INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA

MAESTRÍA EN SISTEMAS COMPUTACIONALES

“MODELO DE CLASIFICACIÓN DE NEFROPATÍA DIABÉTICA MEDIANTE APRENDIZAJE AUTOMÁTICO”

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRO EN SISTSEMAS COMPUTACIONALES

PRESENTA

FRANCIS SUSANA CARRETO ESPINOZA.

Asesor

MSC. José Antonio Hiram Vázquez López.

MISANTLA, VER. ENERO, 2018

INSTITUTO TECNOLÓGICO SUPERIOR DE MISANTLA
DIVISIÓN DE ESTUDIOS PROFESIONALES
AUTORIZACIÓN DE IMPRESIÓN DE TRABAJO DE TITULACIÓN MAESTRÍA

FECHA: 12 de Enero de 2015

ASUNTO: **AUTORIZACIÓN DE IMPRESIÓN DE TESIS.**

A QUIEN CORRESPONDA:

Por medio de la presente se hace constar que el (la) C:

FRANCIS SUSANA CARRETO ESPINOZA

estudiante de la maestría en SISTEMAS COMPUTACIONALES con No. de Control 152T0731 ha cumplido satisfactoriamente con lo estipulado por el **Lineamiento de Posgrado para la obtención del grado de Maestría** mediante **Tesis.**

Por tal motivo se **Autoriza** la impresión del **Tema** titulado:

MODELO DE CLASIFICACIÓN DE NEFROPATÍA DIABÉTICA MEDIANTE APRENDIZAJE AUTOMÁTICO

Dándose un plazo no mayor de un mes de la expedición de la presente a la solicitud del examen para la obtención del grado de maestría.

ATENTAMENTE

M.S.C. José Antonio Hiram Vázquez López
Presidente

M.I.A. Roberto Ángel Meléndez Armenta
Secretario

M.S.C. Eddy Sánchez de la Cruz
Vocal

Archivo.

VER. 01/03/09 F-SA–39

Dedicatoria

Con profundo amor a ***mis padres****.*
Hilda y Efraín porque me formaron con reglas y
con algunas libertades, pero al final de cuentas,
me motivaron constantemente para alcanzar mis anhelos.
Por sus consejos, su apoyo incondicional y su paciencia.
Todo lo que soy es gracias a ustedes y aquí estoy,
con un nuevo logro exitosamente alcanzado.

A ***mis hermanos,***
Lizbeth y Omar no solo por estar presentes aportando buenas cosas a mi vida,
sino por la gran felicidad que siempre me han causado.
Este logro también es de ustedes.

Con especial afecto a ***Emmanuel***
porque su ayuda ha sido fundamental, estuviste a mi lado incluso en los momentos y
situaciones más complicadas, siempre brindándome tu apoyo incondicional.
No fue sencillo alcanzar esta meta, sin embargo siempre fuiste muy motivador y
nunca dejaste de confiar en mí, me decías que lo lograría perfectamente.
Me ayudaste hasta donde te era posible, incluso más que eso.

A mis ***amigos****,*
por permitirme aprender más de la vida a su lado.
Les agradezco no solo por la ayuda brindada,
sino por los buenos momentos en los que convivimos.

Muchas gracias a aquellos seres que siempre aguardo en mi corazón.

Francis Susana Carreto Espinoza.

Agradecimientos

Dios, te agradezco por darme la vida, salud y las bendiciones necesarias para lograr mis metas. Gracias por la familia tan maravillosa que me diste, quienes siempre han creído en mí y cada día me motivan a ser mejor, dándome ejemplo de superación, humildad y sacrificio; enseñándome a valorar todo lo que tengo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada para poder cursar la maestría.

Agradezco también a **mis profesores**, Hiram y Luis por haberme brindado la oportunidad de recurrir a su capacidad y conocimiento, además por haberme tenido toda la paciencia del mundo para guiarme a lo largo de este camino.

Y a todas aquellas personas que de una u otra manera colaboraron en mi formación dentro y fuera de la institución.

"No estudio por saber más,

sino por ignorar menos"

Sor Juana Inés de la Cruz.

Resumen

La nefropatía diabética es un padecimiento que ocasiona graves problemas de salud en la población adulta con diabetes. Una de las consecuencias de la nefropatía se ve reflejada en la afectación de los riñones para realizar la función de eliminar productos de desecho y agua en exceso. En esta tesis se presenta un modelo de clasificación como apoyo a los médicos para identificar el nivel de afectación en los riñones de pacientes con diabetes mellitus tipo 2. La presente investigación se realizó en el Hospital General de Misantla, ubicado en la ciudad de Misantla durante el periodo comprendido entre Enero-Diciembre 2017. La población de estudio corresponde a 55 pacientes con diagnóstico nefrótico donde 32 son mujeres y 23 hombres. La selección de los expedientes que participaron en la elaboración del conjunto de datos se realizó por conveniencia y su relevancia social ya que servirá para que el médico sin experiencia determine el grado de deterioro renal. El conjunto de datos digital se creó a partir del análisis de los factores de riesgo que determinan la nefropatía diabética. Para el preprocesamiento de los datos con el fin de eliminar la inconsistencia, ruido y redundancia se utilizó la media para datos continuos y la moda para datos discretos. La selección de atributos fue realizada con base en chi2 para encontrar el subconjunto de las variables que tienen mayor correlación. El modelo de clasificación automático de las etapas de la nefropatía fue realizado bajo el algoritmo de aprendizaje: Árbol de regresión y clasificación (CART). Para medir el rendimiento del modelo se procedió a una evaluación mediante un análisis de sensibilidad obteniendo un 90% de efectividad en los casos de prueba.

Abstract

The diabetic nephropathy is an ailing that cause serious problems of health in the old people with diabetes. One of the consequences of the nephropathy is reflected in the kidneys involvement for perform the function of delete waste products and excess water. In this tesis is presented a classification model as support to physician for identify the kidneys involvement level of patients with type 2 diabetes mellitus. This research was perform in the General Hospital of Misantla, located in the Misantla city during the period between January-December 2017. The study population corresponds at 55 patients with nephrotic diagnostic where 32 are woman and 23 men. The selection of the records that participated in the elaboration of dataset it was done for convenience and its social relevance as it will serve for that physician without experience determine the degree of renal deterioration. The digital dataset was created from the analysis of the risk factors that determine the diabetic nephropathy. For preprocessing of data in order to eliminate inconsistency, noise and redundance, the mean was used for continuous data and mode for discrete data. The attributes selection was made based on chi2 to find the subset of the variables with greater correlation. The automatic classification model of stages of nephropathy was carried out under the algorithm of learning; Classification and Regression Tree (CART). To measure performance of the model was carried out using a sensitivity analysis obtaining 90% effectiveness in the test cases.

Índice general

Índice de figuras

Índice de tablas

Índice de ecuaciones

Introducción

La presente tesis se refiere al tema de nefropatía diabética, que se puede definir como el daño renal que padecen las personas con diabetes. O aquella considerada como complicación microvascular y una de las principales causas de muerte en los pacientes con diabetes mellitus tipo 2 que tiene como consecuencia la Enfermedad Renal Crónica (ERC), la cual si no se detecta a tiempo se convierte en una Enfermedad Renal Crónica Terminal (ERCT) donde los tratamientos sustitutivos tienen costos muy elevados [20].

Una de las características principales de la nefropatía diabética es que, comúnmente aparece después de unos 5 años de haber diagnosticado la diabetes y se ha clasificado en cinco estadios que determinan el nivel de afectación al riñón, que se mide de acuerdo a la microalbuminaria en las muestras de orina [1].

Para analizar el tema de tesis es importante mencionar las causas que originan la enfermedad. Algunas de ellas son el mal control de la microalbuminuria, la glucosa, la obesidad y el sobrepeso, además de la hipertensión arterial que ocasionan que el daño renal aumente progresivamente y no se detecte hasta que los daños son graves [40]. Por lo tanto, es importante determinar el nivel de daño renal en pacientes diabéticos para proporcionar un tratamiento adecuado y lograr de esta manera mejorar su calidad de vida.

Como objetivo principal de esta investigación se diseñó un modelo de clasificación para nefropatía diabética utilizando técnicas de aprendizaje automático que sirva como herramienta a los médicos para que se le proporcione al paciente un tratamiento en una etapa adecuada que retrase el progreso de la enfermedad o logre que esta permanezca estable.

Para alcanzar el objetivo de la investigación se llevaron a cabo tareas de gran importancia como elegir un método adecuado para la selección de características según el tipo de datos que se tenían almacenados, realizar el preprocesamiento adecuado a los datos y por último encontrar un algoritmo de clasificación óptimo para el problema que se estaba tratando.

El contenido de la investigación se ha divido en los siguientes capítulos:

En el primer capítulo se encuentra el protocolo de la investigación que está integrado por descripción del problema, justificación, objetivos general y específicos, hipótesis, propuesta de solución y metodología.

El segundo capítulo contiene algunas definiciones, conceptos y trabajos que se han elaborado con relación al tema de esta investigación, los cuales se dividen en dos áreas (i) médica y (ii) tecnológica, en (i) los autores dan a conocer definiciones del padecimiento, la clasificación que existe de la misma así como posibles tratamientos. Por otra parte en (ii) se observan técnicas o algoritmos que se han utilizado para la predicción y clasificación de enfermedades como la diabetes, y en específico sobre nefropatía diabética y la eficiencia que tienen cada uno de los algoritmos utilizados.

En el tercer capítulo se presenta la metodología que se realizó para poder llevar a cabo el diseño y desarrollo del modelo de clasificación de nefropatía diabética.

En el cuarto capítulo se realiza un análisis e interpretación de los resultados obtenidos al utilizar el modelo que se diseñó.

Por último, en el quinto capítulo se muestran las conclusiones y el trabajo a futuro de esta investigación.

Capítulo I

Generalidades

1.1. Descripción del problema.

En la actualidad la diabetes es una de las principales causas de muerte no solo en México sino a nivel mundial, debido principalmente a la obesidad y el sobrepeso. Hasta hace poco 7 de cada 10 individuos con edades mayores a los 20 años tenían problemas de exceso de peso (obesidad y sobrepeso) lo que significa un 72.5% de la población [19].

El Fondo de las Naciones Unidas para la Infancia (UNICEF) reportó que México ocupa el primer lugar a nivel mundial en obesidad infantil. Así mismo, el Instituto Mexicano del Seguro Social (IMSS) en 2010 dio a conocer que el 9.3% de la población en general son pacientes diabéticos con edades que están por debajo de los 15 años.

La diabetes mellitus es una enfermedad frecuente ocasionada por el incremento de glucosa en la sangre debido a que el páncreas no produce la insulina suficiente o el organismo no la administra de manera correcta [20]. Las principales complicaciones de la diabetes son infarto al

miocardio o accidentes cerebrovasculares, pie diabético, retinopatía, neuropatía y nefropatía diabética.

La nefropatía diabética es una complicación microvascular crónica causante principal de la Enfermedad Renal Crónica Terminal (ERCT), además de ser una de las causas básicas de muerte en pacientes diabéticos de los dos tipos (1 y 2). Es una enfermedad que surge aproximadamente 5 años después de la detección de diabetes en el paciente y las causas que ocasionan que los individuos con diabetes estén propensos a la nefropatía diabética son:

- La obesidad o sobrepeso es un problema que se encuentra latente en el país y que día a día se incrementa, debido a los malos hábitos alimenticios, es decir, al consumo en grandes cantidades de grasas y azucares.
- El mal control de la glucosa surge principalmente por la mala alimentación (ingesta de azucares y grasas), falta de actividad física y un tratamiento farmacéutico no adecuado al paciente.
- La hipertensión arterial es uno de los factores de riesgo más importantes para que los pacientes diabéticos padezcan nefropatía.

La microalbuminuria es la primera manifestación clínica de la nefropatía, hay que destacar que es un elemento considerado como un buen predictor para la detección de daño en los riñones y es necesario que en esta etapa se tomen medidas que ayuden a disminuir el daño renal producido en los pacientes diabéticos.

Por otra parte, no todos los profesionales de la salud se encuentran capacitados para realizar un diagnóstico de esta enfermedad, lo que implica un retraso en el diagnóstico debido a la falta de conocimientos sobre dicha enfermedad. Así mismo, no existe una herramienta computacional que realice la clasificación del nivel de afectación de los riñones en pacientes con diabetes mellitus tipo 2 mediante el uso de datos clínicos y demográficos que están al alcance de la mayoría de las personas.

Desde hace ya varios años los expertos en enfermedades renales realizan el diagnóstico de la enfermedad mediante la clasificación KDIGO que se encuentra en las guías prácticas clínicas, donde se definen cinco niveles de afectación basados en el filtrado glomerular, para lo cual es necesario saber parámetros como edad, sexo y creatinina.

Uno de los principales problemas del sector salud en México se manifiesta en la falta de herramientas computacionales para la clasificación, diagnóstico, monitoreo y tratamiento de enfermedades de las personas que acuden a sus centros de salud. Es decir, actualmente las

personas deben asistir con personal capacitado en la enfermedad quien determina el nivel del daño renal a partir de ciertas pruebas clínicas y bioquímicas.

Para la detección de la nefropatía diabética se realizan pruebas bioquímicas como la biopsia renal y el examen general de orina (EGO), además de medidas clínicas dentro de las que se encuentran peso, índice de masa corporal (IMC), perímetro de cintura e hipertensión arterial (HTA). Es importante saber que los pacientes que presentan daño excesivo en los riñones aún tienen posibilidad de tener un control de su enfermedad mediante el uso de tratamientos sustitutivos como son:

- diálisis peritoneal,
- hemodiálisis y
- trasplante de riñón.

Sin embargo, el costo de estos tratamientos ha ido aumentando por lo que es necesario implementar medidas preventivas respecto a este problema de salud. Para esta investigación es importante realizar un análisis de los datos que determine de acuerdo a los factores de riesgo (tipos) el nivel de gravedad de la enfermedad sin recurrir a pruebas invasivas para los pacientes.

El desarrollo de herramientas computacionales para dar solución a problemas de la vida real se refleja en la integridad de los datos y la falta de estándares o reglas para el almacenamiento digital de conjuntos de datos. La secretaria de salud en México aún no cuenta con expedientes digitales y por ende no se tiene un conjunto de datos disponibles para trabajar, lo que significa que los expedientes que se manejan actualmente son físicos, y por tanto no existe una estructura que facilite el manejo de los mismos. Es por ello que a partir de los expedientes y con la autorización de pacientes y médicos responsables se debe determinar cuáles son los factores de riesgo de la nefropatía en pacientes con diabetes tipo 2.

