y actividades de la administración. El servicio de salud en las ciudades inteligentes se enmarca en las áreas de gobierno y servicio, buscando prevenir enfermedades y mejorar la salud de los ciudadanos. Buscando aportar al servicio de salud en una ciudad inteligente, en la actualidad se desarrollan numerosas aplicaciones que permiten analizar datos provenientes de las instituciones de salud para apoyar la toma de decisiones en la ciudad. Dichas aplicaciones son desarrolladas con técnicas de minería de datos, las cuales permiten descubrir conocimiento en grandes volúmenes de datos.

- (Stork, 2001) Por tanto, es interesante desarrollar modelos para predecir la supervivencia de los pacientes que llegan a los servicios de urgencias. Para ello, los médicos aplican habitualmente técnicas que convierten la gravedad de las heridas en un número que representa la probabilidad de supervivencia de los pacientes. Por este motivo, este problema puede ser visto como un problema de clasificación puesto que existen solamente dos valores posibles como salida del sistema: sobrevive y muere. Hoy en día el uso de técnicas de minería de datos se ha extendido en gran medida para abordar problemas de clasificación.

- (Suca, 2016) La obesidad es la base de muchas enfermedades crónicas importantes, y por lo tanto es uno de los mayores problemas de salud en el mundo. El objetivo de este artículo es determinar si una persona de 6-17 años en riesgo de obesidad puede ser identificada a partir de sus registros médicos de crecimiento mediante algoritmos de clasificación. El presente trabajo estudió los pasos previos involucrados en la predicción de obesidad y realizó una comparación de

algoritmos de clasificación de minería de datos para determinar el clasificador más adecuado que mejore la exactitud de predicción en casos de obesidad. Los pasos involucrados en este estudio son: una revisión de los factores de obesidad infantil, colección de datos, pre procesamiento y preparación de los datos, comparación y evaluación de los clasificadores. Los algoritmos de clasificación empleados fueron árboles de decisión (J48), Naive Bayes, SVM y redes neuronales. Los resultados de la evaluación demuestran que el clasificador basado en arboles de decisión (J48) es el clasificador más adecuado para la predicción del tipo de obesidad con una tasa de precisión de 97.23%.

2.1.2. Discusión de los Trabajos

Como se puede apreciar en la revisión bibliográfica, se han propuesto una variedad de sistemas de clasificación para la predicción de la mortalidad hospitalaria. En algunos casos son basados en estándares como el Índice de Comorbilidad de Charlson (CCI) y la clasificación del Grupo Relacionado con el Diagnóstico (DRG) utilizan datos de diagnósticos secundarios para atribuir el riesgo de muerte a los pacientes y pueden aplicarse a bases de datos administrativos.

Otros usan el sistema de la Sociedad Americana de Anestesiología (ASA) se utiliza para clasificar a los pacientes quirúrgicos según su gravedad, desde el riesgo preoperatorio. También están el uso de la sala de cuidados intensivos que usa los sistemas APACHE, APACHE II y APACHE III miden la gravedad de la condición clínica de los pacientes ingresados en unidades de cuidados intensivos

(UCI).

Se puede apreciar también que se usa la puntuación de *Acute Psicologic Score*, el número de síndromes geriátricos y la concentración de hemoglobina plasmática, que son variables específicas en pacientes de la tercera edad. En esta línea también se encuentran trabajos, que demostraron que datos fisiológicos de cada paciente dan mejores resultados en la predicción.

Finalmente tenemos trabajos, que son genéricos en cuanto al uso de sus variables por ejemplo usan los registros de readmisión hospitalaria como un factor importante para la determinación del deceso de un paciente, en esa línea también vimos que en Brasil se hizo un estudio con variables generales como edad, sexo, diagnostico, especialidad, tipo de admisión y uso de UCI, variables que se pueden encontrar en cualquier sistema de información de un hospital.

A continuación, podemos ver un esquema comparativo de los principales trabajos relacionados a la presente tesis.

Tabla 2.1: Esquema comparativo de los puntos importantes a ser considerados como comparación de los trabajos tomados como referencia en la presente tesis.

Autores	Población	Técnicas	Área	Precisión
(Rahmanian, et al., 2010)	Operados del corazón	Regresión Lineal	Cirugía Cardiaca	90%
(Cai, 2016)	Pacientes hospital Sídney	Red Bayesiana	Emergencia	80%
(Gomes, et al. 2010)	Hospitalizados de Rio Grande del Sur	Regresión Lineal	Sistemas hospitalizados en general	95%
(Savastano & Cremaschi, 2009)	Hospitalizados de Mendoza	Coeficiente de correlación	Unidad de cuidados intensivos	72%
(Ticona & Huanca, 2005)	Hospitalizadas embarazadas Perú	Regresión lineal	Perinatal	95%
(Archibald et al, 2012)	Hospitalizados Reino Unido	Regresión lineal	Enfermedades Pulmonares	85%
(Shams et al., 2012)	Re hospitalizados	Regresión lineal	General	83.6%
(Francia & Casademont, 2013)	Intra Hospitalaria	Regresión lineal	Medicina interna	96%
(Ruiz & Benito, 2016)	Ancianos	*Acute Psicologic Score*	Geriatría	95%
(Cardona, 2012)	Servicios de salud en general	Árboles de decisión	Servicios de salud	80%
(Serna et al, 2015)	Hospitalizados Colombia	Bayes, Regresión lineal	Unidad de cuidados intensivos	88%
(Paliouras, 2003)	Enfermos del corazón	Clasificadores supervisados	Cardiología	90%
(Suca, 2016)	Predicción de obesidad	Clasificadores supervisados	Adolescentes (6-17)	97%

2.1.3. Consideraciones finales de la revisión de trabajos relacionados

Como se puede apreciar sobre los diferentes trabajos es importante tomar en consideración el trabajo de (Gomes, 2010) por ser un antecedente de predicción de índice de mortalidad comprobado en Brasil que usa variables que podemos replicar a nuestra realidad dado que esas mismas son registradas en nuestros

sistemas de información de nuestro país, además de dar un modelo para calcular un índice de riesgo que está dado por la siguiente ecuación (1): donde se presenta un coeficiente de correlación lineal donde se pueda dar un índice que indica la mortalidad de los pacientes al ingresar a una hospitalización:

IR = 2 (Sexo) + 6 (edad 40 a 59 años) + 14 (edad 60 años a más) + 13 (diagnóstico I, infección/parásitos) + 8 (diagnóstico II, neoplasia) + 10 (diagnóstico VI, sistema nervioso) + 1 (diagnóstico IX, sistema circulatorio) + 6 (diagnóstico X, sistema respiratorio) + 12 (diagnóstico VIII, signos y síntomas anormales) + 9 (emergencia) + 21 (uso de UCI: uno a dos días) + 17 (uso de UCI tres a siete días) + 23 (uso de UCI ocho días a más). (1)

Este modelo servirá de referencia, para hacer las comparaciones necesarias y así poder demostrar que con técnicas de aprendizaje automático, podemos mejorar las predicciones.

2.2. Bases teóricas de la investigación.

Como se pudo apreciar hay varias técnicas que permiten analizar los datos y que pueden ser replicadas, en la presente tesis se investigará la eficiencia de diversos algoritmos de minería de datos, dado que es importante analizar la velocidad y eficacia de respuesta, la minería de datos es un campo de las ciencias de la computación referido al proceso de intentar descubrir patrones en grandes volúmenes de conjuntos de datos, dicho análisis contemplara diferentes aspectos tales como calidad del conocimiento, margen de error, etc. Así mismo permite analizar e

investigar sobre el modelamiento de almacenes de datos o data warehouse como la información registrada en los sistemas de información.

2.2.1. Regresión Lineal y No Lineal

Regresión es una técnica estadística para el modelado y la investigación de la relación entre dos o más variables. En el análisis de regresión pueden emplearse para construir modelos que permitan predecir o hacer pronósticos en función a los datos a ser analizados. Hay que considerar que los modelos de regresión sólo establecen una relación matemática entre variables los cuales son aproximados por diferentes funciones que pueden ser lineales, cuadráticas, polinomiales, logarítmicas, exponenciales, etc.

Regresión Lineal Simple

El modelo estadístico para este tipo de regresión consiste en aproximar con una línea los valores de la variable dependiente el modelo estadístico para este caso está dada por la siguiente ecuación número (2):

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (2)$$

En este caso podemos ver que los valores de Y dependen de la variable X el cual tiene asociado un conjunto de parámetros que no controlados como el error ε o perturbación aleatoria, factor aun hace que la relación no es perfecta lo que es normal en problemas de regresión, en ese sentido es necesario estimar cual es la distribución error minimice la diferencia de aproximación de las variables

Una forma de poder realizar dicha aproximación es mediante el método de

mínimos cuadrados el cual está dado por la siguiente ecuación número (3):

$$\boldsymbol{SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i)^2} \quad (3)$$

Después se intenta hacer la minimización igualando el error a 0 para obtener las siguientes ecuaciones para la aproximación final de la regresión como se puede apreciar en las ecuaciones (4) y (5):

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} \quad (4)$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (5)$$

Donde los valores de error pueden ser calculados con las siguientes ecuaciones (6) y (7):

$$SS_{xy} = \sum_{i=1}^{n} (\boldsymbol{x_i - \bar{x}})(\boldsymbol{y_i - \bar{y}}) \quad (6)$$

$$SS_{xx} = \sum_{i=1}^{n} (\boldsymbol{x_i - \bar{x}})^2 \quad (7)$$

Finalmente, solo faltaría estimar el error residual el cual está dado por la siguiente ecuación número (8):

$$SR^2 = \frac{SSE}{n-2} \quad (8)$$

Con los parámetros estimados la aproximación de la recta será la que mejor representa los valores de la regresión.

Regresión No Lineal

El uso y aplicación de modelos de regresión no lineales son muy extensos y complejos en general, existiendo una amplia literatura e información sobre el tema. En esta tesis daremos algunas definiciones básicas sobre como aproximar el problema mediante una función no lineal.

Dados *n* pares de observaciones (y_i, x_i), el cual es un modelo de regresión no lineal que se caracteriza por tener un regresor fijo el cual es de la forma $y_i = f(x_i, \theta^*) + \varepsilon_i$ donde los errores $\varepsilon_i$son independientes donde $E(\varepsilon_i)$=0 para $1 \leq i \leq n$. Donde se tienen un vector de parámetros $\theta^* \in \theta \subseteq R$ dichos parámetros son desconocidos. Existen varios métodos para estimar el valor de θ^*, con distintas propiedades los cuales dependen de la información que se tenga sobre el conjunto de ejemplos de entrenamiento igual que en el caso lineal se aplica la minimización de mínimos cuadrados no lineal, teniendo un grado de máxima verosimilitud, cuasi-verosimilitud y otros métodos más robustos.

Uno de los métodos más usados es el de métodos numéricos de Gauss Newton el cual se basa el cual se basa en el método de Taylor de primer orden de la función *f*. Existen otros métodos que pueden ser aplicados como el método de mínimos cuadrados generalizados los cuales pueden ser encontrados con mayor detalle en (Selva, 2013).

2.2.2. Árboles de Decisión

Un Árbol de decisión, es un conjunto de nodos interconectados en el cual mediante el análisis de las ocurrencias e instancias se utilizan para poder determinar la clase a la que pertenece un registro en función a sus atributos que tiene. Existen muchos algoritmos que construyen estos árboles dependiendo de la

técnica que uno elija puede tener buenos resultados algunos se construyen de manera *top down* como lo es el caso del algoritmo ID3, ampliamente utilizado en la solución de muchos problemas (Martínez, 2009).

En este tipo de algoritmo, primero se determina cual es nodo raíz que mejor puede aproximar a la solución general esta elección se basa en la determinación de la ganancia de información, que tienen un atributo al ser usado como nodo de clasificación y en función de las ocurrencias se generan las ramas respectivas del nodo, para calcular dicha ganancia de información estos árboles se basan en el concepto de entropía.

La entropía, mide la heterogeneidad que tienen los datos, eso significa que si tenemos una muestra de unos diez individuos por ejemplo cinco con una enfermedad y otros cinco sanos, podemos decir que existe una máxima heterogeneidad por que la mitad son de una clase y la otra mitad de la otra clase, si por el contrario todos fuesen solamente a una clases por ejemplo si todos están sanos no existirá heterogeneidad porque todos son del mismo tipo, es claro notar que este índice nos puede ayudar a particionar un espacio en función a la cantidad de elementos del mismo. La entropía se calcula mediante la siguiente formula número (9):

$$Entropia(S) = \sum_{i=1}^{n} -p_i log_2 p_i \qquad (9)$$

Donde *S* es la colección de objetos, p_i la probabilidad de los posibles valores *i* las posibles respuestas, para una muestra totalmente heterogenia su valor de entropía es igual a 1 para una muestra por el contrario homogénea es 0.

El valor de la entropía permitirá calcular el valor de ganancia de información que

tiene un atributo el cual está dado por la ecuación número (10):

$$GanInf(S,A) = Entropia(S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} Entropia(S_v) \qquad (10)$$

Donde S es la colección de objetos y A son los atributos del mismo para los diferentes valores que estos puedan tomar

En la siguiente figura 2.1, se puede apreciar un árbol de decisión que muestra los nodos del árbol que son los atributos elegidos para el particionamiento y las ramas las ocurrencias o instancias de dichos atributos. Se puede apreciar también que dicho árbol se puede convertir en reglas lógicas lo que hace más fácil su implementación en lenguajes de alto nivel como el *Prolog*, entre otros.

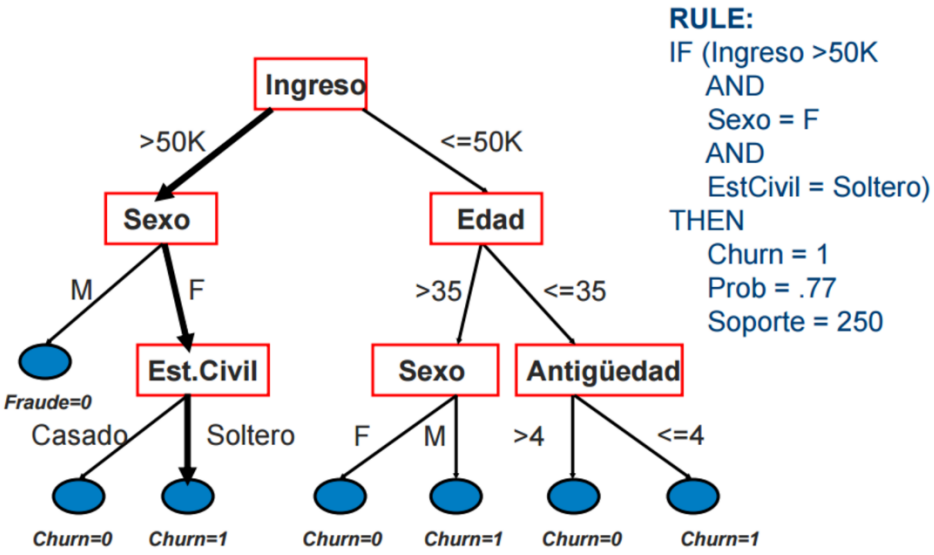

Figura 2.1: Ejemplo de un árbol de decisión que pueden formar reglas (Valdivia, 2015)

Es importante notar también que de dichos árboles se pueden extraer reglas que son importantes para un análisis posterior y más directo de los nuevos datos a clasificar.

