Diagnóstico de diabetes utilizando árboles de clasificación

Equipo 03: Dávila Martínez A. S
Martínez Acosta M.A
Rodríguez Pacheco J.M

UANL

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Introducción

La diabetes es una enfermedad crónica que se origina porque el páncreas no sintetiza la cantidad de insulina que el cuerpo humano necesita.
La importancia de la diabetes mellitus recae en la frecuencia con la que se presenta dicha enfermedad, por ejemplo en México, es la segunda causa de muerte, según el Instituto Nacional De Salud Pública.

Adicionalmente, la diabetes contribuye de manera significativa a otras enfermedades, por lo que es importante el desarrollo de proyectos de investigación que involucren nuevos métodos de diagnóstico.

Hipótesis

Las variables del nivel de glucosa, de insulina sérica, así como la edad, están relacionadas al padecimiento de la diabetes.

- La población estudiada es propensa a padecer diabetes

Objetivos

Principal:

Crear un modelo de predicción para analizar si una persona tiene diabetes o no.

Secundarios:

- Visualizar la distribución de los datos de interés.
- Establecer qué tan correlacionadas están las variables en la base de datos.
- Visualizar gráficamente el rendimiento del modelo.
- Entender si existe una razón médica para tener el valor de 0 en las columnas que lo presenten.

Metodología

La base de datos contiene variables que representan factores de riesgo para la diabetes presentadas en el grupo de mujeres del pueblo Pima en el año de 1990, por ende, cuando se presenta el valor de 0 en alguna medición exceptuando el número de embarazos, se tiene poca fiabilidad en los resultados, por lo tanto decidimos eliminar aquellas filas que contienen ceros.
La base de datos original tenía 768 filas, después de modificarla, se contaban con 392 columnas.

Estadística descriptiva

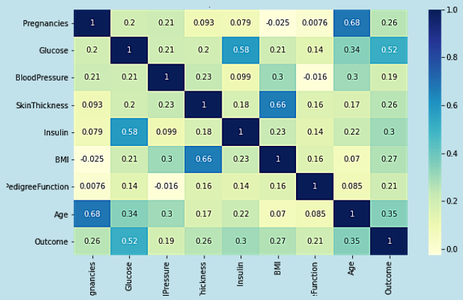
Se tiene como hipótesis determinar si el pueblo Pima es propenso a padecer diabetes mellitus, por lo que utilizamos medidas de tendencia central como la indicador para comparar el grado de riesgo que presenta la media de la población como se muestra en la **figura 1.1.**

- En adición, recurrimos a fuentes confiables de información médica para concluir al respecto.

Figura 1.1: Media de variables que son factores de riesgo para la diabetes.

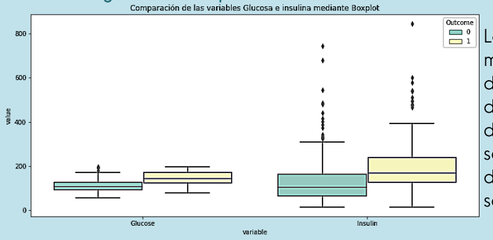
Variable	Media de la variable	Conclusión basada en documentación médica
Nivel de glucosa	122.61 mg/dL	Un nivel de Glucosa entre 100 y 125 mg/dL se considera prediabetes, por lo que concluimos que las mujeres del pueblo Pima son propensas a padecer prediabetes.
Nivel de insulina sérica	155.88 μU/mL	El nivel de insulina sérica promedio está fuera del rango normal.
Índice de masa corporal	33 kg/m^2	Dado que se considera *obesidad de grado I* a un IMC de 30 a 34.9, se concluye que las mujeres del pueblo Pima tienden a padecer obesidad grado I.
Presión arterial diastólica	70.64 mmHg	Un valor inferior a 80 mmHg, se considera una presión arterial normal, entonces se concluye que el promedio de presión arterial diastólica está en un rango normal

Figura 2.1: Mapa de calor



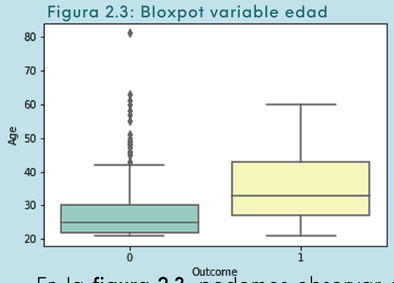
La **figura 2.1** muestra que los niveles de glucosa, la edad y la insulina tienen una correlación significativa con la variable de resultado, por lo que pudimos verificar nuestra hipótesis.
También podemos concluir que la correlación entre pares de características, como la edad y los embarazos, o la insulina y la Glucosa muestran una correlación significativa.