Otro problema frecuente es la falta de datos en los expedientes y la dificultad que esto ocasiona al determinar cuáles son de mayor relevancia para la enfermedad.

Por lo tanto, es necesaria la clasificación de la nefropatía en pacientes diabéticos para ofrecer una mejor calidad de vida mediante intervenciones terapéuticas que cambien o ralentice el curso de la enfermedad.

1.2. Justificación

La nefropatía diabética es una complicación crónica de la diabetes que se presenta aproximadamente después de 5 años de su evolución, es ocasionada por las lesiones que ocurren en los riñones y conlleva a una ERT.

Esta enfermedad se desarrolla en un 30 a 40% en pacientes con diabetes mellitus tipo 1 después de haber transcurrido entre 15 y 20 años desde el diagnóstico de la diabetes y es considerada la principal causa de morbimortalidad en este grupo; mientras que en los pacientes con diabetes mellitus tipo 2 solo se presenta en un 5 a 10%, cifra que en realidad no es baja debido a que este grupo representa el 90% de la población diabética (imagen 1) [7].

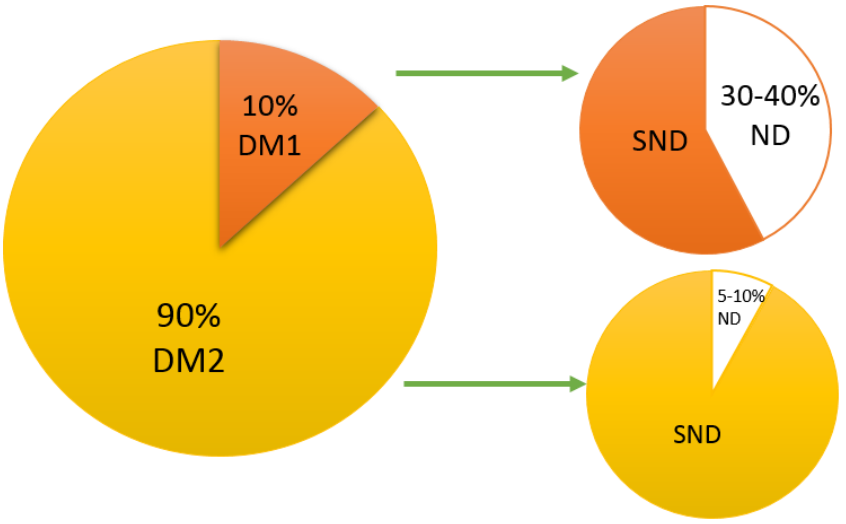

Imagen 1.- Población con nefropatía diabética tipo 1 y 2.

La aplicación de inteligencia artificial en la medicina se ha convertido en un área de gran impacto debido al manejo de una extensa cantidad de datos y el desarrollo de sistemas con comportamientos "inteligentes". Dentro del área médica se tienen gran variedad de aplicaciones que se utilizan para la detección, monitoreo y tratamiento de enfermedades.

Actualmente el sector salud, no cuenta con un sistema de clasificación de nefropatía diabética que determine a partir de ciertas características el estado de afectación en el que se encuentra el paciente. Es por ello que surge la necesidad de elaborar una solución aplicando técnicas computacionales.

En consecuencia, para el desarrollo de esta investigación se diseñará un modelo de clasificación de nefropatía diabética aplicando técnicas de aprendizaje automático. Mediante la aplicación de aprendizaje automático se podrá realizar el análisis de diversas variables que son representativas para determinar el estado nefrótico del paciente.

Así mismo, se espera que el modelo sea una alternativa computacional a la toma de decisiones de los médicos para proporcionar un tratamiento a los pacientes y la identificación de problemas de riñón en diabéticos tipo 2.

1.3. Objetivos

1.3.1. Objetivo general

Diseñar un modelo de clasificación de nefropatía diabética aplicando técnicas de aprendizaje automático como herramienta de apoyo a los médicos para proporcionar un tratamiento adecuado y oportuno a pacientes con la enfermedad dependiendo del grado de afectación.

1.3.2. Objetivos específicos

(1) Crear un conjunto de datos digital a partir de los factores de riesgo que determinan la nefropatía en pacientes diabéticos tipo 2.
(2) Realizar un preprocesamiento de los datos obtenidos en (1) para eliminar inconsistencia, ruido y redundancia en los mismos.
(3) Analizar y elegir un método de selección de características que ayude a encontrar el subconjunto de las variables que tienen mayor correlación o relevancia del conjunto de datos original.
(4) Analizar y elegir un algoritmo de clasificación que a partir del subconjunto encontrado en (3) agrupe los datos en las clases predefinidas dependiendo del caso que se trate.
(5) Realizar un análisis del comportamiento del modelo para determinar la confiabilidad del mismo a través de medidas de desempeño como la sensibilidad.

1.4. Hipótesis

Es posible determinar el grado de afectación renal en 90% de los pacientes con diabetes mellitus tipo 2 del HGM mediante el uso de técnicas de aprendizaje automático, que sirva como

herramienta de apoyo a los médicos para proporcionar un tratamiento adecuado además de mejorar la calidad de vida de los pacientes.

1.5. Propuesta de solución

Para dar solución al problema de clasificación de nefropatía en pacientes diabéticos, se propone el diseño de un modelo de clasificación para determinar el estadio al que pertenece el paciente basado en técnicas de aprendizaje automático que servirán para analizar y clasificar los datos.

El proceso para elaborar el modelo de clasificación consta de las siguientes etapas como se muestra en la imagen 2.

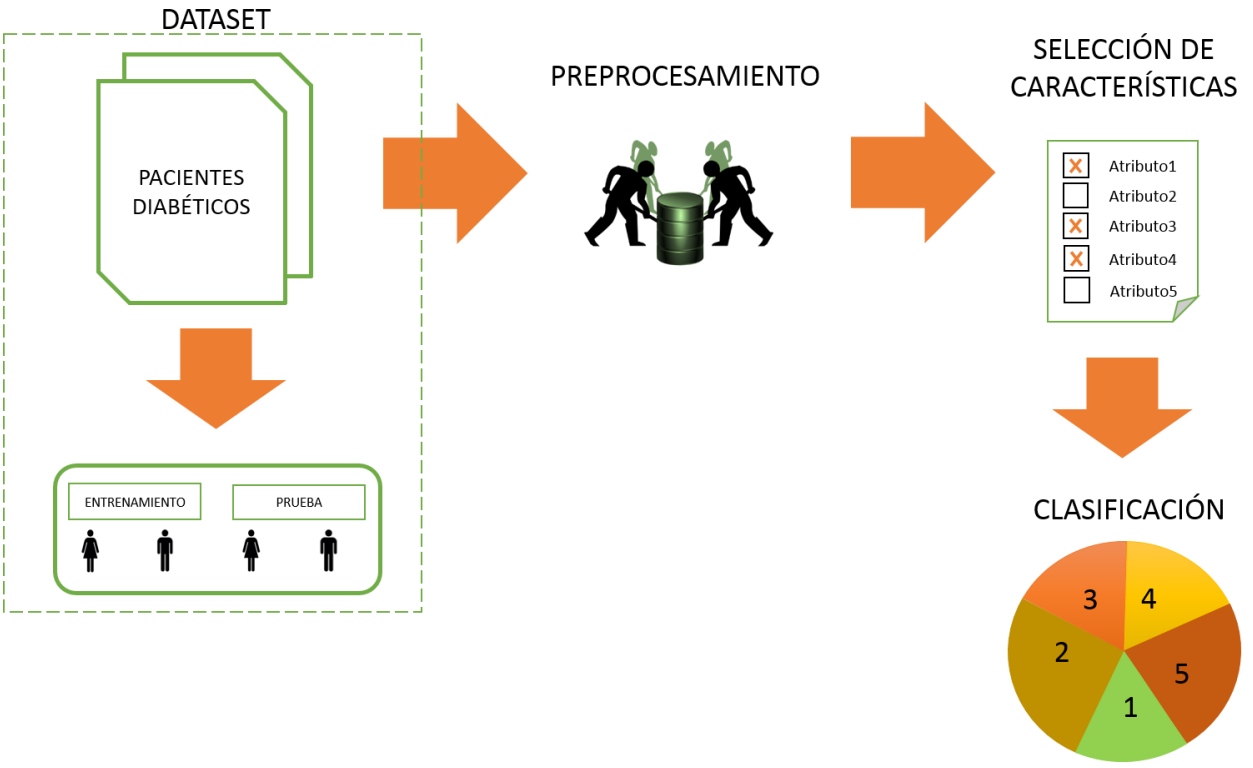

Imagen 2.- Proceso propuesto para obtener el modelo de clasificación.

Enseguida se definen a detalle cada una de las etapas del modelo en general.

1. **Dataset:** El conjunto de datos que se generó se obtuvo de la información de los pacientes diabéticos tipo 2 con diagnóstico nefrótico en cualquier estadio y que acuden a consultas

al HGM ubicado en Misantla, Veracruz. Asimismo, se divide en dos subconjuntos que son el conjunto de entrenamiento y prueba necesarios para la etapa de clasificación.

2. **Preprocesamiento:** Este paso es muy importante debido a que la mayoría de los datos que se recopilan del mundo real contienen ruido o están incompletos. Es por ello que se debe realizar el preprocesamiento necesario al conjunto de datos, dicho proceso consiste en seleccionar, limpiar y transformar los datos que se van analizar, puesto que ayuda a maximizar el rendimiento del modelo.

3. **Selección de características:** Las bases de datos almacenan grandes cantidades de datos, es por ello que esta etapa consiste en establecer una búsqueda para encontrar un subconjunto de características o atributos que sirva para diseñar el clasificador. Al encontrar un subconjunto con mayor relevancia el clasificador se ahorra tiempo y memoria en la gestión de datos que son de poca importancia y por lo tanto podría mejorar el desempeño del clasificador.

4. **Clasificación:** La clasificación consiste en agrupar los datos en clases predefinidas de acuerdo a ciertas características que tienen en común. Para el proceso de clasificación se utiliza el subconjunto de características previamente seleccionadas, ya que a partir de dichas características se encuentran patrones que determinan la clase a la que pertenecen. En la etapa de clasificación se realizan dos tareas importantes:
    - **Entrenamiento:** consiste en alimentar el clasificador con el conjunto de entrenamiento para encontrar patrones representativos del conjunto de datos.
    - **Prueba:** se evalúan los patrones encontrados, es decir, se evalúa el aprendizaje del clasificador utilizando el conjunto de prueba y observando los resultados se determina la precisión del mismo.

1.6. Metodología

Las pruebas médicas que se realizan para diagnosticar nefropatía diabética miden filtración glomerular, hemoglobina glicosilada o creatinina por mencionar algunas. Sin embargo, no todos los médicos son capaces de determinar a partir de esos parámetros que el paciente presenta un alto o bajo grado de nefropatía diabética.

Por lo anterior, se propone el diseño de un modelo de clasificación de nefropatía en pacientes diabéticos que encuentre su estadio de afectación renal, para lograr el objetivo de la investigación se llevaron a cabo las siguientes etapas:

1. **Recolección de los datos.**
   - ❖ Buscar una institución o médico que brinde atención a pacientes diabéticos que presenten la complicación nefrótica.
   - ❖ Investigar en diversas fuentes de salud cuales son las variables determinadas como factores de riesgo en el diagnóstico de la nefropatía diabética, además de solicitar la ayuda de un experto quien se encarga de validar y agregar si es necesario variables a partir de la hoja de hospitalización o ingreso que se realiza al asistir a la consulta.

2. **Digitalización de datos.**
   - ❖ Realizar la digitalización de los datos que serán almacenados en la base de datos ya que la institución que los proporcionará realiza el seguimiento y control de los pacientes mediante un registro en expedientes físicos.
   - ❖ **Análisis de los datos.**
     - o Definir los subconjuntos de entrenamiento y prueba que se utilizan en el clasificador para la fase de aprendizaje y evaluación. El conjunto de entrenamiento consta de los datos que se utilizan para alimentar el clasificador y obtener los patrones representativos del modelo. Por otra parte, el conjunto de prueba sirve para evaluar los patrones encontrados en la fase de entrenamiento y determinar la clasificación a la que pertenecen.
     - o Realizar una investigación sobre los trabajos relacionados en el estado del arte que sirva como ayuda para seleccionar los algoritmos o técnicas a utilizar para la resolución del problema planteado.

3. **Preprocesamiento de datos.**
   - ❖ Los datos que se recopilan de problemáticas de la vida real son inconsistentes, ruidosos o incompletos, es por ello que en esta etapa se realiza un proceso de tratamiento o transformación de los datos. El proceso consiste en completar datos perdidos, disminuir el ruido, eliminar valores atípicos y resolver inconsistencia de los datos, además de tratar la redundancia de los mismos.

4. **Selección de características.**
   - Investigar y analizar el comportamiento de los métodos de selección de características para seleccionar uno que ayude a dar solución al problema planteado. Mismo que se utiliza para obtener un subconjunto de características según la importancia que cada característica represente en el conjunto de datos, ya que servirá como entrada del algoritmo clasificador y proveerá mejores resultados en cuanto a la clasificación de la enfermedad.

5. **Clasificación.**
   - Investigar y analizar los algoritmos de clasificación y seleccionar uno para utilizar como clasificador en el modelo.
   - Utilizar el algoritmo clasificador seleccionado en el paso anterior y alimentar el algoritmo con el conjunto de datos de entrenamiento preprocesados para encontrar patrones.
   - Poner a prueba el conocimiento adquirido en el paso anterior mediante la utilización de los datos de prueba, es decir, se realiza una comparación con los patrones encontrados y se clasifica para determinar la clase a la que pertenecen.

6. **Desempeño del clasificador.**
   - Por último se mide el desempeño del clasificador, para lo cual se utiliza la matriz de confusión y la sensibilidad que presenta el modelo después de realizar las pruebas.

Capítulo II

Marco Teórico

La detección temprana o clasificación de enfermedades crónicas como la diabetes es de vital importancia para ofrecer tratamientos precoces, oportunos y adecuados que permitan a los pacientes tener una mejor calidad de vida.

En los últimos años el uso de las ramas de la computación en la medicina ha incrementado debido a que existen diversas áreas en las que es necesario adoptar nuevas tecnologías dentro de las que se encuentran monitoreo, control, predicción, clasificación y tratamiento de algunas enfermedades [37]. Con respecto a la diabetes y sus complicaciones existen gran variedad de soluciones informáticas que se han llevado a cabo para ayudar a personas y médicos que tratan retinopatía, nefropatía y enfermedades cardiovasculares por mencionar algunos además de la diabetes, las cuales se realizan a partir de inteligencia artificial y sus derivados.