2.2.3. Máquinas de Vectores de Soporte

En el artículo de (Vázquez, 2016) muestra el uso de técnicas de máquinas de vectores de soporte (SVM) en aplicaciones médicas comparándolas con redes neuronales y otras más. El uso de SVM, es muy reconocido en el área de aprendizaje automático.

Las SVM son algoritmos de aprendizaje supervisado propuestos por (Vapnik, 1988) y su equipo en los laboratorios AT&T. Dichos métodos están relacionados a problemas de clasificación y regresión de datos. Dado que es supervisado se necesitan de un conjunto de muestras de entrenamiento para así poder etiquetar las clases esto nos permitirá calibrar el modelo lo que es llamo de entrenamiento para luego construir un modelo que prediga la clase de una nueva muestra.

Una SVM puede ser definido como un modelo que, mediante unos puntos de muestra en un espacio dado, este es separando por las clases en 2 partes siendo lo más amplios posibles mediante el uso de un hiperplano de separación el cual es definido como un vector entre los 2 puntos de 2 clases, los más cercanos a la separación se llama vector soporte. Cuando las nuevas muestras de clasificación se ponen en correspondencia con la superficie de separación de dicho modelo, en función de los espacios a los que pertenezcan, pueden ser clasificadas a una o la otra clase.

En este modelo, es importante el concepto de "separación óptima" el cual es la característica principal donde reside la característica fundamental de las SVM:

dado que este tipo de algoritmos lo que hacen es buscar el hiperplano de separación que tenga la máxima distancia (el cual es denominado margen) con los puntos que estén más cerca de él mismo. Por eso también a las SVM a veces se les conoce como clasificadores de margen máximo. De esta forma, toman como referencia el margen de separación de los puntos del vector que son etiquetados con una categoría y estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado de la superficie de separación. Los SVM inicialmente fueron considerados como pertenecientes a la familia de los clasificadores lineales.

En la literatura de los SVMs, los vectores son llamados de atributos a la variable a predecir y características a los atributos que son transformados y que serán usados para la definición del hiperplano. De tal manera que se realiza la elección más adecuadas de la representación del universo analizado, y se realiza dicho proceso mediante la selección de las principales características.

Los puntos más cercanos al hiperplano se le llaman vector de soporte. Los modelos basados en SVMs siguen los mismos principios de los clasificadores lineales y están directamente relacionados con las redes neuronales. Una de las dificultades es cuando los datos no son linealmente separables pero dicho problema es superado mediante el uso de una función denominada kernel, los cuales resultan de un método de entrenamiento alternativo para clasificadores, que permiten el cambio de dimensión del problema original mediante funciones especiales que pueden ser: polinomiales, funciones de base radial y e incluso el mismo perceptrón multicapa. Este *kernel* realizara la separación es mediante una

línea recta, un plano recto o un hiperplano N-dimensional. Desafortunadamente los datos a estudiar no se suelen presentar en casos de dos dimensiones como en el ejemplo anterior, sino que un algoritmo SVM debe tratar con, más de dos variables predictoras, curvas no lineales de separación, casos donde los conjuntos de datos no pueden ser completamente separados, y clasificaciones en más de dos categorías.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio de funciones *Kernel* ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal. Como se puede ver en la siguiente figura 2.2 se muestra un ejemplo de separación lineal donde se observan dos conjuntos de datos los de color negro y las de color blanco, como se puede ver están bien diferenciadas y separadas lo que indica que con una línea es posible separar el espacio el problema es que en un inicio no se sabe dónde será la línea correcta, como se aprecia hay 3 hiperplanos de separación el de color verde no consigue separar las dos clases lo que nos indica que no es una solución buena, por otro lado el hiperplano dos si consigue hacerlo, pero si uno aprecia la separación entra clases no es maximizada como si lo hace el hiperplano tres, que es lo que SVM intenta garantizar al maximizar la separación entre clases.

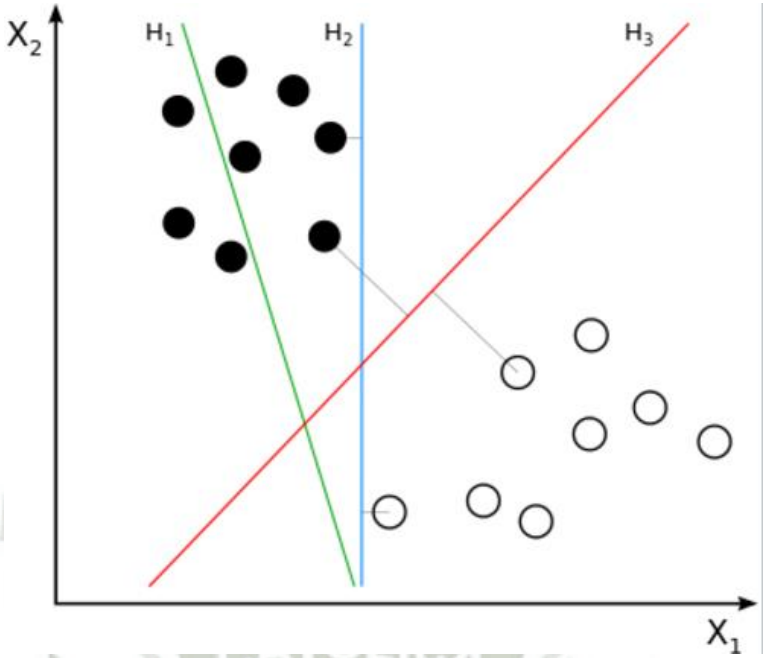

Figura 2.2: Ejemplo de Kernel lineal con relación a otros hiperplanos de separación (Wang, 2005)

Es importante notar que la característica principal de SVM es que no tiene mínimos locales al maximizar dicha separación no se tendría otra solución que fuese mejor que la que encuentra SVM para un caso de clases bien separadas como se pudo ver en el gráfico.

Para poder realizar la separación SVM realiza una optimación para encontrar los vectores frontera entre clases como se puede ver en la ecuación número (11):

$$\min_{f,\xi_i} \|f\|_K^2 + C\sum_{i=1}^{l}\xi_i$$

$$Yi, f(xi) \geq 1 - \xi i, \text{ for all } i \quad \xi \geq 0 \tag{11}$$

SVM para la clasificación hay una información dual que debe ser resuelta en la ecuación número (12):

$$\min_{\alpha_i} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \qquad (12)$$

$$0 \leq \alpha i \leq C$$

Las variables ξi son llamadas de *slack* y miden el error cometido para un punto (*xi*,*yi*). El entrenamiento es realizado con todos los puntos llegando a ser una complejidad muy alta para optimizar la distancia de todos con referencia al plano existen varias propuestas para mejorar dichos algoritmos de optimización.

2.2.4. Clasificador Bayesiano

Es un clasificador probabilístico, basado en el teorema de Bayes (probabilidad condicional de que *A* ocurre si ocurre *B*, de *B* si ocurre *A*, o si simplemente ocurre *A*. Una ventaja de este modelo es que no requiere más que una pequeña cantidad de datos para "entrenar" y determinar lo que ocurrirá basado en las mismas características o variables predictoras, y asume que las variables predictoras son independientes entre sí. En general, se determinan probabilidades para cada clase y escoge la clase con la probabilidad más alta.

(Guareno, 2016) El aprendizaje estadístico desempeña un papel clave en muchas áreas de la ciencia, las finanzas y la industria. Algunos ejemplos de problemas de aprendizaje son: predecir si un individuo que ha tenido un accidente de tráfico volverá a tener otro accidente, en base a su conducción; predecir el valor de una acción en seis meses desde el momento del estudio, sobre la base de las medidas de rendimiento de la empresa y los datos económicos; o identificar los factores de

riesgo para una enfermedad, basándose en variables clínicas y demográficas. En particular, el aprendizaje automático es una parte importante del aprendizaje estadístico, con gran influencia en los campos de la minería de datos y la inteligencia artificial.

(Torres, 2016) analiza la deserción de los estudiantes mediante técnicas de minería de datos y obtener un modelo que fuese capaz, de clasificar estudiantes desertores a partir de los datos socioeconómicos y académicos de los estudiantes de carreras de pregrado en la Universidad Arturo Prat, Chile. Para el desarrollo de este proyecto se utilizó CRISP-DM basado en métodos bayesianos, con el fin de evaluar su comportamiento, encontrándose que el algoritmo Naive Bayes resulto ser el más adecuado para dar respuesta a los objetivos del negocio, dados los niveles de sensibilidad alcanzados.

2.2.5. Redes Neuronales

Una red neuronal artificial (RNA), es un modelo computacional inspirado en las redes neuronales biológicas y puede ser considerada como un sistema de procesamiento de información con características salientes, tales como aprendizaje a través de ejemplos, adaptabilidad, robustez, capacidad de generalización y tolerancia a fallas (D. Hush, 1993).

La RNA puede ser definida como una estructura distribuida, de procesamiento paralelo, formada de neuronas artificiales (o llamados elementos de procesamiento), interconectados por un gran número de conexiones (sinapsis), los cuales son utilizados para almacenar conocimiento y que está disponible para ser utilizado (Haykin, 2001 (D. Rumelhart, 1986)). Una red neuronal es caracterizada por

propiedades de sus neuronas, por la arquitectura de la red (la topología que la red puede poseer) y por los algoritmos de aprendizaje.

Una neurona artificial, es una unidad de procesamiento de información de redes neuronales. El modelo de neurona más conocido es la neurona de McCulloch-Pitts (McCulloch and Pitts, 1943). En esa figura podemos ver ***N*** señales de entradas representadas por las variables *X1, X2, X3...XN*. La conexión de una neurona con índice *i* a la neurona *j* es representada por un peso w_{ij}, que determina el nivel de influencia de la neurona *j* para la neurona i. Si el valor de w_{ij} es positivo es dicho que la sinapsis es ex citatoria y es inhibitoria se fuera negativo. Existen dos etapas de procesamiento para cada neurona: suma y activación. En la figura 2.3, se puede apreciar el modelo de red neuronal donde la salida de una neurona se conecta con la entrada de otra.

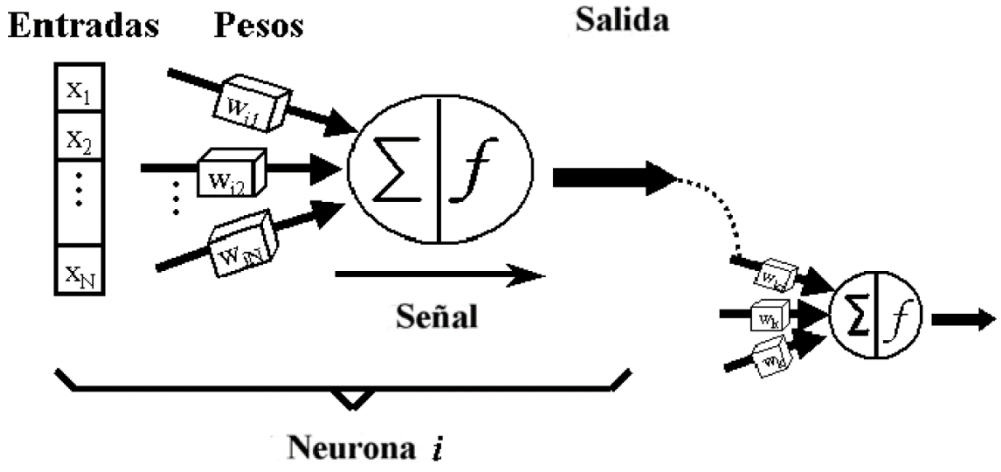

Figura 2.3: Ejemplo de modelo de una neurona artificial (Gutierrez, 2006)

En la primera etapa, las señales de entrada x_j y los pesos w_{ij} son combinadas por la sumatoria de entradas por los pesos donde la salida es y_i es llamado *estado interno* de la neurona *i*.

En la segunda etapa, la salida de la neurona es generada a través de aplicación de una función de activación: $x_i\ (t + 1) = f\ (y_i\ (t))$ donde la salida de la neurona es representada por x_i y f es la función de activación aplicada al estado interno de la neurona, que limita el nivel de activación de la neurona. Generalmente, $x_i \in [-1,1]$ o $x_i \in [0,1]$, en el caso de x_i será un valor continuo y $x_i \in \{-1,1\}$ o $x_i \in \{0,1\}$, en el caso discreto.

La definición de la arquitectura, es un punto importante en el modelaje de una red neuronal, porque ella restringe el tipo de problema que puede ser tratado. Una red también puede estar formada por múltiples capas, las cuales pueden ser clasificadas en tres grupos: capa de entrada, capa intermedia u ocultas y capa de salida como se puede ver en la siguiente figura 2.4 donde esta una arquitectura totalmente conectada de una red neuronal.

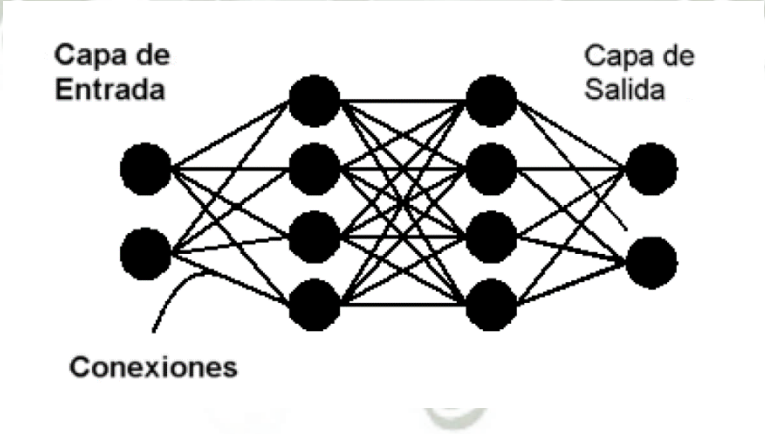

Figura 2.4: Modelo de Arquitectura FeedForward (Haykin, 2001)

Basado en flujo de las señales, las redes neuronales también pueden ser clasificadas en dos tipos: *FeedForward* y redes *Recurrentes*.