Figura 2.2: Bloxpot de variables de interés



La **figura 2.2** muestra que las mujeres mayores a 21 años, del pueblo Pima que son diabéticas poseen un nivel de Glucosa y de insulina sérica más elevado a diferencia de las que no lo son.

Figura 2.3: Bloxpot variable edad



En la **figura 2.3**, podemos observar que la edad de las mujeres diabéticas es mayor que la de mujeres no diabéticas

Figura 2.4: Histograma de la variable Glucose

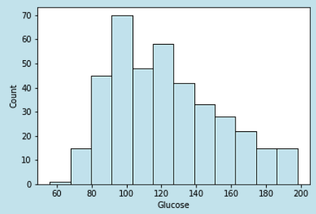


Figura 2.5: Histograma de la variable Insulin

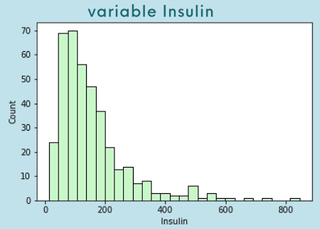
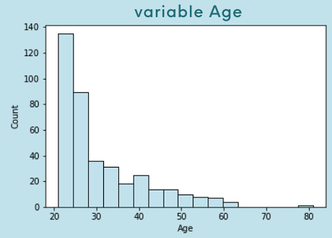


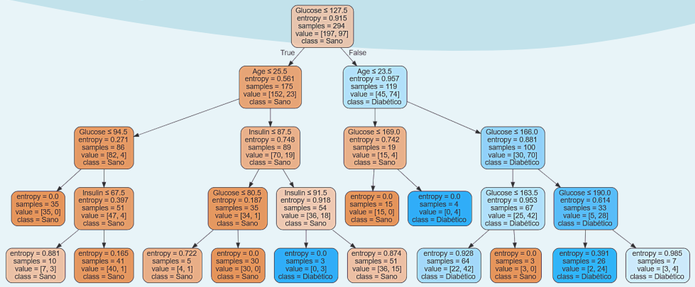
Figura 2.6: Histograma de la variable Age



De la **figura 2.4 a la 2.6** se muestra la distribución de las variables que están altamente correlacionadas con la variable Outcome, esto para entender la frecuencia con la que se presentan los niveles de riesgo para cada variable.

Resultados

Figura 3.1: Árbol de clasificación



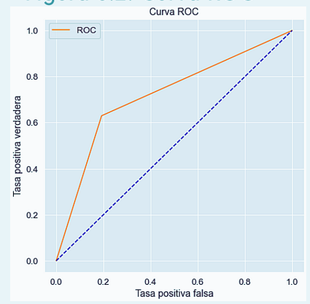
Para el **árbol de clasificación** (figura 3.1) se contemplaron las variables Glucose, Age e Insulin, esto basándonos en su correlación.

AUC=0.76

El valor ideal de un AUC es 1, es este caso obtuvimos 0.76 lo que nos indica que nuestro modelo es 76% exacto para su clasificación

Por otro lado, la curva ROC muestra que es un modelo bueno, sin embargo aún se puede mejorar la exactitud de dicho modelo.

Figura 3.2: Curva ROC



Recursos

La base de datos usada es **Pima Diabetes Database** proporcionada por la página kaggle.

Base de datos

Figura 4: Herramientas usadas para el desarrollo del proyecto.

Github de proyecto

Referencias

Conclusiones

A través de estadística descriptiva, determinamos que el 66.8% de las mujeres en el pueblo Pima NO tiene diabetes, en cambio el 33.2% si padece diabetes, así mismo determinamos que el nivel promedio de Glucosa 122.61mg/dL corresponde a la clasificación de prediabetes y el nivel promedio de insulina sérica está fuera del rango normal, por ende, vemos que el grupo estudiado es propenso a padecer diabetes.

Adicionalmente, determinamos con el mapa de calor que el nivel de glucosa, insulina sérica y la edad influyen significativamente en la diabetes.
En cuanto al desarrollo del árbol de clasificación, obtuvimos que este es capaz de clasificar al 76% correctamente.

Finalmente, como trabajo a futuro tenemos el desarrollar un modelo predictivo que sea capaz de predecir con una exactitud mayor, esto auxiliándonos en técnicas de minería de datos. En nuestro desarrollo del proyecto nos percatamos que una máquina de soporte vectorial podría ayudar a lograr este cometido, así como reconsiderar las variables empleadas en la técnica.