Los métodos que más se han utilizado para la predicción y clasificación de enfermedades es la minería de datos, la cual a partir de uno o un conjunto de algoritmos permite el diseño de los modelos o sistemas predictivos y/o de clasificación, y el aprendizaje automático que aprende cada vez que se incrementa o crece el conjunto de datos ya que los patrones que se encuentran son cada vez mejores.

2.1. Definiciones y conceptos

2.1.1. Diabetes mellitus

La diabetes mellitus es una enfermedad crónico-degenerativa y hereditaria, ocasionada por el incremento de glucosa en la sangre debido a que el páncreas no produce la insulina suficiente o el organismo no la administra de manera correcta, además es una de las principales causas de muerte, debido principalmente a la obesidad y el sobrepeso.

Las complicaciones de la diabetes pueden ser de dos tipos: (i) macrovasculares y (ii) microvasculares; en el primer tipo (i) se encuentran las enfermedades cardiovasculares como infarto al miocardio, accidentes cerebrovasculares e insuficiencia circulatoria en los miembros inferiores, mientras que los padecimientos del tipo (ii) son retinopatía, pie diabético, neuropatía y nefropatía [20].

2.1.2. Nefropatía diabética

La diabetes mellitus es una enfermedad crónica que consiste en la mala administración de la glucosa en el cuerpo. Los riñones se encargan de filtrar la sangre, es decir, toda la sangre que viaja a lo largo del cuerpo pasa muchas veces por los riñones. Por lo anterior, se tiene que si la glucosa se queda en la sangre en lugar de metabolizarse surge la nefropatía diabética que es considerada una de las complicaciones de la diabetes, misma que si no se toman los cuidados y tratamientos necesarios a tiempo puede convertirse en una enfermedad renal crónica hasta llegar a una etapa terminal.

Se han realizado diversas clasificaciones de la evolución del daño renal en pacientes diabéticos de ambos tipos, la clasificación aportada por Mongensen en 1983 clasifica la nefropatía diabética en 5 etapas [1]:

1. Hipertrofia renal - hiperfiltración glomerular: Esta etapa es reversible si tiene un buen control, es común que aparezca después de 5 años de que la diabetes ha sido detectada y produce frecuentemente hiperfiltración en pacientes con diabetes tipo 1.
2. Normoalbuminuria: La excreción de albuminuria (EUA) se encuentra por debajo de los 20 mg/min y se puede tener un filtrado glomerular normalizado a partir de un control estricto de la diabetes.
3. Nefropatía incipiente: La EUA está entre 20-200 mg/min y aparece en un rango de 6-15 años de evolución de la diabetes, es reversible si se tiene un buen control glucémico óptimo más inhibidor de la enzima convertidora de angiotensina (IECA).

4. Nefropatía clínica: También llamada nefropatía establecida donde la EUA es superior a los 200 mg/min y aparece después de 15-25 años de evolución.
5. Insuficiencia renal crónica (IRC): El daño que se ha causado a los riñones es irreversible a pesar de un buen control diabético, sin embargo existen tratamientos sustitutivos que se emplean en estos casos como diálisis peritoneal, hemodiálisis y trasplante de riñón.

En 2009, surge una nueva clasificación para la nefropatía diabética que mide el daño renal mediante la filtración glomerular y forma parte de las guías prácticas clínicas de Enfermedad renal: Mejora de los resultados mundiales (KDIGO, por sus siglas en inglés *Kidney Disease: Improving Global Outcomes*) clasificación que se muestra en la Tabla 1 [14].

ESTADIO ERC		FG (ml/min/1.73 m2)	DESCRIPCIÓN
1		>=90	Daño renal con FG normal.
2		60-89	Daño renal y ligero descenso del FG.
3ª	ERC* IRC*	45-59	Descenso ligero – moderado del FG.
3B		30-44	Descenso moderado de FG.
4		15-29	Descenso grave de FG.
5		<15	Prediálisis.
5D		Diálisis	Diálisis.

Tabla 1.- Clasificación KDIGO.

En un estudio realizado en [7] se identifica como pilares del tratamiento de la nefropatía diabética los siguientes:

- control metabólico.
- plan alimentario: proteínas, hidratos, grasas, sodio.
- control lipídico.
- control de la presión arterial.
- control de la proteinuria.
- control del tabaquismo.

2.1.3. Filtrado glomerular

La tasa de filtrado glomerular (TFG) es una medida del funcionalismo de los riñones. Esta prueba utiliza el resultado de la determinación de creatinina en sangre, incluyéndolo en una fórmula de estimación del filtrado glomerular, cuyo resultado refleja el grado de funcionamiento de los riñones. La medida de la TFG se considera como la forma más exacta de detectar cambios en el

estado de los riñones. Sin embargo, la medida directa de la TFG es complicada y requiere personal muy entrenado. Por esta razón se utiliza a menudo una estimación, la tasa estimada de filtrado glomerular o TEFG.

La TEFG es un cálculo basado en el resultado de la medida de la creatinina sérica. La creatinina es un producto de desecho del músculo que se filtra por los riñones y se excreta en la orina a una tasa relativamente constante. Cuando la función renal disminuye, se excreta menor cantidad de creatinina por la orina de manera que sus concentraciones en sangre aumentan. Con el resultado de la prueba de la creatinina se obtiene una estimación razonable de la TFG real [41].

2.1.4. Minería de datos

Minería de datos (DM, por sus siglas en inglés *Data Mining*) es una etapa del descubrimiento del conocimiento en base de datos (KDD, por sus siglas en inglés *Knowledge Discovery in Databases*) y se encarga de extraer de manera automática información a partir de los datos recopilados. La DM se clasifica en aprendizaje supervisado o predictivo y no supervisado o descriptivo. Las técnicas no supervisadas se dividen a su vez en agrupamiento donde se encuentran los métodos numérico, conceptual y probabilístico, y asociación donde el método es a priori. Por otra parte, están los predictivos como: regresión, árboles de predicción y estimador de núcleos, y los de clasificación que pueden ser: tablas de decisión, árboles de decisión, inducción de reglas Bayesianas, Redes Neuronales (NN, por sus siglas en inglés de *Network Neural*), lógica difusa y técnicas genéticas [25].

2.1.5. Aprendizaje automático

Una de las técnicas que ha causado mayor impacto no solo en la medicina sino en diversas áreas, es el aprendizaje automático o aprendizaje máquina (ML, por sus siglas en inglés *Machine Learning*) es el proceso que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas [38]. El ML es una de las técnicas que trata de emular el aprendizaje humano, por lo que se dice que un programa de computación aprende de la experiencia E con respecto a una tarea T y alguna medida de rendimiento P, si es que el rendimiento en T medido por P, mejora la experiencia E [39]. Los diferentes algoritmos de ML se pueden agrupar en aprendizaje supervisado, aprendizaje no supervisado, aprendizaje semisupervisado, aprendizaje por refuerzo, transducción y aprendizaje multitarea [4].

2.1.6. Selección de características

El término selección de características se utiliza para hacer referencia a las herramientas o técnicas disponibles para reducir la dimensión de los datos a un tamaño apropiado para su tratamiento. Es una estrategia de búsqueda de una cantidad definida de atributos para el diseño del clasificador que reducen tiempo computacional y mejoran el aprendizaje del modelo [35].

2.1.7. Árbol de regresión y clasificación

El modelo de clasificación de nefropatía diabética se diseñó a partir de un árbol de regresión y clasificación (CART, por sus siglas en inglés *Classification and Regression Trees*), el cual utiliza datos históricos para realizar tareas de clasificación o predicción. El resultado de un CART es un árbol de decisión donde las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos históricos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos (Breiman, 1984) [34].

2.1.8. Matriz de confusión

La matriz de confusión permite observar el desempeño de un algoritmo donde cada columna de la matriz representa un número de predicciones de cada clase, mientras que las filas representan las instancias en la clase real como se muestra en la tabla 2 [33].

		Valor calculado	
		Positivo	Negativo
Valor real	Positivo	Positivos Ciertos	Negativos Falsos
	Negativo	Positivos Falsos	Negativos Ciertos

Tabla 2.- Matriz de confusión general.

2.1.9. Sensibilidad

La sensibilidad de una prueba es la proporción de los individuos clasificados como positivos que se identifican correctamente por la prueba en estudio. El valor que puede asumir la sensibilidad varia de 0-1 (100%), cuanto más alto es el valor, hay una mejor capacidad en la detección de enfermos [31].

2.1.10. Precisión

La precisión hace referencia a la proporción de la clasificación de casos positivos que fueron correctos, es decir la confiabilidad que se obtendrá en los resultados.

2.2. Estado del arte

En esta sección se dan a conocer algunos de los trabajos de investigación en el ámbito médico y social acerca de la nefropatía diabética, mientras que la siguiente categoría hace referencia a las soluciones computacionales que se han llevado a cabo para la predicción, tratamiento y monitoreo de la diabetes y algunas de sus complicaciones como son retinopatía, pie diabético, neuropatía y la misma nefropatía, posteriormente se muestran algunos sistemas de clasificación de diabetes y nefropatía, y por último un análisis de los trabajos mencionados.

2.2.1. Nefropatía diabética soluciones médicas

En [22], clasificaron las acciones preventivas en primarias, secundarias y terciarias, para las etapas de nefropatía incipiente, de incipiente a clínica y de incipiente a insuficiencia renal crónica, respectivamente. Los factores de riesgo de nefropatía diabética que determinan la predicción de la misma son los siguientes: a) antecedentes familiares de hipertensión arterial, b) antigüedad de la diabetes (mayor a 10 años), c) compensación metabólica habitual regular (Hemoglobina glicosilada (HbA1c) < 8%), d) presión arterial normal alta, e) tabaquismo y f) genotipo de predisposición a nefropatía diabética. Ruiz llegó a la conclusión de que a todos los pacientes de los países donde se realizaron dichos estudios se les aplicó insulinoterapia desde el diagnóstico de la diabetes.

Los autores en [10], mencionaron que la nefropatía diabética está caracterizada principalmente por proteinuria creciente, hipertensión arterial e insuficiencia renal. De los pacientes con diabetes mellitus tipo 1 y 2, solo 1/3 y 1/5 respectivamente, son los que desarrollan esta enfermedad. Los altos costos humanos, sociales y económicos, justifican su prevención y correctos tratamientos.

Dan a conocer cinco medidas terapéuticas: 1) modificar favorablemente el estilo de vida, además de evitar los factores de riesgo, cardiovascular y la educación del diabético y sus familiares, 2) control estricto y permanente de la glicemia, 3) control de la tensión arterial, 4) control de los trastornos lipídicos y 5) modificación del contenido proteico de la dieta.

La detección de microalbuminuria (MI) proporciona la posibilidad de evitar la progresión del daño renal, ya que es una señal temprana de nefropatía diabética. El grado de albuminuria se relaciona directamente con el avance del deterioro renal. En los pacientes diabéticos con diagnóstico de MI el tratamiento debe estar dirigido a disminuir la albuminuria o evitar que progrese y, además, a tomar las medidas necesarias para mantener la presión arterial, la glicemia y los lípidos en límites recomendados, para evitar complicaciones vasculares y renales a mediano y largo plazo [23].

Es importante detectar la prevalencia e incidencia de la patología y sus principales factores de riesgo. En el estudio realizado en [9], la población que se tomó en cuenta fueron 110 niños y adolescentes con diabetes mellitus tipo 1 que asisten a la consulta cada tres meses al hospital de endocrinología pediátrica, mediante la entrevista se recopilaron datos y una muestra de orina para determinar el grado de microalbuminuria en los pacientes. Los resultados que se obtuvieron es que la nefropatía diabética es más frecuente entre los 11 y 16 años (65%) seguido del grupo 17 y 20 años (35%).

2.2.2. Inteligencia artificial en la Diabetes y sus complicaciones

El objetivo de los autores en [12], es detectar la nefropatía en pacientes con diabetes tipo 2 a partir de la clasificación por género y mediante el uso de parámetros clínicos y genéticos. El modelo desarrollado en esta investigación utiliza las técnicas de F-score e information-gain para la selección de características mientras que para el clasificador emplean máquinas de vector soporte (SVM), bosques al azar (RF), Naive Bayes y árboles de decisión (DT). El resultado de este trabajo de investigación es la importancia de utilizar factores genéticos y clínicos para realizar el diagnóstico de la nefropatía diabética.

En [15], los autores buscan obtener y validar un conjunto de modelos para evaluar el riesgo de presentar complicaciones en pacientes diabéticos, utilizando 51 parámetros clínicos obtenidos en el Ensayo de Control de Diabetes y Complicaciones (DDCT, por sus siglas en inglés *Diabetes and Complication Control Trial*) como factores de riesgo potenciales. Los métodos que se utilizan para la selección de características son: selección de variables bayesianas (SVB), regresión Lasso Cox y selección directa y univariable, además de los algoritmos para el ajuste del modelo

que fueron la regresión Cox, Regresión ridge cox, tiempo acelerado de falla, bosque de supervivencia al azar (RSF) y regresión censurada de máquinas de vector soporte (SVMCR). Asimismo para medir el desempeño del modelo predictivo se utilizó la validación cruzada nested-cross y como resultados se obtuvieron una serie de modelos que evalúan el riesgo de desarrollar enfermedades que se derivan de la diabetes.

En el estudio [2], los autores diseñaron un método para predecir la detección de retinopatía diabética en una etapa temprana. Para el diseño del modelo se utilizaron técnicas de minería de datos donde los algoritmos clasificadores son Naive Bayes y SVM, y como método de selección de características el F-score. Ananthapadmanaban y Parthiban lograron un resultado de 83.37% en exactitud del clasificador usando el algoritmo Naive Bayes.

El objetivo de la investigación en [6], es encontrar un conjunto de características clínicas críticas mismas que ocasionan enfermedades cardiovasculares en pacientes con diabetes tipo 2. El modelo que se diseñó para la predicción de enfermedades cardiovasculares en pacientes diabéticos está compuesto por un algoritmo genético, el cual en su función fitness utiliza el clasificador del vecino más cercano y sus variantes. El conjunto de características que se encontraron como factores de riesgo potenciales para ocasionar este tipo de enfermedades son la edad al diagnóstico, duración de la diabetes, hemoglobina glicosilada, concentración de colesterol y hábitos de fumador. Teniendo como resultado del clasificador una exactitud de 96%.