- Redes *FeedForward.*

  La estructura de una red *FeedForward en* la figura 2.5 consiste en capas de neuronas en la cual la salida de una neurona de una capa, alimenta todas las neuronas de la capa siguiente. El aspecto fundamental de esta estructura es que no existen lazos de realimentación. La red *MultiLayer Perceptron* (MLP) es un tipo de red *feedforward* (Rumelhart, 1986), (Freeman J, 1991).

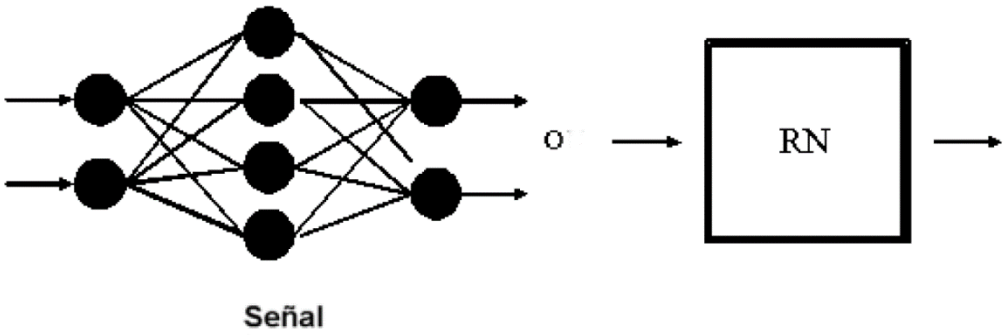

Figura 2.5: Arquitectura Feedforward (Hilera González, 2000)

- Redes *Recurrentes*

  Las redes recurrentes son aquellas que poseen conexiones de realimentación, como puede ser vista en la Figura 2.6, las cuales proporcionan comportamiento dinámico. El modelo de Hopfield es un ejemplo de red neuronal recurrente.

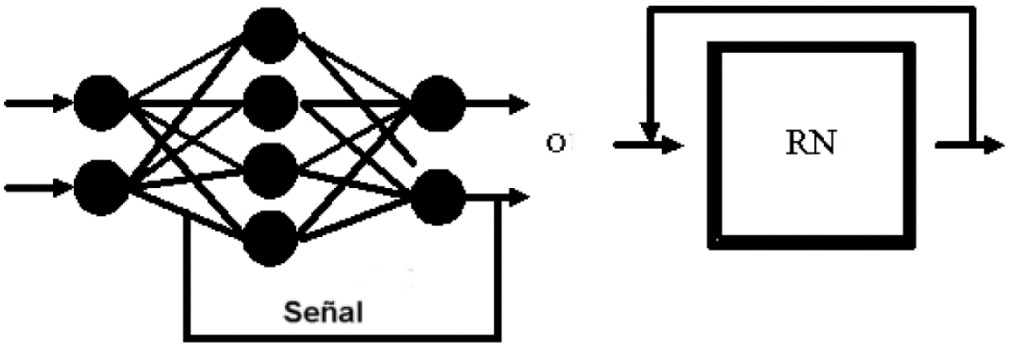

Figura 2.6: Arquitectura Recurrente (Hilera González, 2000)

En general, los siguientes parámetros son importantes para definir la arquitectura de una red neuronal: número de capas, número de neuronas en cada capa y tipo de conexión entre las neuronas, que define la red si es *feedforward* o *Recurrentes*

Algoritmos de Aprendizaje de una RNA

Una propiedad importante, de las redes neuronales es la habilidad de aprender a partir de su ambiente. Eso es hecho a través de un proceso iterativo de ajustes aplicado a sus pesos de conexiones entre las neuronas, denominado *entrenamiento*. Existen muchos algoritmos de aprendizaje. Cada uno sirve para determinadas redes neuronales. Entre las principales se tienen:

- Aprendizaje por Corrección de Error: algoritmo muy conocido basado en la regla Delta, que busca minimizar la función de error usando el gradiente descendente. Este es el principio usado en el algoritmo *BackPropagation*, muy utilizado para el entrenamiento de redes de múltiples capas como la *Multilayer-Perceptron* (MLP) (Freeman, 1991).

- Aprendizaje Competitivo: en el cual las neuronas de una capa compiten entre sí

por el privilegio de permanecer activos, tal que la neurona con mayor actividad sea el único a participar del proceso de aprendizaje. Es usado en los Mapas de Kohonen (Kohonen, 1988) y redes ART (Carpenter G., 1992).

- Aprendizaje Hebbiano: si dos neuronas están simultáneamente activas la conexión entre ellos debe ser fortalecida caso contrario será debilitada (Hebb, 1949). Utilizada en el Modelo de la red Hopfield (Hopfield, 1982).

- Aprendizaje de Boltzmann: es una regla de aprendizaje estocástica obtenida a partir de principios de la teoría de la información y de la termodinámica. El objetivo del aprendizaje de Boltzmann es ajustar los pesos de las conexiones de tal forma que el estado de las unidades visibles satisfaga una distribución de probabilidades deseada en particular (Ackley D., 1985).

Otro factor importante es la manera por la cual una red neuronal se relaciona con el ambiente. En ese contexto existen los siguientes paradigmas de aprendizaje:

- Aprendizaje Supervisado: es utilizado un agente externo que indica a la red la respuesta deseada para el patrón de entrada.
- Refuerzo: es una red variante del aprendizaje supervisado en la cual el aprendizaje sucede de la interacción de la red con su ambiente, indicando a la red solamente una crítica de la corrección de la salida de la red y no a la respuesta correcta en sí.

- Aprendizaje No Supervisado (auto organización): no existe un agente externo indicando la respuesta deseada para los patrones de entrada. Este tipo de

aprendizaje es utilizado en los modelos de Mapas de Kohonen (Kohonen, 1988), redes ART1, AT2 (Carpenter G., 1992) (Sánchez, 2016) (Tase, 2016).

Propagación hacia atrás (*Back-Propagation*)

"*Back-propagation*" es una técnica para resolver el problema de asigna miento de crédito mencionado por Minsky y Papert en su libro *Perceptrons*. David Rumelhart, es una persona asociada con la invención de las redes de back-propagation. David Parker introdujo un algoritmo similar casi al mismo tiempo y otros han estudiado redes similares.

Una red de perceptrones es capaz de entrenar los nodos de salida para aprender a clasificar patrones de entrada, dado que las clases son "linealmente separables". Las clases más complejas no-linealmente separables pueden ser separadas con una red multicapa. La responsabilidad del error se fija al propagar el error de salida hacia atrás a través de las conexiones de la capa previa. Este proceso se repite hasta que se llegue a la capa de entrada. El nombre "*back-propagation*" se deriva de este método de distribuir la culpa por el error.

La red de Propagación hacia atrás

La típica red de propagación hacia atrás se tiene una capa de entrada, una capa de salida y por lo menos una o más capas ocultas. No hay límite teórico para el número de capas ocultas, pero típicamente serán una o dos para aproximación de funciones. Algún trabajo se ha hecho que indica que un máximo de cuatro capas (tres ocultas y una de salida) son requeridas para resolver problemas más complejos de clasificación de patrones. Cada capa está totalmente conectada con su posterior donde las variables serán las siguientes.

- $x_j^{[s]}$ actual estado de salida de la j-ésima neurona en la capa *s*.
- $w_{ji}^{[s]}$ peso en la conexión entre la i-ésima neurona de la capa *(s-1)* y la j-ésima neurona en la capa *s*.
- $I_j^{[s]}$ suma ponderada de las entradas hacia la j-ésima neurona en la capa *s*.

Un elemento de la red de propagación hacia atrás transfiere entonces sus entradas como sigue en la ecuación número (13):

$$x_j^{[s]} = f\left(\sum_i (w_{ji}^{[s]} \cdot x_i^{[s-1]})\right) = f\left(I_j^{[s]}\right) \tag{13}$$

Donde f es tradicionalmente la función sigmoidal, pero puede ser cualquier función diferenciable. La función sigmoidal se define como se ve en la ecuación número (14):

$$f(z) = \left(1.0 + e^{-z}\right)^{-1} \tag{14}$$

Propagando hacia atrás el error local

La red tiene una función de error global "*E"* asociada a ella, que sea una función diferenciable de por los pesos en la red. La función de error actual no es importante para entender el mecanismo de propagación inversa. El parámetro crítico que se pasa hacia atrás a través de las capas se define como en la ecuación número (15):

$$e_j^{[s]} = -\partial E / \partial I_j^{[s]} \qquad ;)$$

Veremos después que esto se puede considerar como una medida del error local en el j-ésimo nodo en el nivel *s*.

Utilizando la regla de la cadena dos veces seguidas nos da una relación entre el error local y un nodo en particular en el nivel *s* y los errores locales en los niveles *s+1* como se ve en la ecuación número (16):

$$e_j^{[s]} = f'\left(I_j^{[s]}\right) \cdot \sum_k \left(e_k^{[s+1]} \cdot w_{kj}^{[s+1]}\right) \tag{16}$$

Si la función sigmoidal se define la salida, entonces su derivada puede ser expresada como una función simple de sí misma como sigue en la ecuación número (17):

$$f'(z) = f(z) \cdot \left(1.0 - f(z)\right) \tag{17}$$

Por lo tanto, el error puede ser reescrita como en la ecuación número (18):

$$e_j^{[s]} = x_j^{[s]} \cdot \left(1.0 - x_j^{[s]}\right) \cdot \sum_k \left(e_k^{[s+1]} \cdot w_{kj}^{[s+1]}\right) \tag{18}$$

Dado que la función de transferencia es una sigmoidal. El término de la sumatoria que es utilizado para retro-propagar errores es análogo al término de suma que es utilizado para propagar hacia delante las entradas por la red. Entonces el mayor mecanismo en una red back-propagation es el propagar las entradas por las capas hasta la capa de salida, determinar el error en la salida, y luego propagar los errores hacia atrás desde la capa de salida hasta la capa de entrada. La multiplicación del error por la derivada de la función de transferencia escala el error.

Minimizando el error local

El objetivo del proceso de aprendizaje, es el minimizar el error global "*E"* del sistema al modificar los pesos. Esta subsección mostrará cómo hacer esto basándonos en el conocimiento del error local en cada nodo, ya que dicho error

es considerado como local debido a que cada nodo es calculado en función a los errores de las capas anteriores en el cálculo de su función de salida.

Dado el actual conjunto de ponderaciones $w_{ji}^{[s]}$, necesitamos determinar cómo incrementar o decrementar de manera que decrezca el error global. Esto puede hacerse usando una regla de descenso por el gradiente como sigue en la siguiente ecuación número (19):

$$\nabla w_{ji}^{[s]} = -lcoef \cdot \left(\partial E / \partial w_{ji}^{[s]}\right) \tag{19}$$

Donde *lcoef* es un coeficiente de aprendizaje. En otras palabras, cambiar cada peso de acuerdo al tamaño y dirección del gradiente negativo en la superficie de error.

Las derivadas parciales pueden calcularse directamente del valor del error local discutido en la anterior subsección, porque, por la regla de la cadena como se ve en la siguiente ecuación número (20):

$$\begin{aligned} \partial E / \partial w_{ji}^{[s]} &= \left(\partial E / \partial I_j^{[s]}\right) \cdot \left(\partial I_j^{[s]} / \partial w_{ji}^{[s]}\right) \\ &= -e_j^{[s]} \cdot x_i^{[s-1]} \end{aligned} \tag{20}$$

Combinando tenemos los valores de la siguiente ecuación número (21):

$$\nabla w_{ji}^{[s]} = lcoef \cdot e_j^{[s]} \cdot x_i^{[s-1]} \tag{21}$$

La función de error global

La discusión previa ha asumido la existencia de una función de error global. Esta función es necesaria para definir los errores locales en la capa de salida para que estos puedan ser retro-propagados hacia el interior de la red.

Supongamos que un vector es presentado en la capa de entrada de la red y supongamos que la salida deseada *d* es especificada por un instructor. Sea *o* quien denote la salida producida por la red con su actual conjunto de pesos. Entonces una medida del error al lograr esa salida deseada está dada por la siguiente ecuación número (22)

$$E = 0.5 \cdot \sum_k \left((d_k - o_k)^2 \right) \tag{22}$$

Donde el subíndice *k* indexa los componentes de *d* y *o*. Aquí, el error local primario está dado por $d_k - o_k$. De la anterior, el "error local escalado" de cada nodo de la capa de salida está dado por la siguiente ecuación dado por la siguiente ecuación número (23):

$$\begin{aligned} e_k^{(o)} &= -\partial E / \partial I_k^{(o)} \\ &= -\partial E / \partial o_k \cdot \partial o_k / \partial I_k \\ &= (d_k - o_k) \cdot f'(I_k) \end{aligned} \tag{23}$$

E, como se definió anteriormente, se define el error global de la red para un vector particular *(i, d)*. Una función en conjunto del error global puede definirse como la suma de todas las funciones de patrones específicos del error.

Cada vez que un particular vector *(i, d)* es mostrado, el algoritmo de back-

propagation modifica los pesos para reducir ese componente en particular de la función de conjunto global de error.

2.2.6 Algoritmos Genéticos

Es un algoritmo matemático altamente paralelo que transforma un conjunto de objetos matemáticos individuales con respecto al tiempo usando operaciones modeladas de acuerdo al principio Darwiniano de reproducción y supervivencia del más apto, y tras haberse presentado de forma natural una serie de operaciones genéticas de entre las que destaca la recombinación sexual (cruzamiento). Cada uno de estos objetos matemáticos suele ser una cadena de caracteres de longitud fija, que ajusta el modelo de las cadenas de cromosomas, y se les asocia con una cierta función matemática que refleja su aptitud.

Terminología

Genética: es la ciencia que estudia los mecanismos de transmisión de características de una especie de una generación para otra. Población: conjunto de individuos los cuales corresponde a un conjunto de soluciones iniciales. Individuos: también es llamado de cromosoma y corresponden a una posible solución, el cual es asociado a un valor que es llamado de aptitud. Función de aptitud: grado de aptitud de un individuo, y mide la capacidad de la solución (individuo) para el determinado problema.

Características de los AG

- Los AG trabajan sobre un conjunto de parámetros codificados es decir

efectúan las operaciones sobre los cromosomas y no sobre el problema.

- Realizan la búsqueda bajo un conjunto de puntos (paralelismo implícito).
- Usan la información de la función de aptitud, no requiere otros conocimientos.
- Las reglas de transición son probabilísticas no determinísticas.

Representación de soluciones

Cada individuo de una población generada por el AG, representa una posible solución del problema que se desea resolver. Así las variables de la función objetivo deben ser representadas por un individuo. Por ejemplo, considere la siguiente función objetivo $f(x)=x2$. Todos los individuos de cualquier población deben representar la variable x. Una posible representación es una simple binarización de los números, otra solución podría ser usar los mismos valores enteros. Escoger una adecuada representación es un área de interés para los AG.

Representación binaria: Una de las representaciones más utilizadas en los AG, es la representación binaria ideal para problemas donde las variables son discretas. Por lo tanto, si las variables son continuas una conversión es necesaria. Por ejemplo, si tenemos valores en el intervalo [0,1] que debe ser representado como cadenas binarias de tamaño 3

Representación continua: Para casos donde el error ocasionado por la representación binaria es crítico los AG, pueden usar directamente los valores continuos. En este caso ninguna conversión es necesaria. No en tanto se debe implementar operadores genéticos de cruzamiento y mutación adecuados a la representación.

Selección para la reproducción

El objetivo principal del operador de selección es copiar las mejores soluciones

eliminando las soluciones de baja aptitud, mientras el tamaño de la solución es constante. Esto es realizado siguiendo los siguientes pasos.

1. Identificar las mejores soluciones de la población.

2. Realizar múltiples copias de las mejores soluciones.

3. Eliminar soluciones de baja aptitud lo que permite que varias copias de las mejores soluciones puedan ser insertadas en la población.

Selección proporcional

La selección proporcional, está dada la función de aptitud entre el total de la suma total de toda la población siendo *Fi* la aptitud de la solución, *N* es el tamaño de la población. Se llama proporcional porque dependiendo del valor de aptitud su probabilidad de selección es proporcional a la misma.

Selección por torneo

En la selección por el torneo son realizadas varias competiciones entre dos soluciones y la mejor solución es copiada en la lista de soluciones. Este proceso es repetido hasta llenar la lista. Fue demostrado que este método posee una convergencia igual o mejor que las otras estrategias de selección además de poseer una complejidad menor que las otras. Para obtener el número de copias esperado, es necesario obtener la probabilidad de cada solución *Pi*. El número de copias en la lista de soluciones es calculado por $Ci = Pi\ N$. O sea, las soluciones con mejor valor de aptitud tendrán más copias en la lista de soluciones.

Selección por ranking.

Esta solución ordena la población por el valor de su aptitud, desde la peor solución hasta la mejor *N* siendo el número de copias proporcional a su ranking.

Cruzamiento

Este operador genera nuevas soluciones a partir de las soluciones escogidas de la lista de soluciones. El operador de cruzamiento posee diferentes variaciones, muchas de ellas especifican a un determinado problema. La forma más simple de cruzamiento es conocida como cruzamiento de un punto, que consiste en:

1. Escoger arbitrariamente dos individuos de la lista de soluciones.

2. Escoger dentro de la cadena del individuo una posición *k* llamada posición de cruzamiento.

3. Crear nuevos descendientes cambiando las cadenas parciales de cada uno de los individuos.

Mutación

La mutación es un operador que produce una alteración aleatoria en una posición de un pequeño número de individuos. La mutación es la segunda manera de los AGs explorar el espacio de búsqueda. Esta pequeña alteración impide que el algoritmo genético tenga convergencia muy rápida, evitando su estabilización en regiones de mínimos locales.

Elitismo

El operador de elitismo mantiene las mejores soluciones encontradas previamente en las generaciones siguientes.

Pasos para implementar un AG

Para resolver un determinado problema utilizando AG los siguientes pasos deben ser considerados:

1. Definir una representación a ser usada para cada individuo de manera que una solución completa puede ser representada.

2. Definir las estrategias de substitución, selección, cruzamiento y mutación

3. Definir la función de aptitud.

4. Ajustar los siguientes parámetros:

- Tamaño de la población.
- Probabilidad de cruzamiento.
- Probabilidad de mutación.
- Numero de generaciones.

Los algoritmos genéticos permiten optimizar parámetros de funciones en ese sentido son optimizadores por excelencia y también pueden ser usados para seleccionar las características más adecuadas en un problema de clasificación.

CAPITULO 3: MARCO METODOLÓGICO

La metodología a usar es la de KDD (Gilbert, 2006) que comprende las siguientes etapas como se puede ver en la siguiente figura 3.1:

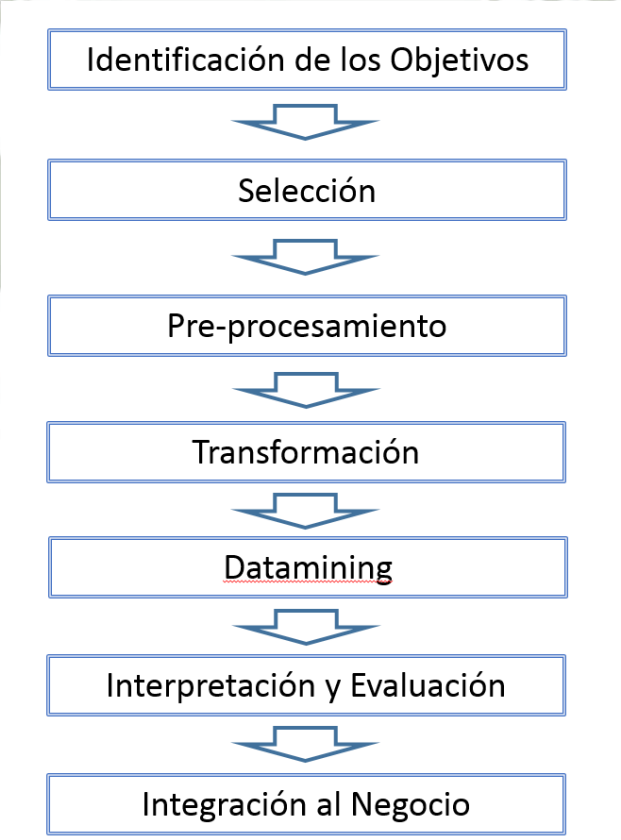

Figura 3.1: Etapas del proceso metodológico desarrollado en la presente tesis

- Identificación de los Objetivos.

  Para la presente tesis, los objetivos ya fueron definidos el cual se basa en la determinación de la mortalidad que un paciente puede tener al ingresar al hospital Honorio Delgado, cabe resaltar que la data utilizada es información que se registra diariamente de forma cotidiana y no tiene mucho detalle de las particularidades del paciente, muy por el contrario, son muy generales, lo que complica que el resultado sea acertado en un 100%.

- Selección de Datos e Información.

  Esta selección de datos en un inicio se basó en función de la revisión bibliográfica, que fue útil para poder tener en cuenta que se usó en otros estudios, dicha revisión fue útil para poder empezar la investigación, pero después se tuvo que adaptar a nuestra realidad, debido a que la información obtenida del hospital no necesariamente tenía el formato para un uso directo para el aprendizaje automático para ser usada en la presente tesis.

- Pre procesamiento.

  El pre procesamiento de los mismos es necesario debido a que la información no es útil en su totalidad haciendo un filtrado de diferentes atributos en esta etapa se utilizó la herramienta de Excel que permitió eliminar data que no es relevante para nuestro propósito, además nos permitió analizar la importancia de los atributos para una posterior utilización de los mismos.

- Transformación.

  La data usada tuvo que ser transformada en muchos casos para poder aprovechar el potencial de cada atributo, además por la rigidez del sistema de información del hospital muchos datos son nominales mas no numéricos como lo ameritan el aprendizaje automático, por eso fue importante usar transformar la data como por ejemplo el caso del atributo sexo no aparece directamente como tal en el archivo proporcionado por el hospital se tuvo que interpretar la información y transformar los datos crear el atributo de masculino y femenino así como es caso fue necesario realizar otros más que son detallados en la experimentación.

- Algoritmos de *Data Mining*.

  En cuanto a los algoritmos de minería se usaron básicamente 4, el primero los algoritmos genéticos para la selección de atributos, y en la clasificación del tipo supervisado para realizar la clasificación se usaron las redes neuronales de tipo perceptron multicapa, el clasificador bayesiano y las máquinas de vectores de soporte, dichos algoritmos de minera son muy usados en problemas de clasificación y son de los que mejores resultados en muchas aplicaciones en este sentido fue también la revisión bibliográfica nos permitió elegir dichas técnicas por su robustez y óptimos resultados en las distintas aplicaciones existentes.

- Interpretación y Evaluación.

  Para la estimación del modelo predictivo, se aplicó el método de validación cruzada

  Utilizando el programa Matlab. Para cada combinación G × A se seleccionaron 2 repeticiones al azar, que permitieron modelar y ajustar los datos a los diferentes

modelos AMMI. Las restantes repeticiones fueron reservadas, como un conjunto de observaciones, hasta el momento de utilizarlas como datos de confirmación o validación.

- Integración al Negocio.

  Finalmente, esta etapa se dará a futuro porque depende de las políticas de la institución.

3.1. Alcances y Limitaciones.

3.1.1. Alcances.

El presente estudio, explorara diversos algoritmos de minería de datos contenidos dentro de la herramienta Matlab, estos previamente identificados, con el fin de evaluarlos en base a diferentes características, como el tiempo de ejecución o calidad de la información obtenida, con el fin de determinar la eficiencia presentada por cada uno respecto a información obtenida.

Es importante, resaltar que la data proporcionada por el Hospital es información confidencial motivo por el cual se utilizarán solo medias estadísticas genéricas usadas en la prueba del algoritmo de clasificación, mas no se darán detalles ni casos particulares por ser información particular del paciente.

3.1.2. Limitaciones.

Se mantendrá en el anonimato a los pacientes, además por ser información genérica se filtrará la información para solamente los casos que ocurran en dicho hospital el cual es un caso particular.

3.2. Aporte.

Una vez culminado el análisis o estudio, se obtendrá un completo análisis de los algoritmos identificados como eficientes, abarcando diversas características tales como calidad de la información y conocimiento obtenido a partir de los datos almacenados o recogidos en el H.R.H.D. así mismo se brindará una guía que facilitará la toma de decisión por un algoritmo a ser implementado.

3.3. Nivel de investigación.

Con este estudio, se pretende realizar una evaluación de diferentes algoritmos en base a su desenvolvimiento una vez aplicados a un almacén de datos en particular, posteriormente se realizará una comparación en base a los resultados obtenidos y se emitirá un cuadro como resultado final.

3.4. Población y muestra o universo.

Para la presente tesis, se tomó en consideración las muestras tomadas del Hospital Regional Honorio Delgado, el cual cuenta con entradas diarias de atención las cuales son registradas en el sistema de información donde en el presente estudio se toma como el número de registros del último año completo que fue el 2016 en el que se cuenta con 24 500 registros de ingresos de pacientes, siendo cada uno de los registros formado por un total de 26 atributos, consideramos que por ser el Hospital más importante de la ciudad es que la data usada en este experimento es suficiente para validar los resultados obtenidos, en la siguiente sección se dará detalles del procesamiento de dicha data.

3.5. Propuesta.

Este trabajo realizo el estudio experimental de 3 algoritmos de clasificación de datos, siendo estos: Clasificador Bayesiano Ingenuo, Maquinas de Vectores de Soporte y Redes Neuronales *Backpropagation*, con el objetivo de clasificar los índices de mortalidad de un paciente del Hospital Honorio Delgado, como se mencionó anteriormente la información usada es limitada a los atributos que el sistema de información registra, básicamente en la ejecución de los algoritmos analizados se presentan dos etapas como se puede apreciar en el siguiente gráfico:

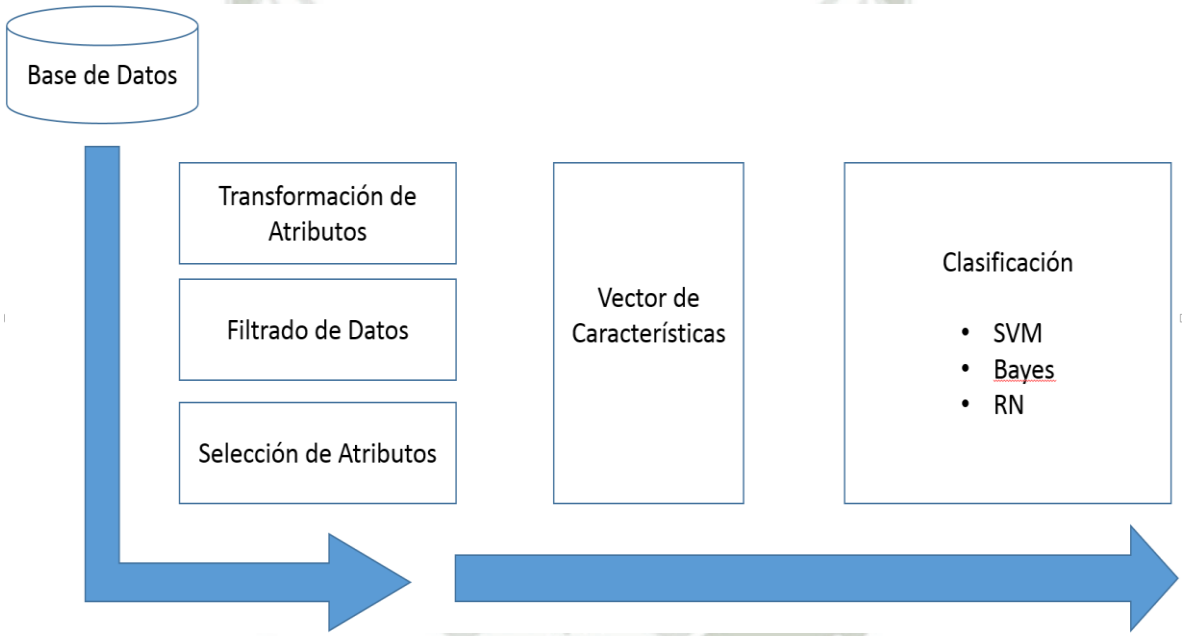

Figura 3.2: Propuesta para el análisis de predicción de mortalidad propuesto (fuente propia)

Podemos ver un esquema del proceso que se realizó en el desarrollo de esta tesis donde se distinguen de forma resumida las etapas realizadas en el análisis de los datos que se pasaran a detallar a continuación.