Identificar los factores de riesgo cardiovascular que tengan mayor relación con la predicción de albuminuria en pacientes con diabetes tipo 2. Los autores de [17], propusieron una red neuronal con cuatro capas ocultas y un bias, además de un modelo de regresión logística condicional como herramienta para predecir el nivel de albuminuria en los diabéticos, las variables de entrada son sexo, duración de la diabetes, presión arterial diastólica y sistólica, lipoproteína de alta y baja densidad, triglicéridos, alta densidad en relación lipoproteína/triglicéridos, colesterol, glucemia en ayunas y hemoglobina glicosilada, sin embargo para la red neuronal se utilizó además de los anteriores la edad e IMC. Usando la regresión los factores más significativos del inicio de la albuminuria fueron el rango glomerular, tiempo de detección de la diabetes y sexo, mientras que en la red neuronal se encontró que lipoproteína de alta densidad es el indicador más importante para predecir la albuminuria en pacientes diabéticos.

En [5], buscan predecir el inicio de la nefropatía diabética teniendo un conjunto de datos de pacientes diabéticos irregular y desequilibrado donde las técnicas que se emplean para la

selección de características es ReliefF y un análisis de sensibilidad y para clasificar un SVM con 39 atributos seleccionadas obteniendo un 96.9% de exactitud en su clasificador.

Encontrar un diagnóstico de nefropatía diabética para poder proporcionar un tratamiento en tiempo adecuado. El conjunto de datos que se utilizan son clínicos y genotipados (genéticos), tomando en cuenta los datos anteriores se realizó una clasificación basada en cluster que contiene las técnicas de K-means y el árbol de decisión C4.5, por otra parte para medir el desempeño de la clasificación se utilizó validación cruzada con 10-folds. En [13], encontraron que la inclusión de características genéticas produce mejores resultados al realizar la predicción de la nefropatía en diabéticos.

El principal objetivo del estudio realizado en [12], es desarrollar un modelo de evaluación de riesgos haciendo una combinación de datos clínicos y genéticos. Para la elaboración del modelo se utilizaron una combinación de métodos dentro de los cuales se menciona el método de selección de características F-score y el clasificador SVM, y el método information-gain con el árbol de decisión C4.5 por mencionar algunos. El desempeño de la tarea de clasificación se midió con un enfoque de validación cruzada utilizando 5-folds, por otra parte se tienen las herramientas utilizadas para el desarrollo del modelo que son LibSVM y Weka. Los resultados que se obtuvieron son que el árbol de decisión y el random forest pueden superar a otros clasificadores en costo y eficiencia.

En [26], se propone un modelo para la predicción de nefropatía incipiente por contraste antes de llegar a los procedimientos radiológicos entre los pacientes tratados usando medios de contraste, el modelo consiste en la implementación de algoritmos de aprendizaje automático como son el random forest y árboles de decisión implementados en el lenguaje de programación R. Los autores mencionaron que el método random forest se utiliza para describir la relación entre variables dependientes e independientes obteniendo buenos resultados con alta flexibilidad y exactitud. En este caso el clasificador obtuvo 80.8% de exactitud al realizar las operaciones.

Es importante encontrar el mejor método para la predicción de alguna enfermedad, sin embargo en los resultados que se obtienen influye en gran medida los parámetros utilizados y la técnica. Es por ello que en [16], compararon el desempeño de algunos de los métodos de aprendizaje automático dentro de la generación de modelos para la predicción de enfermedad renal diabética. Los algoritmos que se usaron para el entrenamiento, validación y prueba del modelo para predecir la enfermedad renal en los diabéticos son regresión parcial de mínimos cuadrados, árboles de regresión y clasificación, árbol de decisión C4.5, random forest, clasificador de Naive Bayes,

redes neuronales y SVM. A partir de los cuales se obtuvo que mediante el SVM y random forest se pueden construir modelos de alto rendimiento.

Las personas que padecen diabetes anualmente invierten grandes cantidades de dinero en discapacidades causadas por la diabetes, es por ello que es necesario estudiar la relación que existe entre las complicaciones causadas por la diabetes. Para llevar a cabo el modelo se realizó una combinación de la regresión logística y las redes neuronales artificiales en el entorno Matlab [21].

En [30], desarrollan y validan un árbol de decisión basado en los perfiles para predecir el riesgo de nefropatía diabética de antemano por albuminuria con el objetivo de examinar los perfiles proteómicos séricos asociados con el posterior desarrollo de nefropatía en pacientes diabéticos tipo 2. Se utilizó la espectrometría de masa de tiempo de vuelo de desorción / ionización con láser de superficie mejorada para obtener los perfiles proteómicos de muestras de suero de referencia de 84 pacientes con diabetes tipo 2 con albuminuria normal. Un árbol de decisión discriminatorio óptimo para los sujetos de caso creados con cuatro nodos que usan cuatro masas distintas fue cuestionado con el conjunto de pruebas. El valor predictivo positivo fue del 77,8% (7/9), y el valor predictivo negativo fue del 72,7% (8/11).

Las complicaciones de la diabetes afectan la calidad de vida de las personas y, sin el tratamiento adecuado, pueden causar la muerte. Es por ello que los autores en [27] tienen como objetivo desarrollar una aplicación Web para predecir la aparición de complicaciones microvasculares en diabéticos tipo 2, por tanto se utiliza la tecnología JFS (JavaServer Faces) que implementa un modelo arquitectónico Modelo-Vista-controlador; la etapa controlador contiene el FacesServlet de JFS que se encarga de coordinar a la vista y al modelo en el flujo de navegación, la vista, es decir, la interfaz de usuario se genera mediante XHTML y el modelo está compuesto por las clases importante en el dominio del problema y hacen uso de Weka (plataforma de minería de datos). En esta investigación la población está conformada por pacientes con diagnóstico de DMT2, en cualquier etapa de la enfermedad, con y sin presencia de RD o ND o PD. El tamaño de la muestra es de 200 pacientes con DMT2 que acudieron a consulta externa, cirugía o urgencias en el Hospital Regional de Rio Blanco en el periodo 2012-2014. Las variables que representan presencia de alguna complicación son sexo, edad, años de evolución, colesterol, insulina, triglicéridos, glucosa, hemoglobina, leucocitos y proteína por mencionar algunos, sin embargo el conjunto consta de 53 variables. Los algoritmos utilizados para la predicción son Bayes Ingenuo, árbol de decisión (ID3 y C4.5), vecino más cercano, perceptrón multicapa y máquinas de vector soporte.

2.2.3. Sistemas de clasificación de diabetes y nefropatía.

Narasimhan y Malathi clasificaron riesgos de nefropatía en mujeres con diabetes mediante la técnica de lógica difusa que contiene cómputo suave. Los parámetros de entrada para el sistema difuso son concentración de glucosa, presión arterial diastólica, IMC y edad, es importante mencionar que el sistema se realizó con Mandami, funciones de membresía triangulares, centro de gravedad para la defuzzificación y selección de clasificación de atributos (ART). En esta investigación se utilizaron la sensibilidad, exactitud y especificidad para medir el desempeño del clasificador, mismos que arrojaron los siguientes resultados 99.61%, 98.88% y 71.42% respectivamente [18].

La diabetes es una enfermedad que va aumentado progresivamente y tiene una alta probabilidad de muerte en los pacientes de este grupo. En [24], realizaron una clasificación para tratamiento y diagnóstico de la diabetes mellitus usando un algoritmo híbrido compuesto por optimización de enjambre de partículas modificadas (MPSO, por sus siglas en inglés *Modified Particle Swarm Optimization*) y mínimos cuadrados de máquinas de vector soporte (SL-SVM, por sus siglas en inglés *Least Squares Support Vector Machine*). El resultado que se obtuvo referente al clasificador fue de 97.833% de exactitud.

En [3], los autores buscaron demostrar la utilidad de las técnicas de clasificación para predecir la enfermedad de riñón con diferentes herramientas de minería de datos. Para lograr lo anterior mencionan un panorama sobre los síntomas y factores de riesgo que pueden ocasionar alguna de las enfermedades en los riñones, además de las enfermedades que sufren los riñones principalmente. Por otra parte, muestran varias técnicas de minería de datos que se han aplicado en la predicción y diagnóstico de varias enfermedades renales.

2.2.4. Análisis de trabajos relacionados

Desde hace varios años hay autores trabajando en desarrollar una solución computacional al problema de predicción de la diabetes y algunas de sus complicaciones. En la revisión del estado del arte, se encontró que los trabajos realizados coinciden en algunas de las técnicas y/o herramientas utilizadas, es por ello que se hace una comparación con el modelo de clasificación propuesto en esta investigación para dar a conocer los resultados obtenidos.

NA.- Numero de atributos

TM.- Tamaño de la muestra.

TA.- Tipo de atributos

MSC.- Método de selección de características.

TC.- Técnica de clasificación.

RES.- Resultados.

OBJ.- Objetivo.

H.- Herramientas.

TD.- Tipo de diabetes.

En la revisión de la literatura respecto al estado del arte relacionado con el modelo de clasificación de nefropatía diabética mediante aprendizaje automático se encontró que, el método más utilizado para la selección de características es el F-Score o Information Gain, la técnica de clasificación que predomina es el Árbol de decisión en cualquiera de sus derivados y la Máquina de Vector Soporte y el desempeño que presentan los trabajos relacionados va de 50% a 100% de sensibilidad en las tareas de clasificación y/o predicción (Tabla 3).

Cabe mencionar que algunos de los trabajos analizados realizan tareas de clasificación binaria, es decir, predicen enfermedades como la diabetes o sus complicaciones pero solo determinan si el individuo padece o no la enfermedad. Sin embargo, la solución propuesta realiza la clasificación de etiquetas multiclase que ayuda a determinar el nivel de afectación del paciente.

El modelo de clasificación de nefropatía diabética propuesto en este trabajo de tesis se muestra en la última posición de la tabla 3, el cual utiliza el método chi2 para la selección de atributos, CART para la tarea de clasificación y como resultado obtiene aproximadamente 90% de sensibilidad al realizar la clasificación que define el nivel de afectación renal que padece el paciente en estudio.

	Elementos a considerar								
Autor	**NA**	**TM**	**TA**	**MSC**	**TC**	**RES**	**OBJ**	**H**	**TD**
Huang 2015a	32	345	Clínicos. Genéticos.	F-score. Information gain.	SVM. RF. NB. DT.	Exactitud 61%-66%	Nefropatía.	LibSVM. Java. Weka.	2
Vicenzo 2015	51	1441	Clínicos.	Selección de variables Bayesianas. Regresión Lasso Cox. Selección directa y univariable.	Regresión Cox. Regresión ridge cox. Tiempo acelerado de falla. Bosque de supervivencia al azar. Regresión censurada de máquinas de vector soporte.	Conjunto de modelos computacionales.	Predicción de Enfermedades cardiovasculares.		1 2
Ananthapadmanaban 2014	16	300	Médicos.	F-score.	NB. SVM.	Exactitud 64%-84% Sensibilidad 96.65% Especificidad 95%	Retinopatía.	RapidMiner.	2
Narasimhan 2014	8	768	Clínicos	Técnica de clasificación de atributos.	Lógica difusa.	Sensibilidad 99.61%. Exactitud 98.88%. Especificidad 71.42%.	Nefropatía.	Matlab.	2
Dalakleidi 2014	32	560	Clínicos.		GA. Función fitness KNN.	Exactitud 96%.	Cardiovasculares.		2

Soliman 2014	8	768	Clínicos.		MODIFIED-PSO. LS-SVM	Exactitud 97.833 %	Tratamiento y diagnóstico de la diabetes.		2
Morteza 2013	13	1104	Clínicos.		NN. Modelo de regresión logística.	HDL-colesterol	Predecir nivel de albuminuria.		2
Cho 2007	184	292	Clínicos.	ReliefF. Análisis de sensibilidad.	SVM	Exactitud 96.9%.	Inicio nefropatía.		2
Huang	17	345	Clínicos. Demográficos. Genéticos.		K-means. DT C4.5	Exactitud 64%-85%. Especificidad 34%-95%. Sensibilidad 61%-88%.	Diagnóstico de nefropatía.	Java. Weka.	2
Huang 2015b	33	527	Clínicos. Genéticos.	F-Score Information Gain	SVM. DT C4.5. NB. RF.	Exactitud 70%-79%. Especificidad 73%-86%. Sensibilidad 53%-83%.	Evaluación de riesgos.	LibSVM. Weka.	
Yin 2017	83	8800	Clínicos. Demográficos.		RF. DT.	Exactitud 80.8% Sensibilidad 82.7% Especificidad 78.8%	Predicción de nefropatía incipiente.	R.	
Leung 2013	87	673	Clínicos. Genéticos.	SVMRadial Cforest	Regresion parcial de mínimos cuadrados. CART. DT C4.5. RF	Exactitud 70%-95%	Predicción renal diabética.	SPSS Statistics 17.0. R.	2

					NB. NN. SVM.				
Parasto 2016	8	180	Clínicos.		Regresión logística. NN.	La función de error de la red neuronal es 0.1 y el modelo de red neuronal híbrida es igual a 0.0002.	Predicción de diabetes.	Matlab.	1 2
Yang 2014	18	84	Clínicos.	Espectrometría de masa de tiempo de vuelo de desorción / ionización con láser.	DT.	Valor predictivo positivo 77.8%. Valor predictivo negativo 72.7%	Predecir nefropatía.	SELDI.	2
Sahir 2014	53	200			NB. DT ID3 y C4.5. KNN. Perceptrón multicapa. SVM.		Predicción de nefropatía, retinopatía y pie diabético.	Java. Weka.	2
Francis 2017	27	55	Clínicos.	Chi2	CART	Sensibilidad-90%	Predicción nefropatía	Python. Weka. Matlab.	2

Tabla 3.- Comparación de trabajos relacionados y modelo propuesto.

Capítulo III

Descripción del modelo de clasificación de nefropatía diabética.

En el presente trabajo se propone un modelo de clasificación de nefropatía diabética, para que los médicos sin experiencia en esta área puedan clasificar en etapas adecuadas el daño renal de los pacientes con diabetes mellitus tipo 2. La imagen 3 se trata de un gráfico enriquecido de tipo Rich Picture que proyecta las fases necesarias para llevar a cabo el modelo.

La recolección de datos consiste en definir y establecer un conjunto de factores de riesgo que determinan la nefropatía diabética, el cual se elabora a partir del conocimiento adquirido de la literatura y la ayuda del experto. Debido a que los datos que se van a extraer son de expedientes físicos es necesario realizar el almacenamiento en un dataset digital. Una vez que se tiene el dataset se procede a la elaboración de los subconjuntos de entrenamiento y prueba. La siguiente fase se encarga de limpiar y transformar los datos para eliminar valores atípicos y evitar redundancia de los mismos. En la selección de características se obtiene un subconjunto, es decir, se reducen o eliminan los atributos irrelevantes que permiten aumentar la precisión del modelo. Así mismo, los subconjuntos de entrenamiento y prueba preprocesados se utilizan como entrada para el clasificador, el cual tiene como salida la clasificación del subconjunto de pruebas.