3.6. Base de datos.

La base de datos utilizada en la presente tesis fueron los registros del hospital Honorio

Delgado, el cual es el consolidado del año 2016, el cual cuenta con 24 500 registros de datos, que corresponde a los siguientes atributos:

- Fecha: corresponde a la fecha del registro de un paciente
- Id: corresponde al código del paciente
- Historia Clínica: es un valor numérico de la historia clínica del paciente
- Fecha de Egreso: es la fecha que el paciente le dan de alta
- Días de Estancia: son la cantidad de días que el paciente estuvo en el hospital
- Edad: es la edad del paciente
- Año – Meses: valor en palabras que puede ser años, meses y días
- Edad Masculino: edad del paciente cuando este es hombre
- Edad Femenino: edad del paciente cuando este es mujer
- Vivo o Fallecido: 1 si el paciente está vivo y 2 si falleció
- Diagnostico 1: codificación de enfermedad 1
- Diagnostico 2: codificación de enfermedad 2
- Diagnostico 3: codificación de enfermedad 3
- Diagnostico 4: codificación de enfermedad 4
- Diagnostico 5: codificación de enfermedad 5
- Diagnostico 6: codificación de enfermedad 6
- Cirugía Laparoscópica: valor numérico de la cirugía realizada
- Operación 1: si se realiza una operación 1
- Operación 2: si se realiza una operación 2
- Operación 3: si se realiza una operación 3
- Operación 4: si se realiza una operación 4
- Código Anestesia: valor numérico de la anestesia
- Apellidos del Cirujano que Opero

- Apellidos Obstetra
- Apellidos Obstetriz
- Servicios: describe el servicio que fue usado

Siendo un total de 26 atributos como se puede ver mucha de la información registrada no es de utilidad para nuestro propósito por eso es necesario transformar los datos, así como filtrar algunos.

Un caso particular es el atributo diagnostico el cual sigue una codificación según la siguiente tabla, en dicha tabla se puede apreciar los códigos que son usados en el sistema de información del H.R.H.D., que sigue los estándares internacionales de Clasificación Internacional de Enfermedades que permite a los pacientes registrar según dicha codificación los diagnósticos y enfermedades que estos puedan tener al momento de ingresar al hospital.

Tabla: Código CIE los diagnósticos.

ap.	Códigos	Título
I	A00-B99	Ciertas enfermedades infecciosas y parasitarias
II	C00-D48	Neoplasias
III	D50-D89	Enfermedades de la sangre y de los órganos hematopoyéticos y otros trastornos que afectan el mecanismo de la inmunidad
IV	E00-E90	Enfermedades endocrinas, nutricionales y metabólicas
V	F00-F99	Trastornos mentales y del comportamiento
VI	G00-G99	Enfermedades del sistema nervioso
VII	H00-H59	Enfermedades del ojo y sus anexos
VIII	H60-H95	Enfermedades del oído y de la apófisis mastoides
IX	I00-I99	Enfermedades del sistema circulatorio
X	J00-J99	Enfermedades del sistema respiratorio
XI	K00-K93	Enfermedades del aparato digestivo
XII	L00-L99	Enfermedades de la piel y el tejido subcutáneo
XIII	M00-M99	Enfermedades del sistema osteomuscular y del tejido conectivo
XIV	N00-N99	Enfermedades del aparato genitourinario
XV	O00-O99	Embarazo, parto y puerperio
XVI	P00-P96	Ciertas afecciones originadas en el periodo perinatal
XVII	Q00-Q99	Malformaciones congénitas, deformidades y anomalías cromosómicas
XVIII	R00-R99	Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
XIX	S00-T98	Traumatismos, envenenamientos y algunas otras consecuencias de causa externa
XX	V01-Y98	Causas externas de morbilidad y de mortalidad
XXI	Z00-Z99	Factores que influyen en el estado de salud y contacto con los servicios de salud
XXII	U00-U99	Códigos para situaciones especiales

Como ejemplo, un valor de dicho diagnóstico es el siguiente O47el cual indica que el diagnostico está en la codificación O, dentro de la cual está en la

subcategoría 47, en la presente tesis no se tomó detalle de los diagnósticos, solo se consideró la cantidad de diagnóstico que tiene un paciente ya que eso indica las condiciones en las que se encuentra el paciente y en consecuencia el riesgo que corre por tener más una enfermedad.

De forma similar se trató el número de operaciones, ya que no todos se operan y los que si lo hacen también tienen una cantidad de operaciones realizadas eso también indica el riesgo de muerte, que corre un paciente cada vez que se realiza más de una operación.

También hay que notar, que no todos los atributos están llenos existe en su mayor parte celdas vacías indicando que el sistema de información que registra dicha data no controla índices ni relacionamiento, eso dificulta el uso de la información para los propósitos de la presente tesis, por otro lado en alguno atributos existe información que no tiene relacionamiento a lo que debería mostrar como por ejemplo, en días de estancia que aparece una fecha e incluso códigos alfa numéricos, no existe el atributo sexo pero si están dos atributos que puede servir para inferir dicho atributo como lo es la edad masculino y la edad femenino. Finalmente podemos decir que los atributos de índices, así como los nombres son descartados por defecto por no proveer ningún tipo de información útil para determinar el riesgo de muerte de un paciente.

3.7. Transformación y eliminación de atributos.

No todos los atributos proveen información, para los fines de la investigación en ese sentido se eliminaron los siguientes atributos:

- Fecha: corresponde a la fecha del registro de un paciente
- Id: corresponde al código del paciente
- Historia Clínica: es un valor numérico de la historia clínica del paciente
- Fecha de Egreso: es la fecha que el paciente le dan de alta
- Cirugía Laparoscópica: valor numérico de la cirugía realizada
- Código Anestesia: valor numérico de la anestesia
- Apellidos del Cirujano que Opero
- Apellidos Obstetra
- Apellidos Obstetriz

Algunos atributos tuvieron que ser transformados, para luego poder ser utilizados, se tuvieron que transformar los atributos:

- Edad Masculino
- Edad Femenino

Para obtener el atributo Sexo, como uno solo y no separado como se presenta originalmente ya que esto ocasionaba problemas debido a que si era hombre no existía registro en el atributo Edad Femenino y viceversa, una vez realizada dicha transformación los atributos originales de Edad femenino y Edad masculino fueron eliminados.

De los atributos:

- Diagnostico 1: codificación de enfermedad 1
- Diagnostico 2: codificación de enfermedad 2
- Diagnostico 3: codificación de enfermedad 3
- Diagnostico 4: codificación de enfermedad 4

- Diagnostico 5: codificación de enfermedad 5
- Diagnostico 6: codificación de enfermedad 6

Fueron fusionados en uno solo el cual fue simplemente el conteo de cuantos diagnósticos tiene un paciente creándose un nuevo atributo llamado:

- Número de Diagnósticos; el cual no es más que el simple conteo de cuantos diagnósticos tienen una persona y una vez calculado este atributo los originales fueron desechados.

De forma similar sucedió en el número de operaciones:

- Operación 1: si se realiza una operación 1
- Operación 2: si se realiza una operación 2
- Operación 3: si se realiza una operación 3
- Operación 4: si se realiza una operación 4

Creando se un nuevo atributo llamado:

- Número de operaciones que simplemente el conteo de la cantidad de operaciones que tiene registrado un paciente el cual una vez calculado hizo que los anteriores fueran descartados para nuestro propósito.

Es necesario notar que dichas eliminaciones y transformaciones fueron necesarias porque las técnicas de aprendizaje automático, que utilizaremos se basan en atributos numéricos y los eliminados no lo eran y los transformados ahora sí lo son.

3.8. Filtrado de datos.

Después de eliminar atributos y transformar otros, quedarían:

- Días de Estancia
- Edad
- Sexo
- Vivo o Fallecido

- Numero de Diagnósticos
- Número de Operaciones
- Servicios

A pesar de tener pocos atributos, fue necesario filtrar registros de los atributos a utilizar primero porque muchos de ellos no registran información coherente, entonces en primera instancia fueron eliminadas los registros que tengan valores anómalos a lo que les corresponde por ejemplo, en días de estancia se registraban algunas fechas, también se eliminaron los registros que tenían algún atributo vacío, ya que al no contener la información completa dificultarían el diagnóstico. A continuación, mencionaremos que filtrado fue realizado en cada uno de los atributos.

En cuanto al atributo días de estancia, no se realizó ningún tipo de filtrado y fue utilizado tal con la información que registraba originalmente, salvo los errores que fue mencionado anteriormente que nos imaginamos fue motivo de algún error involuntario.

En el atributo edad, no se tomaron en consideración a los recién nacidos y todos aquellos que no cumplieron un año.

El atributo Vivo = 1 o Fallecido = 2, además de esos valores se encontraron registros con valores 3 e incluso 4, dichos registros fueron eliminados quedándose la base con solo los vivos y fallecidos.

El atributo de numero de diagnósticos fue usado tal cual no fue necesario realizar ningún tipo de filtrado

El atributo número de operaciones también no tuvo problemas y fue utilizado sin ningún filtrado

El atributo de servicio, ayudo a filtrar 13 tipos de servicios los cuales se presentan tanto para personas vivas como para fallecidos, las demás fueron descartadas porque no abrían ocurrencias de la clase vivos y la clase fallecidos para el diagnóstico. Es importante mencionar que este atributo solo sirvió para filtrar mas no fue usado con sus valores dados para la clasificación, ya que por cada servicio existen ambos casos de personas vivas y fallecidas y eso dificultaría la clasificación final.

Finalmente se usaron cinco atributos, para la discriminación y un atributo de clasificación, hay que n otra que también se tuvo que eliminar registros repetidos lo cual no aporta a la mejora de las técnicas de clasificación, por el contrario, demorarían los procesos de aprendizaje automático y prueba.

La base de datos del año 2016 consta de 24 500, después del filtrado los datos que tienen error y los casos mencionados anteriormente, 6 229 registros son los que utilizaremos para nuestro experimento.

3.9. Selección de atributos.

Para la selección de atributos se usaron los algoritmos genéticos, el cual usa una representación binaria donde cada valor de la cadena que tenga valor 1 significa que dicha característica será usada como atributo para la clasificación por ejemplo tenemos las siguientes representaciones de unos individuos:

I 1: 0 1 1 0 1

I 2: 1 1 0 1 0

El I1 representa un individuo donde los valores donde están los unos indican que fueron usados los atributos 2, 3, y 5, por otro lado el individuo dos indica que fueron usados los atributos 1, 2 y 4, con dichos atributos se usa un clasificador para poder decidir cuál combinación de atributos es la más adecuada creando una población inicial y luego proceder a realizar una selección por torneo para poder luego proceder al cruzamiento, mutación y finalmente el elitismo así garantizar la convergencia del algoritmo genético.

3.10. Vector de características.

Finalmente, nuestro vector de características estaría conformado de cinco atributos a ser usados para la clasificación los cuales son:

- Días de Estancia
- Edad
- Sexo
- Numero de Diagnósticos
- Número de Operaciones

Y la clase a la que pertenece que sería:

- Vivo o Fallecido

3.11. Clasificación.

3.11.1. Métodos de validación de aprendizaje automático.

La validación cruzada de K iteraciones o del inglés *K-fold cross-validation*, los datos de muestra se dividen en K partes. Y uno de las partes se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El

proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado.

Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba, pero aun así tiene una desventaja, y es que, a diferencia del método de retención, es lento desde el punto de vista computacional. En la práctica, la elección del número de iteraciones depende de la medida del conjunto de datos. Lo más común es utilizar la validación cruzada de 10 iteraciones.

El objetivo de la validación cruzada, es estimar el nivel de ajuste de un modelo a un cierto conjunto de datos de prueba independientes de las utilizadas para entrenar el modelo. Estas medidas obtenidas pueden ser utilizadas para estimar cualquier medida cuantitativa de ajuste apropiada para los datos y el modelo. Por ejemplo, en un modelo basado en clasificación binaria, cada muestra se prevé como correcta o incorrecta (si pertenece a la temática o no), de forma que en este caso, se puede usar la tasa de error de clasificación para resumir el ajuste del modelo. Así mismo, se podrían utilizar otras medidas como el valor predictivo positivo. Cuando el valor a predecir se distribuido de forma continua se puede calcular el error utilizando medidas como: el error cuadrático medio, la desviación de la media cuadrada o la desviación absoluta media.

3.11.2. Precisión y exhaustividad.

La precisión y exhaustividad (denominado a veces como exhaustividad y precisión) es una métrica empleada en la medida del rendimiento de los sistemas de búsqueda y recuperación de información y reconocimiento de patrones. En este contexto se denomina precisión (denominado igualmente valor positivo predicho) como a la fracción de instancias recuperadas que son relevantes, mientras exhaustividad (denominado igualmente sensibilidad o exhaustividad) es la fracción de instancias relevantes que han sido recuperadas. Tanto la precisión como la exhaustividad son entendidas como medidas de la relevancia.

3.11.3. Sensibilidad y especificidad.

Dado un estimador para una variable estadística discreta binaria se definen dos valores asociados importantes:

La sensibilidad nos indica la capacidad de nuestro estimador para dar como casos positivos los casos realmente enfermos; proporción de enfermos correctamente identificados. Es decir, la sensibilidad caracteriza la capacidad de la prueba para detectar la enfermedad en sujetos enfermos (24).

La especificidad nos indica la capacidad de nuestro estimador para dar como casos negativos los casos realmente sanos; proporción de sanos correctamente identificados. Es decir, la especificidad caracteriza la capacidad de la prueba para detectar la ausencia de la enfermedad en sujetos sanos (25).

$$\boldsymbol{Sensibilidad} = \frac{\boldsymbol{Verdaderos\ Positivos}}{\boldsymbol{Total\ Casos\ Positivos}} \quad (24)$$

$$\boldsymbol{Especificidad} = \frac{\boldsymbol{Verdaderos\ Negativos}}{\boldsymbol{Total\ Casos\ Negativos}} \quad (25)$$

3.11.4.Matriz de confusión

En el campo de la inteligencia artificial una matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en el aprendizaje automático. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

CAPÍTULO 4: ANÁLISIS Y DISCUSIÓN

4.1. Análisis de los Atributos Filtrados.

Después de haber realizado los respectivos filtros y transformaciones a los atributos a continuación mostraremos en tres dimensiones la distribución de los registros, que ayudara a ver cómo están distribuidos los datos en grupos de 3 en 3 atributos, no fue considera para la agrupación el atributo de hombre o mujer pues el interés es que sea diagnostico en general de cualquier paciente sin considerar el sexo

- **Atributos Días de Hospitalización, Edad, Numero de Diagnósticos:** La primera combinación de datos será el uso de las variables de días de hospitalización, con la edad y el número de diagnósticos los de color verde corresponde a las personas que fallecieron mientras que los de color negro son las personas vivas, que se hicieron algún registro en el sistema, podemos ver que existe no hay mucha diferenciación de que tienen más de un diagnostico con relación a los otros, como se puede aprecias en la siguiente figura 4.1.

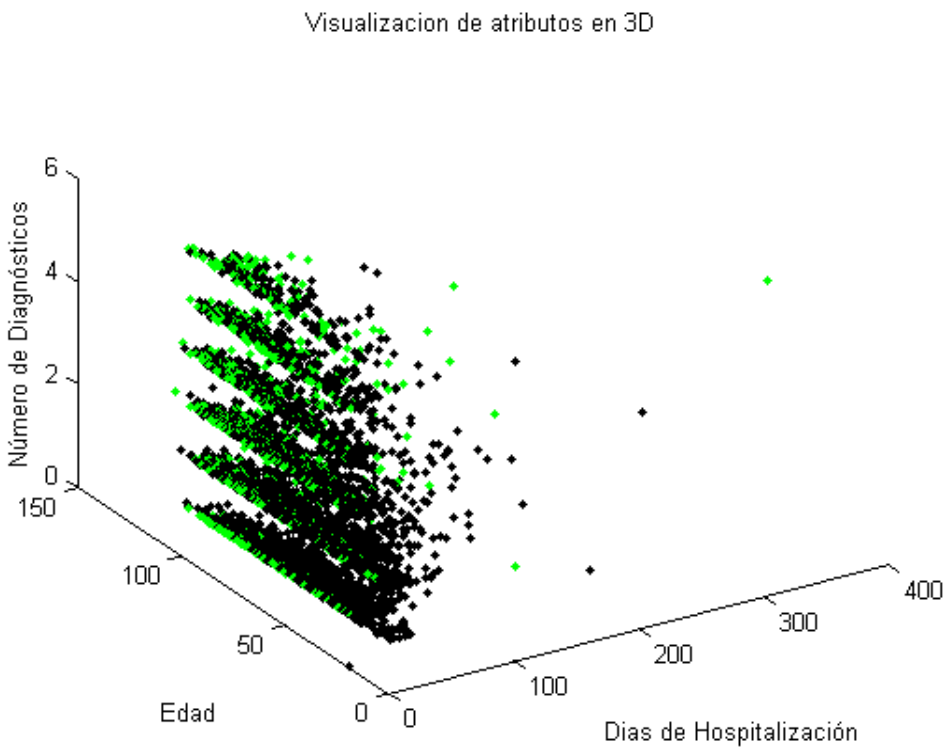

Figura 4.1: Visualización de los registros en relación a los siguientes atributos Días de hospitalización, edad y numero de diagnósticos

- **Atributos Días de Hospitalización, Edad, Numero de Operaciones:** Igual que en el caso anterior los de color verde son los pacientes fallecidos y los de color negro los vivos, en este segundo grafico podemos apreciar que hay una diferenciación entre los que tiene más de cuatros operaciones que es un buen indicio como un buen atributo que ayudara a la discriminación de predicción de mortalidad de pacientes como se puede ver en la siguiente figura 4.2.

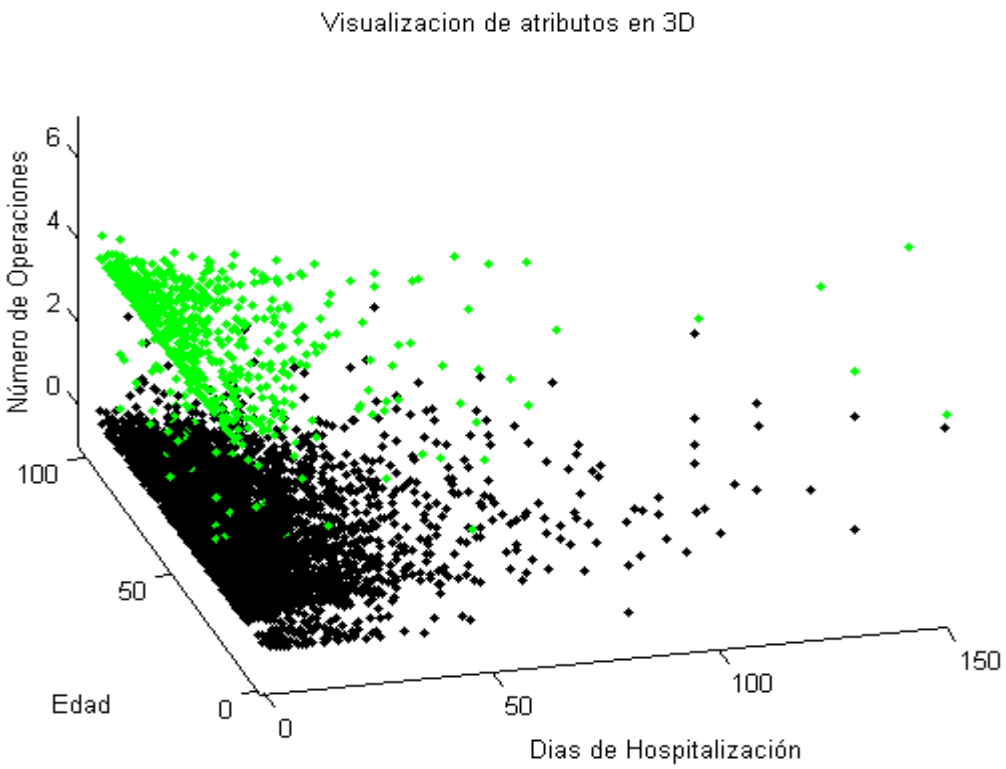

Figura 4.2: Visualización de los registros en relación a los siguientes atributos: Edad, número de operaciones y días de hospitalización

- **Atributos Días de Hospitalización, Numero de Diagnósticos y Numero de Operaciones:** También los de color verde son los pacientes fallecidos y los de negro los vivos, podemos apreciar que el atributo de operaciones consigue diferenciar los vivos de los fallecidos pero también no es difícil notar que aún existe una separación clara de las clases por eso es importante usar todos los atributos para la separación de los mismos en se sentido es importante usar aun todos los atributos para la clasificación, como se puede apreciar en la siguiente figura 4.3.

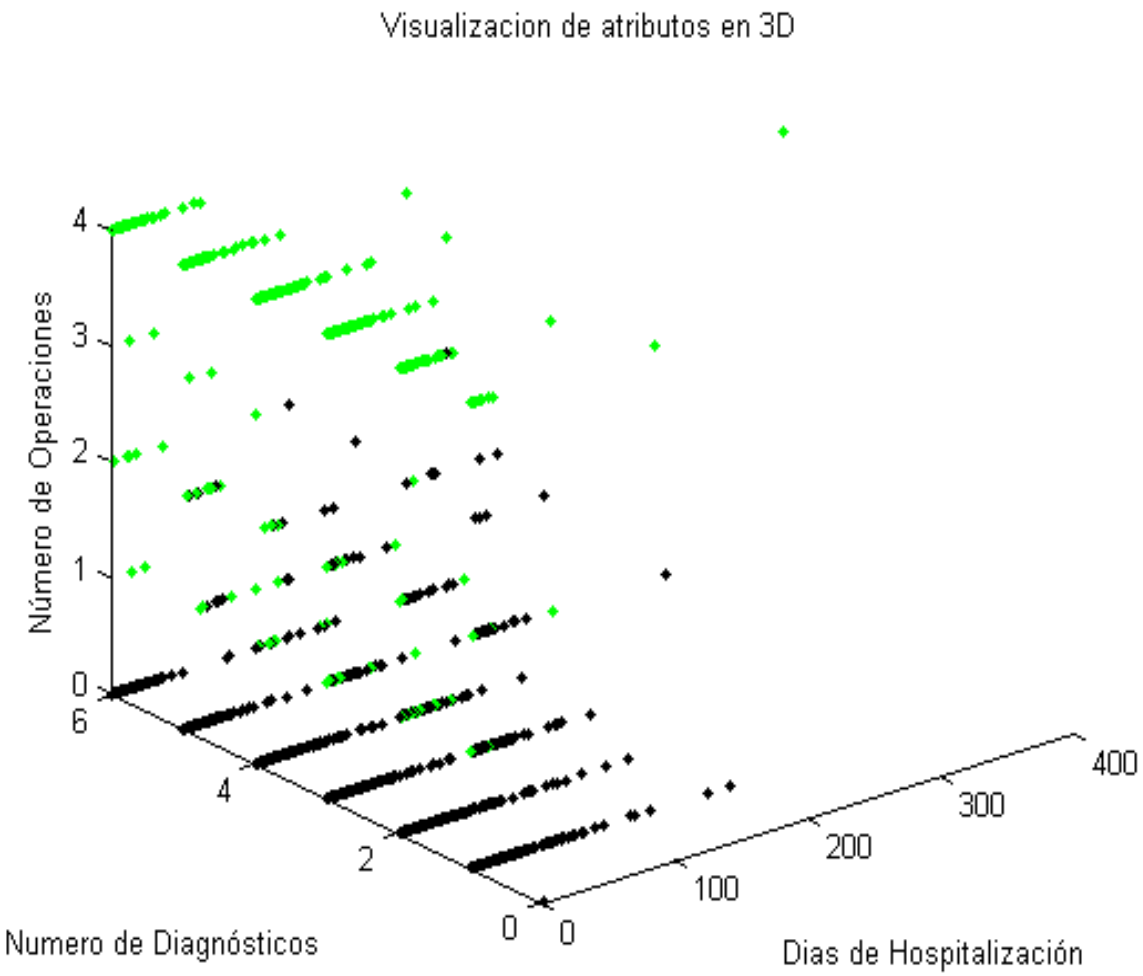

Figura 4.3: Visualización de los registros en relación a los siguientes atributos: Número de Operaciones, Número de diagnósticos y días de hospitalización

- **Edad, Numero de Operaciones y Numero de Diagnósticos:** También los de color verde son los fallecidos y negro los vivos, podemos ver que los diagnósticos más las operaciones separa mejor las clases, pero, también podemos ver que aún existen elementos que no se consiguen separa correctamente lo cual indica que debemos usar todos los atributos para la clasificación ya que no hay forma de separar de forma trivial

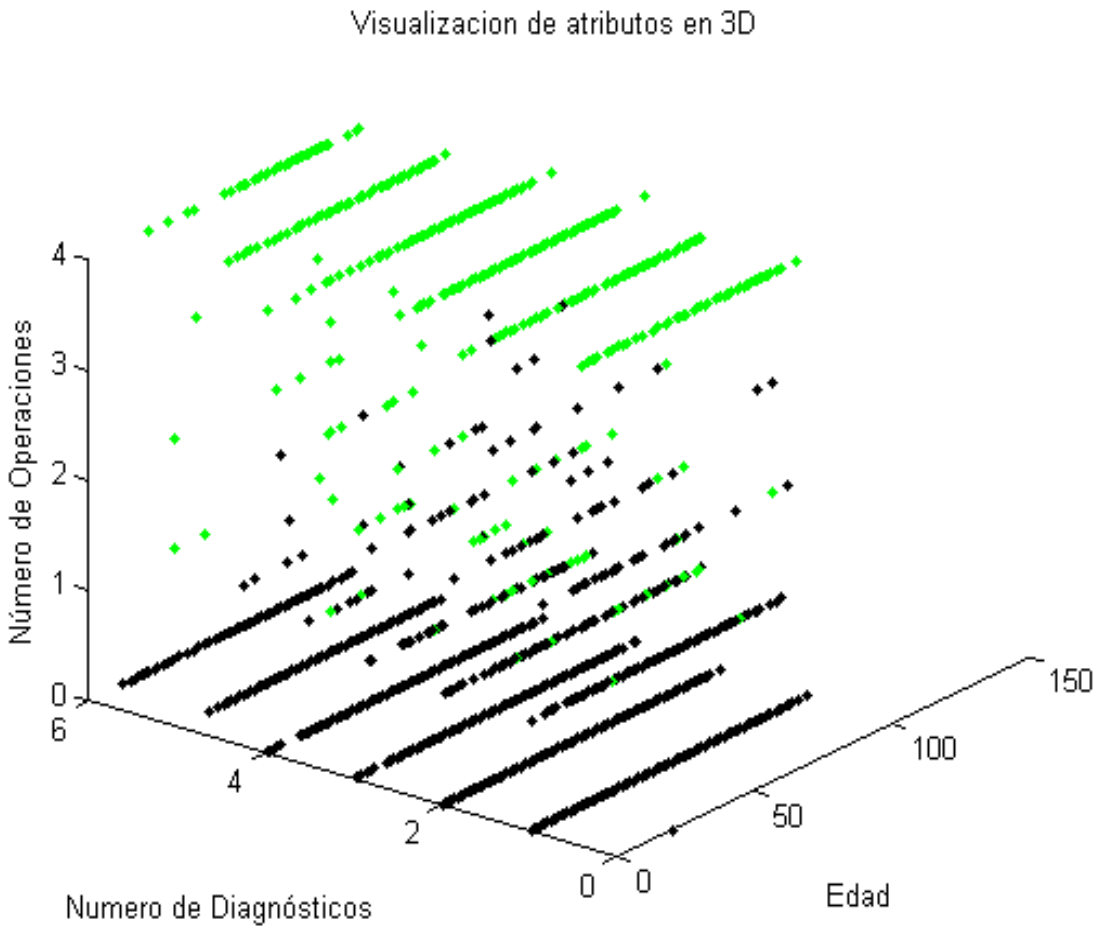

Figura 4.4: Visualización de los registros en relación a los siguientes atributos: Número de Operaciones, Número de diagnósticos y edad

- **Análisis de relevancia de atributos:** Es importante notar que existen atributos relevantes para el pronóstico, para validar dicha importancia se puede hacer al usar un clasificador como máquinas de vectores de soporte, para clasificar los vivos y fallecidos usando solamente un atributo, esto dará una idea de la importancia y relevancia de los atributos y así poder tener una idea de ranking de atributos según la influencia en el diagnostico final, así como se pudo apreciar visualmente en las figuras 4.1, 4.2, 4.3 y 4.4 donde también se ve algo de esa relevancia, además podemos cuantificar numéricamente la importancia de cada uno de los atributos y así tener una idea más clara de cada atributo:

Atributo número de días	40.04 %
Atributo edad	61.78 %
Atributo Sexo	51.82 %
Atributo número de diagnósticos	64.30 %
Atributo número de operaciones	78.00 %

Esto ratifica la importancia de los atributos de número de operaciones y diagnósticos

4.2. Selección de atributos

Los parámetros usados que dieron mejores resultados para la codificación de los algoritmos genéticos fueron los siguientes:

- Tamaño de la población 10 individuos
- Selección por torneo
- Tasa de cruzamiento 0.8
- Probabilidad de mutación es 0.1
- Y un elitismo de 2 individuos
- Numero de generaciones 100

Como se mostrará en las siguientes secciones, el potencial atributo que pudo ser eliminado fue el atributo de sexo, pero es importante notar que el resultado global no es mejor que los resultados obtenidos, pero sí de todos los atributos es el que menos aporte tiene al resultado de clasificación en ese sentido se presenta una comparación de los resultados con el uso de dicho atributo y sin el uso del mismo por cada clasificador.