Para finalizar el proceso se mide el desempeño del clasificador mediante pruebas estadísticas y la matriz de confusión.

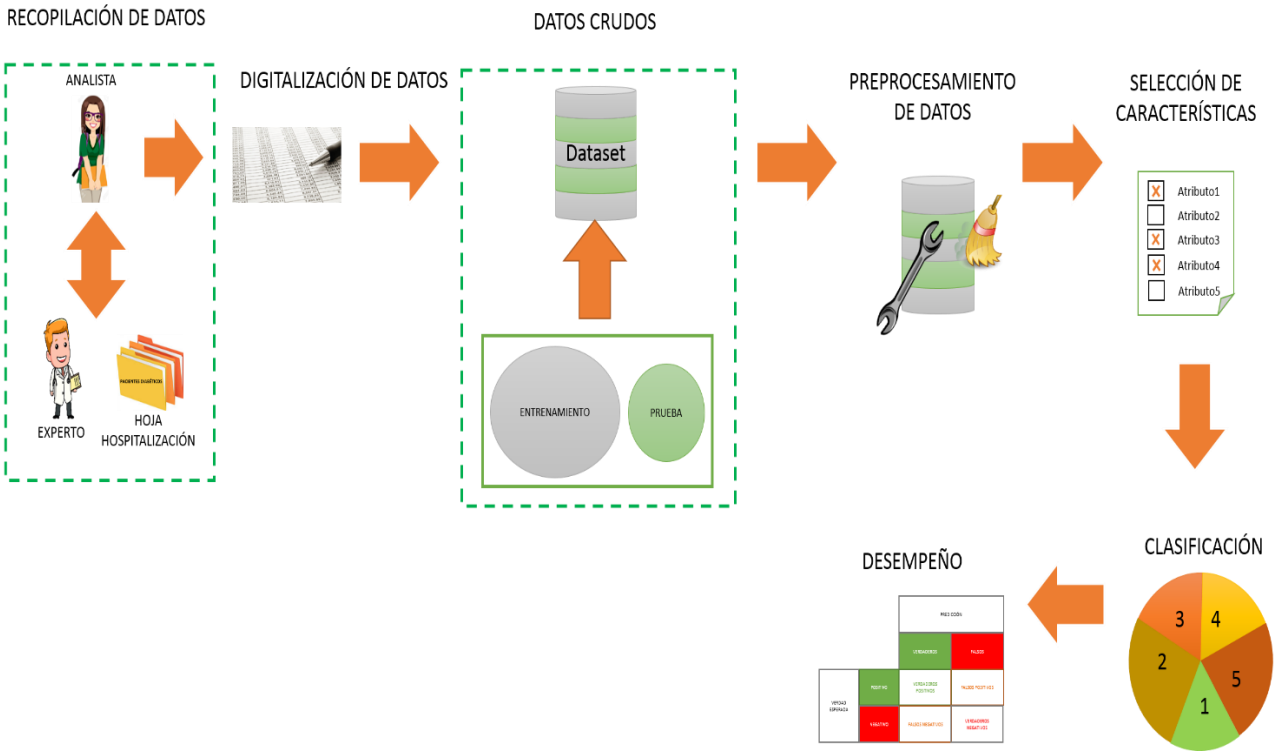

Imagen 3.- Etapas del modelo de clasificación de nefropatía diabética.

3.1. Elaboración del modelo

Para el diseño y desarrollo del modelo de clasificación de nefropatía diabética se emplearon los siguientes elementos de apoyo: un conjunto de datos extraídos de los expedientes médicos del Hospital General de Misantla obtenidos mediante un estudio retrospectivo del periodo 2015-2016, las guías de prácticas clínicas KDIGO para el manejo de la enfermedad renal crónica, y la ecuación para calcular el filtrado glomerular que determina el nivel del daño renal y por ende el estadio al cual pertenece el paciente.

3.1.1. Sujetos de estudio

En la presente investigación participaron 55 pacientes con diagnóstico nefrótico y diabetes mellitus 2, 32 de los cuales son mujeres y 23 hombres que asisten a consulta al Hospital General de Misantla (HGM) en Misantla, Veracruz, México en el periodo 2015-2016.

3.1.2. Formula de filtrado glomerular

Debido a que la institución de salud no realiza ninguna prueba de filtrado glomerular es necesario a partir de los datos y pruebas que se tienen encontrar este valor, mismo que según la clasificación que se va utilizar es el que determina el estadio al cual pertenece el paciente. La ecuación 1 es la que se utiliza para medir la filtración glomerular:

$$FG = 175 * Cr^{-1.154} * e^{-0.203} * (si\ es\ f\ (0.742))$$

Ecuación 1.- Filtrado Glomerular.

Donde:

Cr: Creatinina.

e: edad del paciente.

f: femenino (sexo).

3.1.3. Conjunto de datos

El conjunto de datos que se recopiló contiene cinco clases y 55 instancias, donde cada clase hace referencia a un estadio de la clasificación KDIGO para medir el daño renal en las personas. Cada instancia tiene 26 atributos o características, además de contener anomalías. Los cuales son de tipo numérico y texto, y el atributo que se utilizará como predictor es el estadio que mide el daño renal (Tabla 4).

No. Atributo	Nombre	Descripción	Tipo	Valor
1	no_Exp	Número de expediente del paciente	NUMERICO	
2	Edad	Edad del paciente	NUMERICO	
3	Sexo	Sexo del paciente	TEXT	F/M
4	Talla	Estatura del paciente	NUMERICO	
5	Peso	Peso del paciente	NUMERICO	
6	IMC	Índice de masa corporal del paciente	NUMERICO	
7	CM	Otras enfermedades además de diabetes	TEXTO	SI/NO
8	terapéutica	Cuenta con algún tratamiento	TEXTO	SI/NO
9	HbA1c	Hemoglobina glicosilada del paciente	NUMERICO	
10	glucosa	Glucosa del paciente	NUMERICO	
11	Urea	Urea del paciente	NUMERICO	

12	Cr	Creatinina del paciente	NUMERICO	
13	BUN	Nitrógeno ureico del paciente	NUMERICO	
14	URIC	Ácido úrico del paciente	NUMERICO	
15	CHOL	Colesterol del paciente	NUMERICO	
16	TRIG	Triglicéridos del paciente	NUMERICO	
17	prot_tot	Proteínas totales del paciente	NUMERICO	
18	albumina	Albumina del paciente	NUMERICO	
19	proteinuria	Proteinuria del paciente	NUMERICO	
20	dep_cr	Depuración de creatinina del paciente	NUMERICO	
21	eritrocitos	Eritrocitos del paciente	NUMERICO	
22	HGB	Hemoglobina del paciente	NUMERICO	
23	HCT	Hematocrito del paciente	NUMERICO	
24	tipo	Tipo de diabetes del paciente	NUMERICO	
25	evolución	Años de que el paciente es diabético	NUMERICO	
26	FC	Filtrado glomerular del paciente	NUMERICO	
Clase	estadio	Daño renal de paciente	TEXTO	G1/G2/G3/G4/G5

Tabla 4.- Atributos clínicos usados en este estudio.

Cabe mencionar que el conjunto de datos presenta un desbalance, el desbalance en los datos hace referencia a que el número de observaciones o instancias no es la misma para todas las clases.

3.1.3.1. Estructura del dataset

Para una mejor manipulación y organización de los datos se almacenaron en un archivo separado por comas (.csv). Los archivos en formato CSV son documentos en formato abierto utilizados comúnmente para la representación de datos en forma de tabla, en este formato las columnas se separan por una coma (,) y las filas por un salto de línea.

En la imagen 4 se muestra un segmento del archivo .csv que contiene el conjunto de datos, en el cual cada línea representa un paciente con sus respectivos valores numéricos y cadenas según sea el atributo al que se hace referencia, también se pueden observar las anomalías que dicho dataset presenta.

```
1  1357101601,58,"F",1.54,60,25.3,"si","si",,461,220.4,18.36,99.2,4.4,172,242,,1.8,200,,2.81,8.1,23.
   7,2,14,1.981,"5"
2  568071901,46,"F",1.47,55,25.45,"si",,7.1,244,24,0.7,,,178,133,,,150,,4.5,,,2,8,90.086,"1"
3  568073001,48,"M",,76,,"si","si",,102,406.6,18.82,21,8.4,142,437,8.1,4,30,19.09,2.69,6.6,21.9,2,,2.
   697,"5"
4  557021201,59,"M",1.54,77.5,32.68,"si","si",,320.8,40.4,1.71,18.9,,207,164,,,500,50.76,4.25,12.
   7,41,2,3,41.179,"3"
5  1634051701,83,"F",,,,"si","si",,309.4,224.7,9.66,105,,197.2,250,,,100,,3.46,9.8,31,2,21,3.865,"5"
6  839101501,76,"M",,,,"si",no,,106,250,11.05,,,,,,,,2.79,8.4,25.1,2,,4.541,"5"
7  1667041101,48,"F",,,,"si",no,,770,47.7,1.08,22.3,,93.4,150.1,,,30,,2.61,7.7,22.7,2,8,54.148,"3"
8  1666021301,51,"M",1.65,95,34.89,"si",no,,71,133.1,1.2,62.2,,,,,1.75,,,2.47,7.8,23.9,2,,63.83,"2"
9  645101701,70,"F",1.48,66,30.13,"si","si",6.8,190,25,0.9,8.1,,,,,,,64.8,3.6,9.9,28.1,2,7,61.9,"2"
10 965122501,50,"M",,62,,"si","si",,108,234.1,19.67,109.4,6.6,253,139,6.8,2,200,38.65,2.1,5.6,17.5,,15,2.
   542,"5"
11 855941801,61,"M",1.54,65,27.41,"si","si",,1000,38,0.86,,,,,,,500,,3.76,10.5,31.9,2,3,90.408,"1"
12 835061301,80,"F",,,,"si","si",,49,31.24,0.47,14.6,5.8,88,93,,,,,3.59,8.5,29.2,2,20,127.501,"1"
13 737081802,80,"M",1.52,74.5,32.25,"si","si",,130,114,3.2,54.1,9.1,110,69,,2.6,100,,2.77,7.7,24.4,2,3,18.
   783,"4"
14 657050901,59,"M",,,,"si","si",,240,235,11,,7.7,109,94,,,300,,2.84,7.4,21.5,2,18,4.806,"5"
15 1355080101,63,"M",1.6,54,21.09,"si","si",,204,45,1.2,18.5,9.1,87,116,,,,26.38,3.84,10.4,34.4,2,20,61.
   15,"2"
16 1368080701,47,"F",,,,"si","si",7.2,113,155,9.5,62.8,7.9,209,225,,3,300,8.5,2.75,7.3,22.8,2,15,4.423,"5"
17 1254031301,62,"F",,51,,"si","si",,175,47,0.5,22,6.1,209.9,149.2,,,30,,5.06,13.9,42.1,2,10,125.019,"1"
18 1638092001,78,"F",,,,"si","si",7.2,109,37,1,17.3,,,,,,,,3.7,11.2,34,2,16,53.622,"3"
19 1636102001,79,"F",,,,"si","si",,395,54,1.6,16,,,,,,30,24.81,3.96,10.4,32,2,15,31.093,"3"
20 1571092501,44,"F",,,,"si","si",,127,97.37,3.19,45.5,,116,215,,,100,20.76,3.05,7.8,26,2,3,15.792,"4"
21 1570082501,45,"M",1.57,68.5,27.79,"si","si",8.3,139,136.75,7.1,63.9,9.8,203,121,,,500,33,2.51,7.5,23.
```

Imagen 4.- Conjunto de datos en formato .csv.

3.1.3.2. Conjunto de entrenamiento y prueba

Realizar la separación del conjunto de datos en los subconjuntos de entrenamiento y prueba es una tarea importante para encontrar una función o modelo que mejor generalice los datos y a partir de la misma se clasifique los valores de prueba, regularmente el conjunto de entrenamiento contiene el mayor número de datos posible y el conjunto de prueba contiene una fracción menor del conjunto de datos. Ambos conjuntos como su nombre lo indica se utilizan para la fase de aprendizaje y prueba del modelo.

Para determinar qué elementos formarán parte de cada uno de los conjuntos se ha establecido que para el conjunto de entrenamiento se utilizó 70% del conjunto original que corresponde a 38.5 registros y el 30% restante representa el conjunto de prueba que corresponde a 16.5 registros, teniendo un total de 55 registros en el conjunto de datos (Imagen 5).

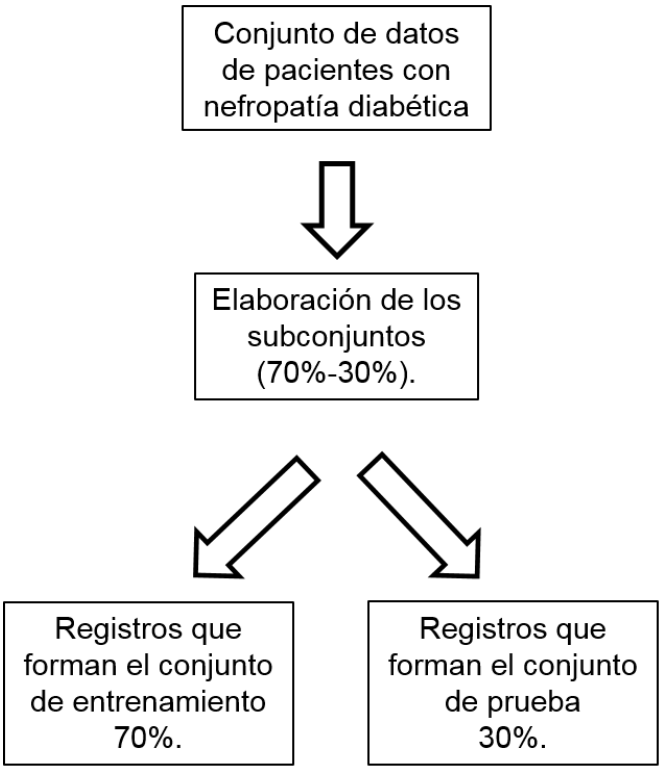

Imagen 5.- Selección de los elementos para los conjuntos de entrenamiento y prueba.

3.1.4. Preprocesamiento de los datos.

Los datos obtenidos mediante el estudio retrospectivo del periodo 2015-2016 son los que conforman el conjunto de datos a utilizar para la elaboración del modelo, sin embargo presentan ciertas anomalías y es por ello que se debe realizar el preprocesamiento ya que esto puede mejorar el desempeño del modelo.

Un factor de riesgo importante para el diagnóstico de nefropatía diabética es la hemoglobina glicosilada (HbA1c), sin embargo al recolectar los datos se encontró que este dato falta en más del 50% de los registros, debido a que es una prueba que no se practica en la institución de salud y no todo los pacientes tienen las posibilidades para realizarlo en un laboratorio particular, por lo tanto se tuvo que omitir del conjunto de datos al igual que la columna prot_tot (proteínas totales), albumina y dep_cr (depuración de creatinina).

El dataset presenta escasez de datos en sus registros, por lo tanto es necesario aplicar una técnica que verifique el tipo de variable (discreta o continua) correspondiente a dicha columna y realice la operación necesaria para llenar datos faltantes.

Para dar solución a este problema se utilizan las medidas de tendencia central como son media y moda, donde la media se utiliza cuando la variable es de tipo continuo y la moda cuando es discreta. La representación matemática de la media está dada por la ecuación 2 y la moda hace referencia al valor que mayor número de veces se repite en el conjunto de datos.

$$\bar{x} = \frac{suma\ de\ todos\ los\ valores}{cantidad\ total\ de\ datos} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{N}$$

Ecuación 2.- Media Aritmética.

Otra tarea necesaria sobre el conjunto de datos consiste en la limpieza de estos, debido a que existen atributos que no aportan información relevante que ayude a mejorar el desempeño del modelo que se está buscando, este problema se solucionó omitiendo atributos que representan el número de expediente, tipo de diabetes, co-morbilidad (otras enfermedades además de la diabetes) y terapéutica (fármaco proporcionado para este padecimiento).

3.1.5. Selección de características

En esta fase se realizó un análisis de la información encontrada en la literatura acerca de las causas y factores de riesgo que ocasionan en las personas diabéticas la nefropatía, y con ayuda de un profesional de salud en esta enfermedad se estableció un conjunto de datos clínicos considerados importantes y determinantes con la aparición de la enfermedad, lo anterior debido a que no existe un conjunto de factores de riesgo establecidos para predecir nefropatía diabética.

Para dar solución a lo planteado se aplicó el algoritmo chi-cuadrada (X^2) que consta de una función evaluador tipo filtro, es decir, la función criterio es independiente del algoritmo de aprendizaje y utiliza medidas de dependencia para medir la correlación entre los atributos y el resultado es un subconjunto de características. La prueba chi2 es una prueba estadística que se utiliza para probar la independencia de dos eventos, es decir, los dos eventos son la aparición de la característica y la clase; si los dos eventos son dependientes se puede usar la ocurrencia de la característica para predecir la ocurrencia de la clase [36].

En la imagen 6 se puede observar el proceso que se realiza al aplicar un algoritmo de selección de características.

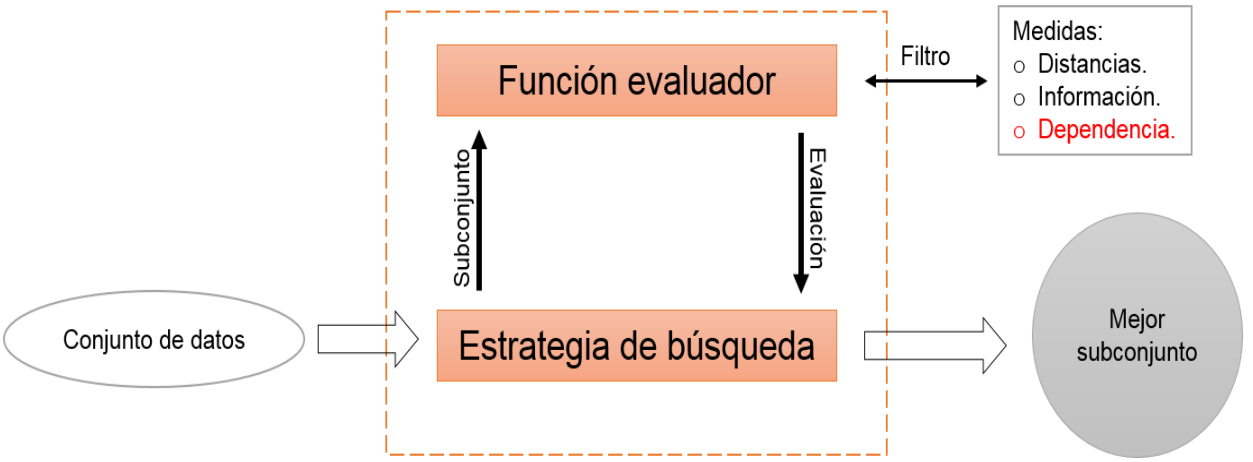

Imagen 6.- Proceso de selección de características.

Una vez que se aplicó el algoritmo de selección de atributos se obtuvieron los atributos *Glucosa, urea, cr, BUN, CHOL, TRIG, prot_tot, proteinuria, dep_cr y FG* como elementos del subconjunto de características que se utilizará para alimentar el clasificador.

3.1.6. Clasificador.

Un clasificador es un algoritmo que pertenece a la categoría de algoritmos de aprendizaje supervisado ya que contiene un atributo clase que define cada instancia, es decir, se tiene un conocimiento a priori. Los problemas de clasificación pueden ser binarios o multiclase y su tarea consiste en asignar a una clase conocida una instancia no etiquetada.

Las máquinas de soporte vectorial (SVM, por sus siglas en inglés *Support Vector Machine*), redes neuronales, árboles de decisión y Naive Bayes por mencionar algunos son algoritmos clasificadores.

3.1.7. Árbol de clasificación y regresión.

Los árboles tipo CART pretenden explicar y/o predecir una variable respuesta a partir de un conjunto de variables predictoras mediante un conjunto de reglas sencillas.

Para llevar a cabo la correcta aplicación del árbol tipo CART al conjunto de datos es necesario tener en cuenta que el problema que se está trabajando pertenece a una predicción multiclase, por lo que la variable respuesta puede tomar un valor de entre estos cinco G1, G2, G3, G4 y G5, además de ser un conjunto de datos donde sus clases no se encuentran balanceadas.

3.1.8. Ajustes del clasificador CART.

El objetivo del algoritmo CART es particionar repetidamente los datos para formar el árbol y para realizar la división de los datos se requiere un "criterio de particionamiento" el cual lo determinará la medida de impureza.

3.1.8.1. Función de impureza.

La función de impureza es una medida que permite determinar la calidad de un nodo, esta será denotada por i (t). En este caso se eligió el índice Gini que tiende a separar la categoría más grande en un grupo aparte.

La representación matemática del índice Gini se muestra en la ecuación 3.

$$i(t) = \sum_{i \neq j} p(j|t)p(i|t)$$

Ecuación 3.- Índice Gini.

Encontrar la partición que maximice ∆i (t) en ecuación 4.

$$\Delta \mathrm{i} = -\sum_{j=1}^{k} [p_j(t)]^2$$

Ecuación 4.- Maximizar ∆i (t).

3.1.8.2. Peso de las clases.

Debido a que el conjunto de datos presenta un desbalance en las clases es necesario especificar al algoritmo CART el peso que se asigna a cada una de las mismas. Para determinar el peso de las clases se utiliza el parámetro "balanceado" que mediante los valores de las etiquetas ajusta automáticamente los pesos inversamente proporcionales a las frecuencias de clase en los datos de entrada. La ecuación 5 muestra el proceso matemático que se realiza para determinar el peso para cada una de las clases que se tienen en el conjunto de datos.

$$\mathrm{n\_samples} \,/ \left(\mathrm{n\_classes} * \mathrm{np.bincount(y)}\right)$$

Ecuación 5.- Balance de las clases.

Donde:

n_samples: número de instancias en el conjunto de datos.

n_classes: número de clases en el conjunto de datos.

np.bincount(y): es una función perteneciente al lenguaje Python que obtiene el número de ocurrencia de cada una de las clases, usando como parámetro el arreglo que contiene las etiquetas de las clases.

Después de haber realizado las configuraciones necesarias al algoritmo, el siguiente paso consiste en la fase de aprendizaje y en la imagen 7 se muestra el proceso de aprendizaje del algoritmo clasificador.

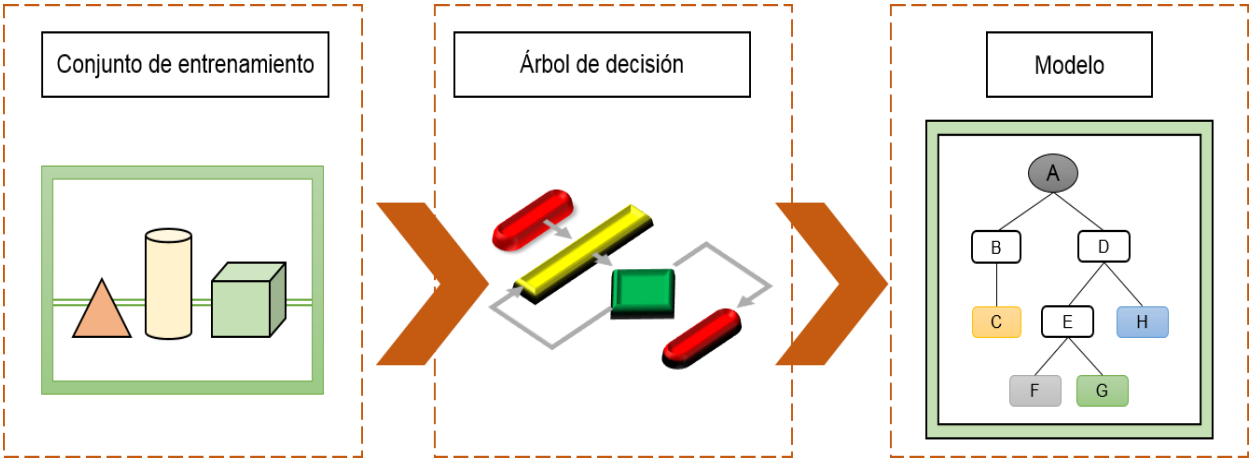

Imagen 7.- Proceso de aprendizaje del clasificador.

El proceso de aprendizaje consiste en alimentar el clasificador, en este caso el árbol, a partir del conjunto de entrenamiento para obtener una serie de reglas que determinen el conjunto de patrones o modelo generalizado del conjunto de datos que se han encontrado. Los patrones encontrados mediante el proceso de aprendizaje se utilizan para realizar pruebas y observar el comportamiento del modelo.

El modelo generalizado del conjunto de datos de nefropatía diabética obtenido a partir del árbol tipo CART se puede observar en la imagen 8.

La raíz del árbol corresponde a la variable FG con un valor menor o igual a 81.0365 si la condición se cumple se crea un nuevo nodo y en caso contrario la hoja determina estadio 1, el nuevo nodo corresponde a la variable Cr con un valor menor o igual a 3.65 si dicha condición es verdadera

se crea un nuevo nodo y en caso contrario la hoja indica estadio 5, el nuevo nodo contiene la variable FG y un valor menor o igual a 58.7675 misma que si cumple genera un nuevo nodo y en caso contrario se clasifica en estadio 2, al nuevo nodo corresponde FG con un valor menor o igual a 30.098 si se cumple la hoja del árbol indica que el estadio es 3 y en caso contrario 4.

Las reglas que corresponden al árbol son las siguientes:

- Si FG <= 81.0365 entonces
    - Si Cr <= 3.65 entonces
        - Si FG <= 58.7675 entonces
            - Si FG <= 30.098 entonces
                - *estadio* = 3
            - de lo contrario *estadio* = 4
        - de lo contrario *estadio* = 2
    - de lo contrario *estadio* = 5
- de lo contrario *estadio* = 1

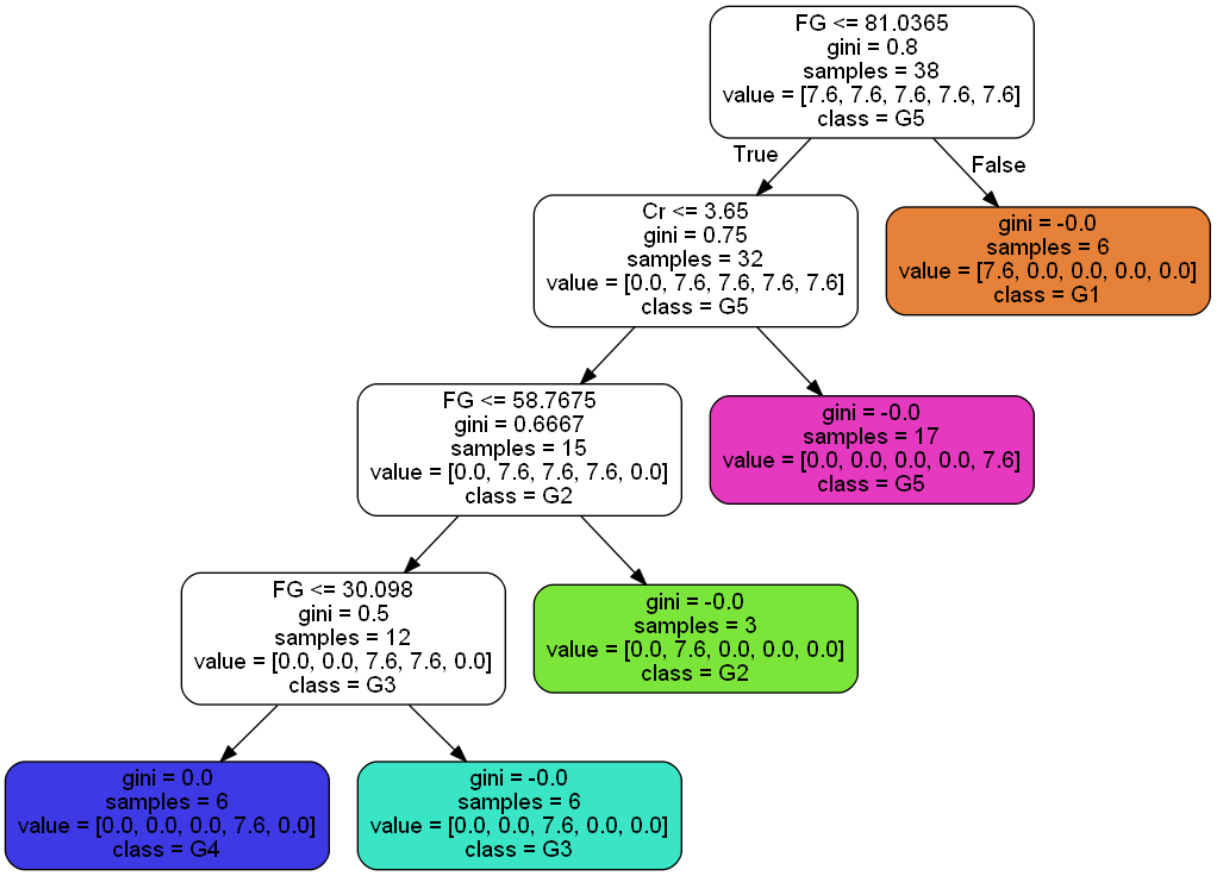

Imagen 8.- Modelo generalizado del conjunto de datos.