4.3. Redes Neuronales *Backpropagation*

La red neuronal utilizada, fue una arquitectura *Perceptron* multicapas en el simulador que fue usado para realizar las pruebas fue del Matlab. Es necesario tomar en cuenta que fueron probadas varias arquitecturas, principalmente incrementando neuronas y capas, pero la arquitectura que se presenta a continuación fue la que mejores resultados han dado en las pruebas realizadas.

- Arquitectura de tres capas.
- Dado que la cantidad de atributos son cinco la cantidad de neuronas en la capa entrada también serán las mismas.
- En la capa intermedia se tiene 10 neuronas en la capa intermedia.
- Por ser la salida binaria de fallecido o vivo solo tenemos una neurona en la capa de salida.
- La función de activación usada en la capa intermedia es la función exponencial y en la capa de salida la función identidad.
- La tasa de aprendizaje es 0.001.
- La cantidad de épocas de entrenamiento es de 1000.

Para la validación de la arquitectura se utilizó la validación cruzada, con diez particiones K = 10, este parámetro es sugerido por muchos trabajos de investigación como valor adecuado, lo que es importante notar que este método de validación es universal y útil en validación de modelos de clasificación.

Hay que mencionar que de los 6229 registros de los cuales 704 son los casos de personas que fallecieron en el hospital, quedando 5525 de registros de personas que están vivas, eso quiere decir que estamos con una base desbalanceada lo que es lógico porque no siempre fallece una persona si entra a un hospital.

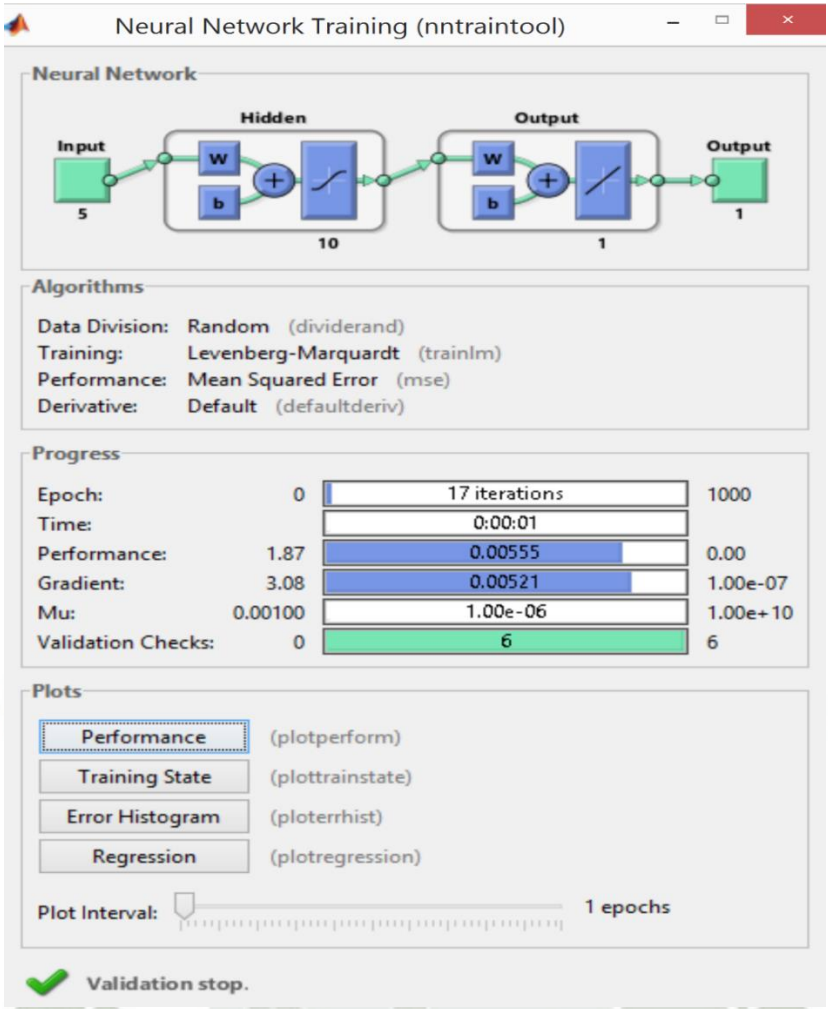

Figura 4.5: Simulador de redes neuronales de Matlab

Como se pudo ver en la figura 4.5 que muestra la arquitectura de la red neuronal del simulador de Matlab, se puede mostrar algunos parámetros mencionados como la cantidad de neuronas escondidas que son 10, entre otros parámetros.

Otra grafica que es importante mostrar en la figura 4.6 es la de la performance de la red neuronal, donde podemos ver las curvas de convergencia del entrenamiento una validación y una curva de test mostrando como el error medio cuadrático (del inglés MSE) converge en las iteraciones del aprendizaje, teniendo un 0.0091251 que corresponde al proceso de entrenamiento.

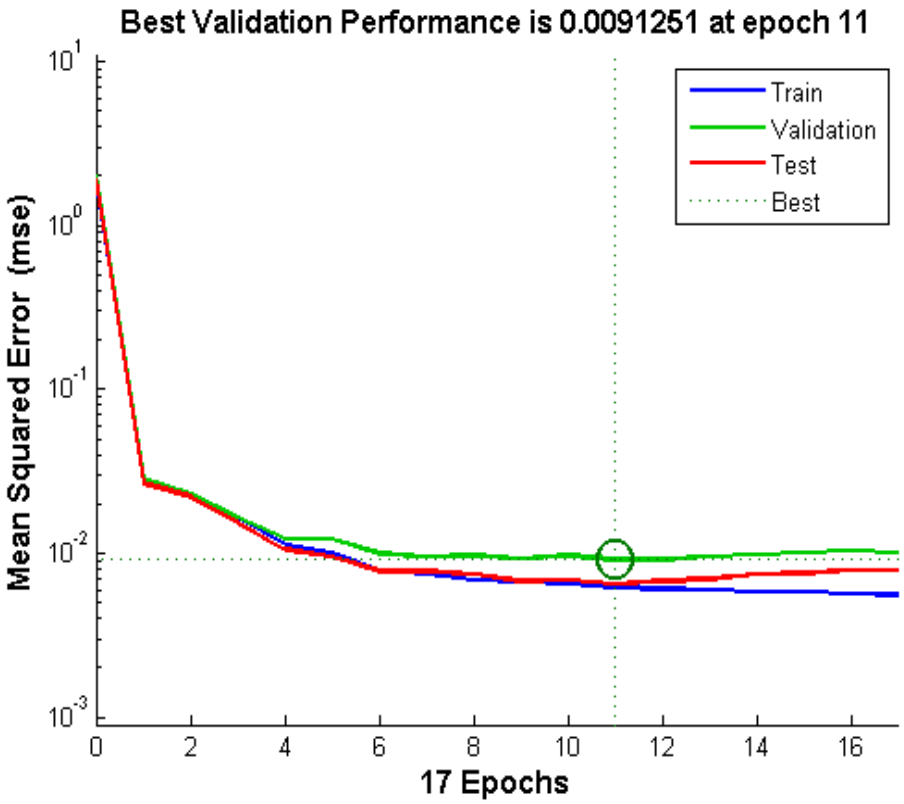

Figura 4.6: Convergencia de la red neuronal en función a la minimización del error

Es importante también analizar, el gradiente descendente en el proceso de entrenamiento, esa evolución nos evidencia que el modelo de red neuronal está realizando un proceso de aprendizaje como se ve en la Figura 4.7.

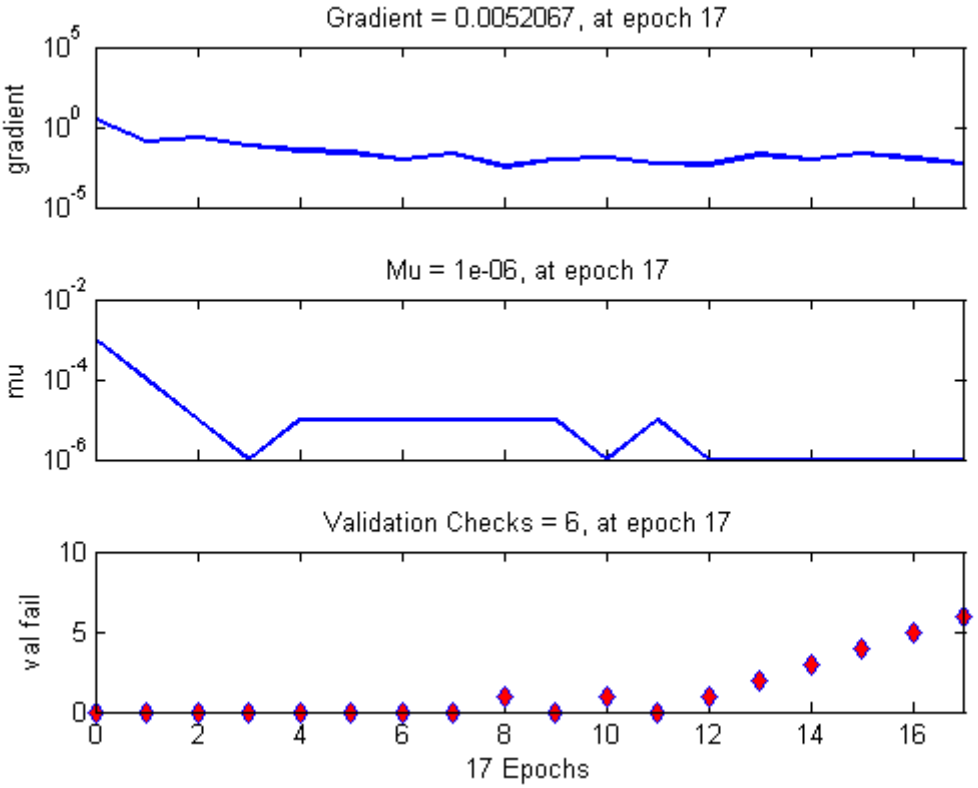

Figura 4.7: Graficas de convergencia de los parámetros de entrenamiento

La matriz de confusión de los resultados obtenidos se puede apreciar en la siguiente tabla:

Tabla 4.1: Matriz de confusión de la Red neuronal

	Fallecido	Vivo
Fallecido	94.74	5.26
Vivo	0.45	99.55

Podemos ver que existe un error de clasificación de 5.26% de falsos positivos que son personas que fallecieron y fueron clasificados con pronóstico de vivos, dado que son menos los casos de fallecidos es por ese motivo que el valor es alto comparado

con el error de clasificación de vivos que no llega ni al uno por ciento siendo de 0.45, pero tenemos que tomar en cuenta en si en el primer caso dichos porcentajes corresponden a 37 y 25 registros equivocados, esto se debe a que las clases son desbalanceadas siendo mayor la cantidad de personas vivas en referencia a las fallecidas.

Por otro lado, otros índices de clasificación de muestran a continuación, con esto podemos notar que el modelo basado en redes neuronales:

- Correctos : 0.990
- Error : 0.010
- Sensibilidad : 0.947
- Especificidad : 0.995

Si eliminamos el atributo de sexo el resultado del modelo será el siguiente:

- Correctos : 0.989
- Error : 0.0106
- Sensibilidad : 0.934
- Especificidad : 0.996

4.4. Clasificador Bayesiano

Para el caso del clasificador Bayesiano ingenuo, también fue utilizado la validación cruzada de K = 10, que es el adecuado para validar las pruebas, en este caso también se utilizó el simulador de Matlab en este existe el algoritmo de clasificación

bayesiana implementada el cual es usa mediante el comando: *NaiveBayes.fit* el cual es usado para la creación de clasificador bayesiano, el cual debe recibir como parámetro el conjunto de entrenamiento que será utilizado el entrenamiento y calibración del modelo, además se debe pasar como parámetro las clases que serán usadas para etiquetar las clases que en este caso son dos para los vivos y para fallecidos.

También es necesario tener en cuenta que se tiene varias distribuciones para la iniciación del clasificador, fueron consideradas en la presente tesis siendo las que mejor resultado obtenidos las siguientes: distribución normal y distribución normal con tolerancia a errores.

Tabla 4.2: Matriz de confusión del Clasificador Bayesiano Normal con Tolerancia a Error

	Fallecido	Vivo
Fallecido	97.60	2.40
Vivo	1.12	98.88

En la tabla 4.2 podemos ver la matriz de confusión donde el error muestra un más balanceado si lo comparamos con el de las redes neuronales, pero por otro lado significa que hay mayor error en el diagnóstico del registro de las personas que estaban vivas, pero en el de las fallecidas se mantiene, aunque el bayesiano tiene un mayor error, pero mínimo.

- Correctos : 0.9801
- Error : 0.0199

- Sensibilidad : 0.9540
- Especificidad : 0.9807

Tabla 4.3: Matriz de confusión del Clasificador Bayesiano Distribución Normal

	Fallecido	Vivo
Fallecido	94.60	5.40
Vivo	2.02	97.98

En la tabla 4.3 podemos ver la matriz de confusión donde el error se muestra más balanceado si lo comparamos con el de las redes neuronales, pero por otro lado significa que hay mayor error en el diagnóstico de los registros de las personas que estaban vivas, pero en el de las fallecidas se mantiene, aunque el bayesiano tiene un mayor error, pero mínimo.

- Correctos : 0.9759
- Error : 0.0241
- Sensibilidad : 0.9460
- Especificidad : 0.9797

Por otro lado, si no usamos el atributo de sexo los resultados serían los siguientes:

- Correctos : 0.9761
- Error : 0.0239
- Sensibilidad : 0.9460
- Especificidad : 0.9799

Podemos apreciar que el resultado en el clasificador bayesiano sin el uso del atributo sexo es mejor en unas milésimas, pero estos resultados en general muestran una menor precisión si los comparamos con las redes neuronales.

4.5. Máquinas de vectores de soporte.

En el caso de las máquinas de vectores de soporte es necesario tener en cuenta que un punto determinante es el uso del kernel para poder determinar la mejor clasificación en el simulador de matlab tenemos la función **svmtrain**, que similar al clasificador bayesiano necesita del conjunto de entrenamiento y las clases respectivas, pero también permite elegir los kernels a utilizar en el caso de matlab se dispone los siguientes kernels:

- Lineal
- Cuadrático
- Polinomial
- RBF
- MLP

Todos fueron probados el RBF y el lineal, fueron los que mejores resultados dieron en la clasificación los cuales se presentaran a continuación. En el kernel RBF se utilizó una sigma igual 1.2, con un con un margen de error C igual a 0.25 con estos valores se obtuvo la siguiente matriz de confusión:

Tabla 4.3: Matriz de confusión del SVM - RBF

	Fallecido	Vivo
Fallecido	96.87	3.13
Vivo	1.41	98.59

Como se puede apreciar en la tabla el error de diagnóstico de fallecidos ha disminuido considerablemente comparado con los demás métodos, este es un punto importante porque si se tienen una persona con riesgo de muerte y se le estima como sano ese error es más grave que de un sano pronosticarlo como posible fallecimiento, porque el que si necesita ayuda no la tendría en ese sentido consideramos que SVM sería la mejor solución para el problema de análisis predictivo de muerte y sobre vida al tener el menor error en el diagnóstico de fallecimiento, como se mencionó el otro Kernel que dio buenos resultados fue el lineal. A continuación podemos ver la tabla 4.4

Tabla 4.4: Matriz de confusión del SVM - Lineal

	Fallecido	Vivo
Fallecido	97.02	2.98
Vivo	1.81	98.19

Este Kernel presenta un menor error incluso que el kernel RBF y la parte importante es que la predicción de fallecidos es más acertada como ya habíamos comentado este sería determinante en la toma de decisiones en ese sentido podemos decir que este modelo de SVM – lineal sería el mejor de todas las técnicas utilizas, el cual tiene:

- Correcto : 0.9806
- Error : 0.0194
- Sensibilidad : 0.9819
- Especificidad : 0.9702

Además, es importante mencionar que este modelo es menos costoso computacionalmente dando una respuesta menor en tiempo comparado con el kernel RBF. Si ahora no utilizamos el atributo de sexo los resultados también mejoran como podemos ver a continuación

- Correcto : 0.9859
- Error : 0.0141
- Sensibilidad : 0.9403
- Especificidad : 0.9917

Finalmente podemos decir que para todos los casos de la validación cruzada matlab dispone una función que facilito realizar dichas pruebas mediante el comando:

Cross-valind ('Kfold', Clases, 10)

Como se puede ver se tienen que pasar como parámetro el número de particiones que realizara la función que para esta tesis fue diez particiones, y además las clases para poder balancear los grupos de entrenamiento.

Ahora podemos apreciar una comparación, de los mejores resultados obtenidos por las 3 técnicas siendo las redes neuronales con todos los atributos, la que a primera vista da el mejor resultado y el clasificador bayesiano es el que menor acierto tiene pero ya se mencionó, anteriormente que dichos resultados mejoraron al eliminar el atributo sexo pero no consiguen superar el de los 5 atributos con redes neuronales,

además se hizo una separación de los *outliers* con una función de matlab los cuales se colocaron solo en la clase de test y no en el entrenamiento, y se puede ver que los resultados no varían mucho siendo milésimas los cambios como se puede ver en la tabla.

Tabla 4.5: Comparación de los resultados obtenidos de las técnicas propuestas

Técnicas	% Acierto	% Error	Sensibilidad
SVM	98.06	1.94	0.9819
Redes Neuronales	99.0	1.0	0.947
Clasificador Bayesiano Uniforme	97.59	2.41	0.9460
Clasificador Bayesiano Tolerancia Ruido	98.01	1.99	0.9540
SVM sin atributo	98.59	1.41	0. 9403
Redes Neuronales sin atributo	98.9	1.06	0.9340
Clasificador Bayesiano sin atributo	97.61	2.39	0.9460
SVM sin outliers	98.04	1.96	0.9818
Redes Neuronales sin outliers	98.94	1.06	0.9459
Clasificador Bayesiano sin outliers	97.54	2.46	0.9457

CONCLUSIONES

Primera: Como se pudo apreciar en la sección de resultados se llegó a un 99% de acierto esto indica que, si se puede realizar el pronóstico de fallecimiento y la sobre vida en los pacientes hospitalizados, eso significa que es posible plantear técnicas de aprendizaje automático para apoyar a la toma de decisiones en sistemas de información hospitalarios como se plantea en el objetivo general.

Segunda: Fueron recolectados datos reales del Hospital General Honorio Delgado y tuvieron que ser analizados de forma adecuada para poder ser usadas por los clasificadores, siendo que si es factible usar datos reales en aplicaciones de inteligencia artificial, siendo analizadas tres técnicas automatizadas de aprendizaje automático donde en un sentido general las redes neuronales dieron el mejor resultado, pero después de hacer un análisis podemos ver que las máquinas de vectores de soporte tiene el menor error cuando se intenta clasificar a las personas que podrían fallecer, esta diferencia creemos que es el motivo para afirmar que SVM fue la mejor de las técnicas usadas en esta investigación.

Tercera: Es importante también tomar en cuenta el tratamiento de los datos, difícilmente existen sistemas que usen directamente los datos almacenados, como se presentó en la tesis se tuvieron que tomar en cuenta que existen errores en los registros, también valores vacíos, y atributos que tuvieron que ser abstraídos para poder mejorar la predicción siendo necesario realizar un proceso adecuado para la selección de características relevantes que con los resultados se demostró que si fueron satisfactorios.

Cuarta: A pesar de que el estado del arte indicaba ciertos atributos que pueden ser usados estos no fueron directamente utilizados esto implica que solo se tomaron como referencia y en el desarrollo de esta tesis concluimos en otros atributos que ayudaron a mejorar la predicción siendo que los resultados obtenidos en la presente tesis superaron los resultados del estado del arte lo cual nos permite afirmar que esta propuesta es original y con gran aporte a las áreas del conocimiento de predicción de muerte y sobre vida hospitalaria.

Quinta: En una comparación directa de las técnicas: Redes Neuronales, Clasificador Bayesiano y Máquinas de Vectores de Soporte, se puede ver que las redes neuronales se presentan como la mejor en este proceso de clasificación y el clasificador bayesiano como el peor, lo cual dice mucho de la potencialidad que tienen dichas técnicas.

RECOMENDACIONES

En cuanto a las recomendaciones que podemos dar:

- Es necesario que los sistemas de información hospitalaria tengan un mayor cuidado en la digitación de los registros debido a que por problemas de error fueron desechadas más de la mitad de los datos a usar, además que en algunos registros existían valores totalmente distintos a lo que debe registrar, se sugiere usar técnicas de normalizando en su base de datos para disminuir los errores en la información almacenada.

- Los modelos utilizados fueron usados como una simple simulación dado que este estudio dio resultados interesantes se sugiere que se implemente la propuesta en lenguaje de alto nivel de tal manera que este estudio se vuelva útil en una aplicación práctica.

TRABAJOS FUTUROS

Como trabajos futuros podemos decir

- Se puede implementar otros sistemas de predicción de partes específicas siguiendo la metodología desarrollada en la tesis por ejemplo para el caso específico de pacientes con cáncer, o problemas del corazón etc.

- Los resultados obtenidos demuestran que las técnicas de aprendizaje automático, pueden ser muy útiles no solamente para el caso de estudio si no también puede ser implementada en sistemas hospitalarios de la región y de todo el país.

- Finalmente es interesante validar las bases de datos de las instituciones públicas lo que es importante para poder desarrollar aplicativos y herramientas como la que se proponen en la presente tesis.

BIBLIOGRAFÍA

Ackley D., H. G. (1985). A learning algorithm for boltzman machines. *Cognitive Science*, 147–169.

Archibald, R. C. (2012). Prediction of In-Hospital Mortality in Acute Exacerbations of COPD. *Scottish Universities Medical Journal*, 1(2).

Barber, D. (2012). *Bayesian reasoning and machine learning.* Cambridge University Press.

Cai, X. P.-C.-S. (2016). Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, págs. 23(3), 553-561.

Cao, L. P. (2008). *Data mining for business applications.* Springer Science & Business Media.

Cardona, A. M. (2012). Aplicación de árboles de decisión en la salud pública. . *Revista CES Salud Pública*, 3(1), 94.

Carpenter G., S. G. (1992). *Neural Network for Vision and Image Processing.* Massachusetts Institute of Tegnology.

Christian Suca, A. C. (2016). Comparacion De Algoritmos De Clasificacion Para La Prediccion De Casos De Obesidad Infantil. *Researchgate*.

Cristianini, N. &.-T. (2000). *An introduction to support vector machines and other kernel-based learning methods.* Cambridge university press.

D. Hush, B. H. (1993). Progress in supervised neural networks. *IEEE Signal Processing Magazine*, 10(1):8–39.

D. Rumelhart, G. H. (1986). *Learning representations by back-propagating errors.* Nature (London).

Demuth, H. B. (2014). *Neural network design.* Martin Hagan.

Elloumi, M. &. (2013). *Biological Knowledge Discovery Handbook: Preprocessing, Mining and Postprocessing of Biological Data* . John Wiley & Sons.

Francia Santamaria, E. &. (2013). *Predicción de la mortalidad intrahospitalaria en medicina interna.* Bracelona: Tesis de Doctorado.

Freeman J, M. S. (1991). *Neural Networks: Algorithms, Applications, and Programming Techniques.* Addison Wesley.

Gilbert, K. A. (2006). Mineria de datos: Conceptos y tendencias. . *Asociacion Española para la Inteligencia Artificial (AEPIA)*, 11-18.

Gomes, A. S. (2010). Mortality prediction model using data from the Hospital Information System. *Revista de saude publica*, págs. 44(5), 934-941.

Guareno, M. A. (2016). *Support vector regression: propiedades y aplicaciones.* Sevilla: Tesis de Bachiler.

Gutierrez, J. (2006). *Nueva Red Caótica para el Reconocimiento de Patrones Multivalor.* Arequipa: Tesis Universdiad Nacional San Agustin.

Haykin, S. S. (2001). *Kalman filtering and neural networks* . New York: Wiley.

Hebb, D. (1949). *The organization of behavior; a neuropsychological theory.* Wiley-Interscience.

Hilera González, J. R. (2000). *Redes neuronales artificiales: fundamentos, modelos y aplicaciones.* Alfaomega .

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *PNAS of the U.S.A.*, 79:2554–2558.

Kohonen, T. (1988). *Self-Organization and Associative Memory.* Springer-Verlag.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 249-268.

Lewis, D. D. (1998). *Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning* . Springer, Berlin, Heidelberg.

Martínez, R. E. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana*, 9(2), 19-24.

Paliouras, G. A. (2016). Machine learning and its applications: advanced lectures. *Springer*.

Piateski, G. &. (1991). *Knowledge discovery in databases.* MIT press.

Rahmanian, P. B. (2010). Predicting hospital mortality and analysis of long-term survival after major noncardiac complications in cardiac surgery patients. *The Annals of thoracic surgery*, págs. 90(4) 1221-1229.

Rodriguez Porrero, C. (2016). *Ciudades amigables con la edad, accesibles e inteligentes.* CEAPAT-IMSERSO.

Ruiz Hidalgo, D. &. (2016). *Desarrollo y validación de un modelo predictivo de mortalidad a corto plazo en ancianos ingresados por patología médica.* Barcelona: Tesis Doctoral.

Sanchez, S. E. (2016). Implementacion de Algoritmos de Inteligencia Artificial para el Entrenamiento de Redes Neuronales de Segunda Generación. *Jovenes en la Ciencia*, 6-10.

Sanz, J. A. (2014). Medical diagnosis of cardiovascular diseases using an interval-valued fuzzy rule-based classification system. *Elsevier*, 103-111.

Sanz, J. A. (2015). A compact evolutionary interval-valued fuzzy rule-based classification system for the modeling and prediction of real-world financial

applications with imbalanced data. *IEEE*, 973-990.

Satyani Manthale, P. K. (2017). Survey on Interesting Pattern Classification. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(3), 3833- 3837.

Savastano, L. B. (2009). *Análisis de la mortalidad en la unidad de cuidados intensivos del Hospital Central de Mendoza.* Mendoza: Hospital Mendoza Argentina.

Scholkopf, B. &. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond.* MIT press.

Serna, N. R. (2016). *Predicción de Readmisiones, Mortalidad, e Infecciones en la UCI usando Técnicas de Aprendizaje de Máquinas.* Cali: Universidad de los andes.

Shams, I. A. (2014). A predictive analytics approach to reducing avoidable hospital readmission. *arXiv preprint arXiv*, 1402-5991.

Sierra Araujo, B. (2006). *Aprendizaje automatico: conceptos basicos y avanzados: aspectos practicos utilizando el software Weka.* Madrid: Pearson Prentice Hall.

Stork, D. G. (2001). *Patter Classification* . Wiley.

Suca C., C. A. (2016). Comparación de algoritmos de clasificación para la predicción de casos de obesidad infantil. *ResearchGate*.

Tase, R. O. (2016). Nuevo Algoritmo Multiclasificador Para Flujos De Datos Con Cambios De Concepto. *Capa*, V.2.

Theodoridis, S. P. (2009). *Introduction to pattern recognition: a matlab approach.* Academic Press.

Ticona, R. &. (2005). Mortalidad perinatal hospitalaria en el Perú: factores de riesgo. 70(5), 313-317.

Torres, C. Z. (2016). *Estudio de variables que influyen en la desercion de estudiantes universitarios de primer ano, mediante mineria de datos.* Ciencia Amazonica.

Valdivia, J. (2015). *Recursos, ejemplos y documentación sobre herramientas de Business Intelligence. Ejemplos prácticos en tecnología SAS y otras herramientas.* http://sasybi.blogspot.pe/2015/05/arboles-de-decision-en-sas.html.

Vapnik, V. N. (1988). *Statistical learning theory* . New York: Wiley.

Vazquez, S. R. (2016). Evaluacion de alternativas para la clasificacion de celulas cervicales utilizando solo rasgos del nucleo. *Revista Cubana de Ciencias Informáticas*, 211-222.

Wang, L. (2005). *Support vector machines: theory and applications* . Springer Science & Business Media.