Capítulo IV

Análisis de resultados

Para medir el rendimiento del modelo se llevaron a cabo experimentos que permiten observar su comportamiento al realizar la tarea de clasificación.

Mediante la plataforma WEKA se aplicaron los clasificadores Bayes Net, Simple Logistic, LWL, Filtered Classifier VFI, Decision Table, DTNB y LADTree por mencionar los de mayor precisión. Sin embargo, debido a la falta de información en el conjunto de datos se alteró el comportamiento de los clasificadores obteniendo valores de exactitud entre 70% y 95% (Tabla 5).

CATEGORÍA	ALGORITMO	Instancias correctas	% correcto	% incorrecto
bayes	BayesNet	47	85.45	14.55
function	SimpleLogistic	39	70.91	29.09
meta	FilteredClassifier	53	96.36	3.64
misc	VFI	49	89.09	10.91
rules	DecisionTable	53	96.36	3.64
	DTNB	53	96.36	3.64
trees	LADTree	52	94.55	5.45

Tabla 5.- Algoritmos clasificadores implementados en WEKA.

Enseguida se desarrolló una máquina de soporte vectorial utilizando una función de base radial y una función lineal, mismas que sirvieron para determinar y encontrar que el conjunto de datos no es linealmente separable como se muestra en la imagen 9, ya que el porcentaje de error al clasificar superaba el 50%.

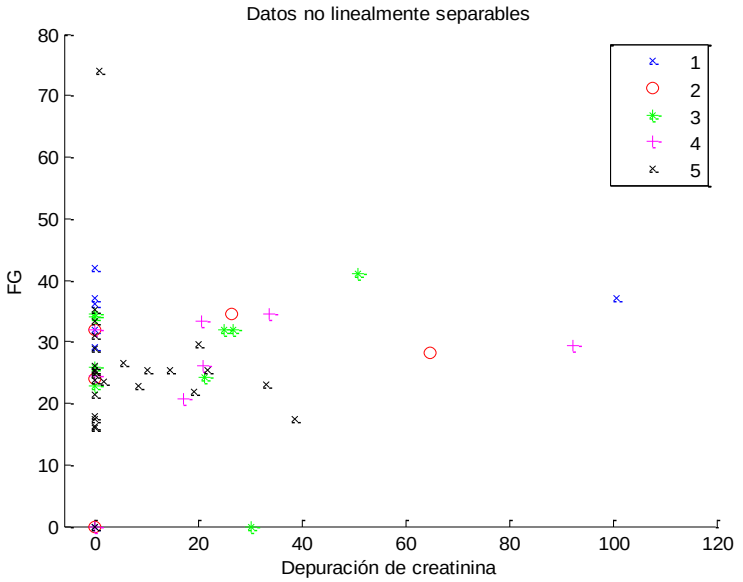

Imagen 9.- Dispersión de los datos.

Por tanto, dicho algoritmo quedo descartado para su implementación y se tomaron como base los resultados obtenidos en la plataforma Weka para formalizar y disponer de los algoritmos que mejor se comportaban respecto al conjunto de datos recopilado.

Para aplicar los algoritmos seleccionados al modelo los experimentos fueron realizados en el lenguaje de programación Python, el cual se utiliza en la distribución Anaconda para adentrarse en el cómputo científico.

Después de haber entrenado el clasificador y obtener un modelo se debe realizar la evaluación del mismo, por lo que se pone a prueba el conocimiento adquirido por el modelo y se evalúa el comportamiento que este presenta.

La precisión y error medio son métricas para medir el desempeño de métodos de clasificación sobre conjuntos de datos balanceados. Sin embargo, no son los adecuados para conjuntos no balanceados. [32]

Las medidas de desempeño que se utilizaron en este caso para evaluar el comportamiento del modelo son:

- Matriz de confusión
- Sensibilidad
- Precisión.

Para llevar a cabo la evaluación del modelo se utilizó un conjunto de prueba que consta de 17 instancias y representa el 30% del conjunto de datos, donde las instancias fueron 1, 2, 3, 2 y 9 pertenecientes a las clases 1, 2, 3, 4, y 5 respectivamente. En la matriz representada en la tabla 6, de 1 instancia real de la clase 1, el sistema predijo correctamente 1, de 2 instancias de la clase 2 predijo 2, de 3 instancias de la clase 3 predijo 3, de 2 instancias de la clase 2 predijo 2 y de 9 instancias de la clase 5 predijo 8 correctos y 1 como clase 4.

		Valor predicho				
		G1	G2	G3	G4	G5
Valor real	G1	1	0	0	0	0
	G2	0	2	0	0	0
	G3	0	0	3	0	0
	G4	0	0	0	2	0
	G5	0	0	0	1	8

Tabla 6.- Matriz de confusión del modelo.

A partir de la matriz se puede observar que el modelo tiene problemas al distinguir entre la clase 4 y 5, pero que puede distinguir suficientemente bien entre las demás clases.

Aplicando la ecuación 6 y tomando como referencia los valores de la matriz de confusión se obtiene que el modelo alcanza una sensibilidad de 94%.

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

Ecuación 6.- Sensibilidad del conjunto de datos.

Para determinar la proporción de la clasificación de casos positivos que fueron correctos se aplica la ecuación 7 a los valores de la matriz de confusión y se obtiene que el modelo tiene 96% de precisión.

$$Precisión = \frac{TP}{(TP + FP)}$$

Ecuación 7.- Precisión del conjunto de datos.

Una vez que se realizaron pruebas con el conjunto de datos se diseñó y desarrolló una versión beta de la aplicación que utiliza el modelo propuesto en esta investigación para demostrar el funcionamiento del mismo.

La interfaz consiste en llenar una serie de campos que hacen referencia a los factores de riesgo que determinan la nefropatía en diabéticos tipo 2 (imagen 10).

Imagen 10.- Versión beta de aplicación de nefropatía diabética.

Para encontrar el estadio al que pertenece el paciente es necesario cargar los datos en la interfaz (imagen 10) para posteriormente pulsar el botón *clasificar,* el cual como su nombre lo indica realiza la tarea de clasificación. Al pulsar el botón *clasificar* se realizarán las operaciones necesarias para llevar acabo la clasificación, es decir, a partir del conjunto de datos que se tiene se seguirá el proceso del modelo propuesto para encontrar el estadio al que pertenece dicho paciente según los valores ingresados y las operaciones realizadas, y por último mostrar el nivel de afectación renal en el campo correspondiente.

4.1. Casos de estudio

Para demostrar el funcionamiento del modelo propuesto se tienen tres casos de estudio, los cuales representan tres pacientes con diagnóstico nefrótico y diabetes tipo 2 que por tanto son individuos que se pueden incluir en el estudio. Es importante mencionar que los datos que se cargaron son ajenos al conjunto de datos almacenado, es decir, los pacientes no son parte del conjunto de datos que se está manejando en el estudio.

CASO # 1: El individuo presenta nefropatía diabética en estadio cinco y los datos que lo representan se muestran en la imagen 11.

Imagen 11.- Paciente femenino con nefropatía diabética.

Por lo anterior, se observa que el modelo propuesto y el experto de la salud han predicho el mismo estadio de afectación renal al paciente que representa el conjunto de datos cargado.

Caso # 2: El siguiente conjunto de datos representan a un paciente del sexo femenino con insuficiencia renal crónica (imagen 12).

Imagen 12.- Paciente femenino con insuficiencia renal.

La insuficiencia renal crónica se encuentra dentro de los estadios 3 y 4, el resultado que obtuvo el modelo indica que el paciente se encuentra en un estadio 3, ya que su nivel de creatinina es bajo y el nivel de filtración glomerular es intermedio.

Caso # 3: Se trata de un paciente del sexo masculino, con un nivel de obesidad elevado y un diagnóstico de insuficiencia renal (imagen 13).

Imagen 13.- Paciente masculino con insuficiencia renal.

Por lo anterior, el resultado obtenido utilizando el modelo indica que el paciente se encuentra en un estadio 3, debido a que presenta un nivel de creatinina bajo y un descenso ligero - moderado del filtrado glomerular.

4.2. Contrastación de la hipótesis

4.2.1. Hipótesis científicas

H_0: El modelo propuesto obtiene un porcentaje menor a 90% al predecir el inicio de nefropatía en pacientes con DM2.

H_1: El modelo propuesto es capaz de predecir el inicio de nefropatía en un 90 % en pacientes con DM2.

4.2.2. Hipótesis estadísticas

K=90%

H_0 -> $\mu_{nefropatíaDM2} < k$.

H_1 -> $\mu_{nefropatíaDM2} >= k$.

4.2.3. Datos de la muestra

Tamaño de la muestra (N) = 30.

Grados de libertad (n) = 29.

Nivel de significancia (μ) = 0.05.

Rechazar si -1.5232 > T < 1.5232.

Aceptar si T ≥ -1.5232 ó T ≥ 1.5232.

Análisis de resultados							
Corrida	Sensibilidad del modelo %	Valor esperado %	Precisión	Matriz de confusión	Aciertos	$x-\bar{x}$	$(x-\bar{x})^2$
1	100	90	100	4 0 0 0 0 0 2 0 0 0 0 0 3 0 0 0 0 0 2 0 0 0 0 0 6	17	11.1930	125.2829
2	94	90	95	2 0 0 0 0 0 1 0 0 0 0 0 1 0 0	16	5.1930	26.9671

				0 0 0 5 0 0 0 0 1 7			
3	82	90	91	2 0 0 0 0 0 1 0 1 0 0 0 2 0 0 0 0 1 0 0 0 0 0 1 9	14	-6.8070	46.3355
4	71	90	68	1 0 0 0 0 0 0 0 1 0 0 2 0 0 0 0 0 1 1 1 0 0 0 0 10	12	- 17.8070	317.0898
5	82	90	95	2 2 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 2 1 0 0 0 0 9	14	-6.8070	46.3355
6	100	90	100	2 0 0 0 0 0 2 0 0 0 0 0 5 0 0 0 0 0 1 0 0 0 0 0 7	17	11.1930	125.2829
7	100	90	100	2 0 0 0 0 0 2 0 0 0 0 0 3 0 0 0 0 0 1 0 0 0 0 0 9	17	11.1930	125.2829
8	82.35	90	89	1 0 0 0 0 0 1 2 0 0 0 0 2 0 0 0 0 0 1 1 0 0 0 0 9	14	-6.4570	41.6931
9	94.11	90	95	3 0 0 0 0 0 1 0 0 0 0 0 4 0 0 0 0 0 1 1 0 0 0 0 7	16	5.3030	28.1216
10	94.11	90	100	3 1 0 0 0 0 0 0 0 0 0 0 3 0 0 0 0 0 3 0 0 0 0 0 7	16	5.3030	28.1216
11	94.11	90	95	2 0 0 0 0 0 2 0 0 0 0 0 2 0 0 0 0 0 1 1 0 0 0 0 9	16	5.3030	28.1216
12	94.11	90	97	2 1 0 0 0 0 1 0 0 0 0 0 4 0 0 0 0 0 3 0 0 0 0 0 6	16	5.3030	28.1216
13	88.235	90	94	1 0 0 0 0 0 1 0 0 0 0 0 2 0 0	15	-0.5720	0.3272

				0 0 0 2 0 0 0 0 2 9			
14	70.588	90	69	1 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 3 0 0 0 0 9	12	-18.2190	331.9326
15	88.235	90	87	1 0 0 0 0 0 1 0 0 0 0 0 4 0 0 0 0 2 2 1 0 0 0 0 6	14	-0.5720	0.3272
16	88.235	90	92	2 0 0 0 0 1 1 0 0 0 0 0 3 0 0 0 0 0 2 0 0 0 0 1 7	15	-0.5720	0.3272
17	58.8235	90	86	1 3 0 0 0 0 0 0 0 0 0 0 2 3 0 0 0 0 1 1 0 0 0 0 6	10	-29.9835	899.0113
18	94.11	90	95	1 0 0 0 0 0 1 0 0 0 0 0 4 0 0 0 0 0 3 1 0 0 0 0 7	16	5.3030	28.1216
19	100	90	100	4 0 0 0 0 0 3 0 0 0 0 0 2 0 0 0 0 0 1 0 0 0 0 0 7	17	11.1930	125.2829
20	64.705	90	67	1 0 0 0 0 0 1 2 0 0 0 3 2 0 0 0 0 0 1 1 0 0 0 0 6	11	-24.1020	580.9072
21	94.11	90	95	1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 3 1 0 0 0 0 10	16	5.3030	28.1216
22	94.11	90	97	1 0 0 0 0 1 2 0 0 0 0 0 4 0 0 0 0 0 3 0 0 0 0 0 6	16	5.3030	28.1216
23	82.352	90	91	2 0 0 0 0 0 3 0 0 0 0 3 1 0 0 0 0 0 3 0 0 0 0 0 5	14	-6.4550	41.6672
24	94.11	90	95	2 0 0 0 0 0 3 0 0 0 0 0 1 0 0	16	5.3030	28.1216

				0 0 0 3 1 0 0 0 0 7			
25	100	90	100	1 0 0 0 0 0 1 0 0 0 0 0 2 0 0 0 0 0 1 0 0 0 0 0 12	17	11.1930	125.2829
26	94.11	90	100	0 0 0 0 0 1 2 0 0 0 0 0 3 0 0 0 0 0 1 0 0 0 0 0 10	16	5.3030	28.1216
27	82.352	90	91	1 0 0 0 0 0 1 2 0 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 10	14	-6.4550	41.6672
28	94.11	90	95	2 0 0 0 0 0 2 0 0 0 0 0 6 0 0 0 0 0 1 1 0 0 0 0 5	16	5.3030	28.1216
29	100	90	100	4 0 0 0 0 0 1 0 0 0 0 0 2 0 0 0 0 0 1 0 0 0 0 0 9	17	11.1930	125.2829
30	88.235	90	90	4 0 0 0 0 0 1 0 0 0 0 0 2 0 0 0 0 1 1 0 0 0 0 1 7	15	-0.5720	0.3272

4.2.4. Operaciones para calcular el estadístico

En la prueba t para una muestra se evalúa que la media de la población en estudio es igual a un valor determinado y permite comprobar si es posible mantener la hipótesis nula que corresponde a la media calculada. La muestra en este caso corresponde a los valores que representan el porcentaje de sensibilidad del modelo.

Para el cálculo de la prueba se debe llevar a cabo el siguiente procedimiento:

1. El primer paso consiste en calcular la media aritmética de la muestra, para lo cual se requiere datos de la muestra (X) y cantidad de datos (N):

$$\bar{\chi} = \frac{\sum X}{N} = \frac{2664.2105}{30} = 88.8070$$

2. Enseguida se calcula la desviación típica, la cual se encuentra a partir del dato de la muestra menos la media elevados al cuadrado ($(\chi i - \bar{\chi})^2$) y los grados de libertad (n):

$$S = \sqrt{\frac{\sum(\chi i - \bar{\chi})^2}{n}} = \sqrt{\frac{3407.8289}{29}} = \sqrt{117.5113} = 10.8403$$

3. Para obtener el valor del estadístico se emplean los valores encontrados en (1) y (2), además del valor de referencia (μ) y el tamaño de la muestra (N):

$$T = \frac{\bar{\chi} - \mu}{\frac{S}{\sqrt{N}}} = \frac{88.8070 - 90}{\frac{10.8403}{\sqrt{30}}} = \frac{-1.1930}{\frac{10.8403}{5.477}} = \frac{-1.1930}{1.9792} = -0.6028 \rightarrow t_{29}$$

4. Tomando como referencia los grados de libertad y el nivel de significancia el valor que corresponde al Estadístico t-student en la tabla es 1.699.

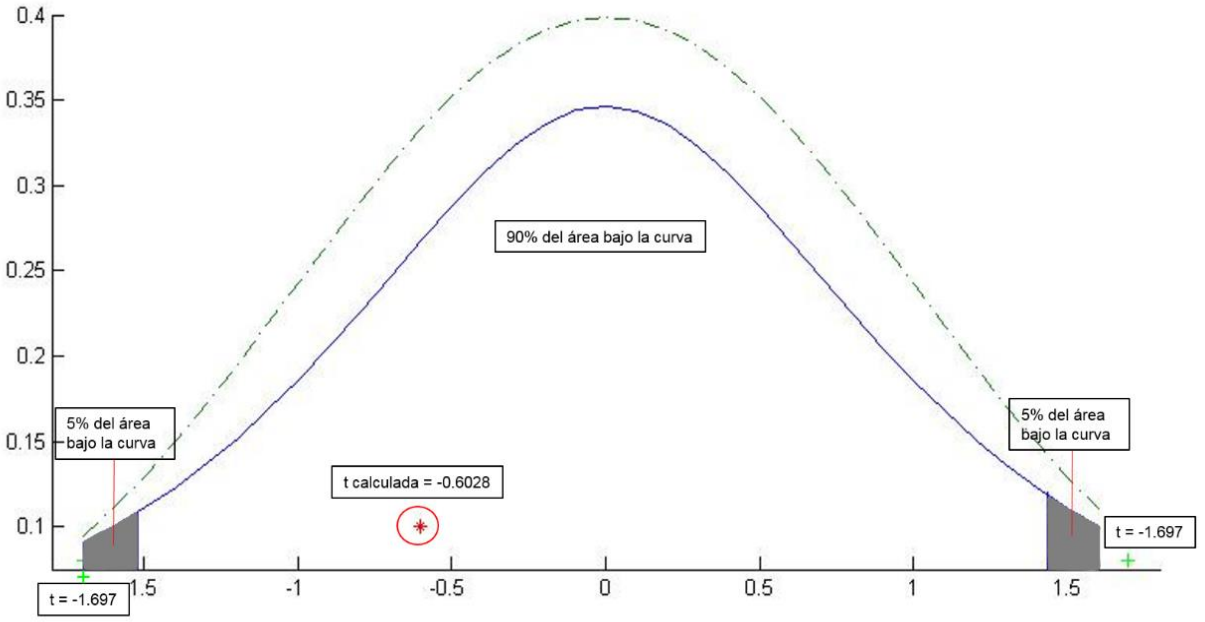

Imagen 14.- Estadístico t-student del porcentaje de sensibilidad del modelo propuesto.

El valor obtenido -0.6028 al calcular el estadístico se encuentra entre -1.5232 y 1.5232 (Imagen 14), es decir, se rechaza la H_0. Por tanto, la evidencia demuestra que de 17 elementos que se clasifican un 90% se hace de manera correcta, lo que indica que la H_1 es aceptada.

Capítulo V

Conclusiones y trabajo futuro

5.1. Conclusiones

La presente tesis tuvo como objetivo el diseño de un modelo para la clasificación de nefropatía en pacientes diabéticos mediante el uso de técnicas de selección de atributos y algoritmos de aprendizaje automático, aplicando la clasificación de las guías prácticas clínicas KDIGO con la finalidad de clasificar la nefropatía en los estadios I, II, III, IV y V.

Para llevar a cabo esto, fueron necesarias algunas tareas importantes dentro de las que se encuentran análisis, recopilación, digitalización y preprocesamiento de datos, selección y clasificación de atributos, y un análisis de resultados que sirvió para determinar el rechazo o aceptación de la hipótesis a partir del comportamiento del modelo.

A lo largo de la investigación una de las tareas en las que se invirtió mayor cantidad de tiempo es la recopilación de datos, ya que actualmente en el Hospital General de Misantla no se cuenta con expedientes clínicos digitales y examinar los expedientes físicos uno por uno requiere de un periodo de tiempo extenso.

Respecto a la tarea de clasificación el modelo es apto para agrupar el daño renal en cinco categorías, lo que significa que el modelo es capaz de clasificar o encontrar el nivel de afectación

renal del paciente tomando como referencia los valores que representan los factores de riesgo determinantes de la nefropatía, obteniendo un 90% de sensibilidad y un tiempo que no rebasa los 60s al ejecutar el modelo.

Mediante el diseño del modelo se pretende aportar a los médicos sin experiencia en la enfermedad una herramienta de apoyo para clasificar la nefropatía en una etapa adecuada, además de ayudar a la toma de decisiones con respecto al seguimiento del paciente.

5.2. Trabajo a futuro

Si bien la presente tesis abordó como la nefropatía afecta a la comunidad diabética en México por la obesidad, mal control de la presión arterial y glucosa, tratamiento farmacéutico inadecuado y falta de herramientas tecnológicas para la clasificación de nefropatía. Para dar seguimiento a la investigación temas interesantes a tratar puede ser:

- Agregar al conjunto de datos el registro de personas sanas para que se pueda realizar la tarea de predicción.
- Alimentar continuamente el conjunto de datos para obtener un buen desempeño de predicción de la enfermedad en una etapa temprana.
- Complementar la selección de atributos aplicando otros algoritmos para realizar un rankeo.
- Diseñar y desarrollar una aplicación que utilice el modelo para alimentar el conjunto de datos.
- Implementar la aplicación en la unidad adecuada de la secretaria de salud y observar su comportamiento.

REFERENCIAS

[1] Alfredo Torres Viloria, Alfredo Torres Viloria, R. Z. C. Nefropatía diabética. Hospital General \Dr. Manuel Gea González" (2002).

[2] Ananthapadmanaban, K. R., and Parthiban, G. Prediction of chances – diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology 7 (2014), 1498-1503.

[3] Bala, S., and Kumar, K. A literature review on kidney disease prediction using data mining classification technique. International Journal of Computer Science and Mobile Computing 3 (2014), 960-967.

[4] Caparrini, F. S. Introducción al aprendizaje automático. http://www.cs.us.es/~fsancho/?e=75, 2015.

[5] Cho, B.-H., Lee, J.-S., Chee, Y.-J., Kim, K.-W., Kim, I.-Y., and Kim, S.-I. Prediction of diabetic nephropathy from diabetes dataset using feature selection methods and SVM learning. Journal of Biomedical Engineering Research 28 (2007), 355-362.

[6] Dalakleidi, K. V., Zarkogianni, K., and Karamanos, V. G. A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in type 2 diabetes patients. In Bioinformatics and Bioengineering (BIBE) (2014).

[7] Elvert, A. Nefropatía Diabética. Tech. rep., Universidad de Buenos Aires, Buenos Aires.

[8] En Gestión de Datos, P. E. Técnicas de minería de datos y modelos predictivos: aspectos a tener en cuenta. http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/tecnicas-de-mineria-de-datos-y-modelos-predictivos-aspectos-a-tener-en-cuenta, 2016.

[9] Flores, C. H., and Godoy, G. V. Nefropatía diabética en pacientes que asisten a la consulta externa de endocrinología pediátrica del hospital escuela. Revista Médica de los Post Grados de Medicina 10 (2007), 79-82.

[10] Gutiérrez, C. G., and Rodríguez, J. C. S. Nefropatía diabética: prevención o retraso por el médico general integral versus lamentos del nefrólogo. Revista Cubana de Medicina General Integral 13 (1997), 19-27.

[11] Huang GM., Chen YC., Weng J.TY. (2015) Construction of a Prediction Model for Nephropathy Among Obese Patients Using Genetic and Clinical Features. In: Li XL., Cao T., Lim EP., Zhou ZH., Ho TB., Cheung D. (eds) Trends and Applications in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, vol 9441. Springer, Cham

[12] Huang, G.-M., Huang, K.-Y., Lee, T.-Y., and Weng, J. T.-Y. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinformatics 16 (2015), S5.

[13] Huang, G.-M., Lee, Y.-C., Weng, J. T.-Y., Chen, Y.-C., and Wu, L. S.-H. Cluster-based classification of diabetic nephropathy among type 2 diabetic patients. In International Congress on Natural Sciences and Engineering (ICNSE).

[14] Improving Global Outcomes (KDIGO) CKD-MBD Work Group. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease–mineral and bone disorder (CKD–MBD). Kidney International 2009; 76 (Suppl 113): S1–S130.

[15] Lagani, V., et al., Development and validation of risk assessment models for diabetes-related complications based on the DCCT/EDIC data, Journal of Diabetes and Its Complications (2015), http://dx.doi.org/10.1016/j.jdiacomp.2015.03.001

[16] Leung, R. K.,Wang, Y., Ma, R. C., Luk, A. O., Lam, V., Ng, M., So, W. Y., Tsui, S. K., and Chan, J. C. Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case{control cohort analysis. BMC Nephrology 14 (2013).

[17] Morteza, A., Nakhjavani, M., Asgarani, F., Carvalho, F. L., Karimi, R., and Esteghamati, A. Inconsistency in albuminuria predictors in type 2 diabetes: a comparison between neural network and conditional logistic regression. Translational Research 161 (2013), 397-405.

[18] Narasimhan, B., and Malathi, A. Fuzzy logic system for risk-level classification of diabetic nephropathy. In Green Computing Communication and Electrical Engineering (ICGCCEE) (2014).

[19] OMENT. Cifras de sobrepeso y obesidad en méxico-ensanut mc 2016. http://oment.uanl.mx/cifras-de-sobrepeso-y-obesidad-en-mexico-ensanut-mc-2016, 2016.

[20] OMS. Diabetes. http://www.who.int/diabetes/action_online/basics/es/index3.html, 2017.

[21] RAHIMLOO, P., and JAFARIAN, A. Prediction of diabetes by using artificial neural network, logistic regression statistical model and combination of them. Bulletin de la Société Royale des Sciences de Liége 85 (2016), 1148-1164.

[22] Ruiz, M. Prevención de la nefropatía diabética. , 2006.

[23] S., J. M., and A., P. D. Microalbuminuria como elemento de predicción de nefropatía y riesgo cardiovascular en pacientes diabéticos. Revista Chilena de Endocrinología y Diabetes 3 (2010), 189-196.

[24] Soliman, O. S., and AboElhamd, E. Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine. International Journal of Computer Trends and Technology 8 (2014), 38-44.

[25] Técnicas de análisis de datos. Aplicaciones prácticas utilizando Microsoft y Weka. Universidad Carlos III de Madrid, 2012, ch. , pp. 96-158.

[26] Yin, W.-j., Yi, Y.-h., Guan, X.-f., Zhou, L.-y., Wang, J.-l., Li, D.-y., and Zuo, X.-c. Preprocedural prediction model for contrast-induced nephropathy patients. Journal of the American Heart Association 6 (2017).

[27] S. Burciaga *et al.* "Desarrollo de una aplicación Web para predecir la aparición de complicaciones en pacientes con diabetes tipo II", *Research in Computing Science., vol. 77, pp 109-119, 2014.*

[28] E. Palma. Modelos predictivos y descriptivos en minería de datos., https://es.slideshare.net/lalopg/mtodos-predictivos-y-descriptivos-minera-de-datos, 2015.

[29] Modelos predictivos: reforzando el valor de una buena decisión., https://blog.es.logicalis.com/analytics/modelos-predictivos-reforzando-el-valor-de-una-buena-decision, 2015. (visto 13-10-17).

[30] Yang, Y., Zhang, S., Lu, B. et al. Wien Klin Wochenschr (2015) 127: 669. https://doi.org/10.1007/s00508-014-0679-1.

[31] Cuevas, C. y Alejo, A. (2010). UNAM facultad de psicología, octubre 2010, Validez y fiabilidad de las medidas de exposición y medición [PDF]. Recuperado de http://www.psicol.unam.mx/Investigacion2/pdf/SENSIBILIDAD%20Y%20ESPECIFICIDAD.pdf

[32] Puente, L. et al. (2014) Método rápido de preprocesamiento para clasificación en conjuntos de datos no balanceados. Research in Computing Science, 73, 129–142.

[33] Wikipedia (2017). Matriz de confusión. Wikipedia. Recuperado de: https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n

[34] Serna, S. (2009). Comparación de árboles de regresión y clasificación y regresión logística (Tesis de Maestría). Universidad nacional de Colombia. Medellín, Colombia.

[35] Troncoso, A. (s.f.) Selección de atributos [PDF]. Recuperado de http://eps.upo.es/troncoso/MaterialDocente/MasterComputing/Tema3-SeleccionAtributos.pdf

[36] Chi Square test for feature selection (2016). Chi Square test for feature selection. Learn for master. Recuperado de http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selection

[37] Faneite A, Pedro. (2003). Las computadoras en la medicina de hoy. Revista de Obstetricia y Ginecología de Venezuela, 63(1), 47-54. Recuperado el 22 de noviembre de 2017 de http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S0048-77322003000100007&lng=es&tlng=es.

[38] Morales, E. & Escalante, H. J. (s.f.). Aprendizaje Computacional [PDF]. Recuperado de https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/intro.pdf

[39] T. Mitchell (1997) Machine Learning, McGraw–Hill.

[40] U.S. Department of Health and Human Services. (s.f.). Diabetic Kidney Disease. Recuperado de https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/diabetic-kidney-disease.

[41] Lab Tests Online-ES. Tasa de filtrado glomerular. (2018). Recuperado de https://labtestsonline.es/tests/tasa-de-filtrado-glomerular.