Universidad Panamericana

Facultad de Ingeniería

Materia

Inteligencia de Negocio y Soluciones basadas en Datos

Profesor

Dra. Karina Ruby Pérez Daniel

Fecha de entrega

28 de Noviembre, 2022

Ciclo

1228

Nombre del proyecto

Factores imprescindibles para el diagnóstico de la diabetes

Integrantes del Equipo

Nombre	ID
Arias Cortina, Carlos	0217787

Factores imprescindibles para el diagnóstico de la diabetes

Este proyecto presenta la búsqueda de un modelo que permitiera determinar con un alto nivel de precisión la probabilidad de diabetes en las personas. Se quiso perseguir esta narrativa dado que la diabetes es de las principales causas de muerte en México, y como prevención se busca poder predecir un diagnóstico temprano de diabetes para así ser tratado antes de que se vuelva irreversible, por tanto, como equipo se tiene una preocupación por el bienestar de la comunidad y, es por lo anterior que se decidió efectuar un análisis de diagnósticos de diabetes según diversas métricas para casos positivos y negativos.

Descripción del caso de negocio

El set de datos utilizado proviene del *National Institute of Diabetes and Digestive and Kidney Diseases*. Este archivo se creó con el fin de poder predecir si un paciente tiene diabetes al basarse en diversas métricas de diagnóstico. Todas las pacientes del set de datos son mujeres de al menos 21 años.

La diabetes es una enfermedad crónica cuyo caso clínico implica un cambio radical en la manera en la cual el cuerpo transforma los alimentos en energía. Cuando se padece de diabetes, el cuerpo carece de las facultades para poder producir las cantidades adecuadas de insulina y, al no poder igualar la demanda de ésta, el exceso de azúcares se mantiene en el torrente sanguíneo; provocando con el paso del tiempo problemas de salud graves como lo son las enfermedades cardiovasculares, renales y puede provocar pérdidas de visión. La diabetes es una enfermedad que se divide en dos tipos. La diabetes tipo 1 es fácilmente identificable y suele diagnosticarse a niños de temprana edad. Por otro lado, la diabetes tipo 2 puede ser muy peligrosa para aquellas personas a las que no se les ha diagnosticado o a quienes se les ha dado un diagnóstico erróneo, debido a que es un proceso más pasivo y los signos y síntomas se presentan en diferentes etapas de tiempo y, para ser controlada debe de modificarse todo el estilo de vida de la persona.

Analizar esta información no sólo permite predecir el diagnóstico de diabetes de una persona según sus características, sino que también ayudará a determinar y analizar la importancia de estas características y su presencia en las personas padecientes de diabetes con la finalidad de generar mayor grado de conciencia en la población, resaltar los riesgos y problemáticas de la posibilidad de tener diabetes, así como concientizar los riesgos en el estilo de vida y cuidados necesarios respecto a esta enfermedad pero sobre todo poder dar un diagnóstico temprano como prevención a que llegue a ser fatal.

Descripción del set de datos

Los datos recabados son:

Atributo	Descripción	Tipo de Dato
Pregnancies	Número de embarazos que tuvo la persona a la fecha.	Numérico / Discreto
Glucose	Nivel de glucosa en la sangre de la persona.	Numérico / Discreto
BloodPressure	Presión sanguínea de la persona al momento de registrarla.	Numérico / Discreto
SkinThickness	Grosor de la piel de la persona.	Numérico / Discreto
Insulin	Nivel de insulina en la sangre de la persona.	Numérico / Discreto
BMI	Índice de masa corporal de la persona.	Numérico / Continuo
DiabetesPedigree Funcion	Porcentaje de probabilidad de tener diabetes	Numérico / Continuo
Age	Edad de la persona.	Numérico / Discreto
Outcome	Expresar diagnóstico de diabetes.	Categórico

Tratamiento de los datos:

- No hay datos faltantes, así que no fue necesario rellenar o quitar registros.
- Todos los datos, excluyendo el de outcome, son de tipo numérico; vienen en escalas diferentes así que en caso de que se quiera hacer un modelo predictivo a parte de la exploración de datos entonces sería necesario reescalarlos (normalizarlos).
- La clase tiene más registros con diagnóstico negativo de diabetes que positivos, así que para un modelo se podrían eliminar aleatoriamente registros negativos para que éste no aprenda más cargado hacia este resultado que el otro.

Metodología de trabajo

Tipo de análisis que se utilizó, es decir, qué pasos se utilizaron para llevar a cabo la solución.

- Exploración: Se entendió el set de datos con el fin de determinar la forma en la que se debería de trabajar en él y también para buscar la información que pueda estar oculta entre sus datos:

  Datos Faltantes: Se comenzó buscando si había datos nulos en el set de datos y, se llegó al resultado de que no había ninguno. Así que no sería necesario rellenar o quitar registros.

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

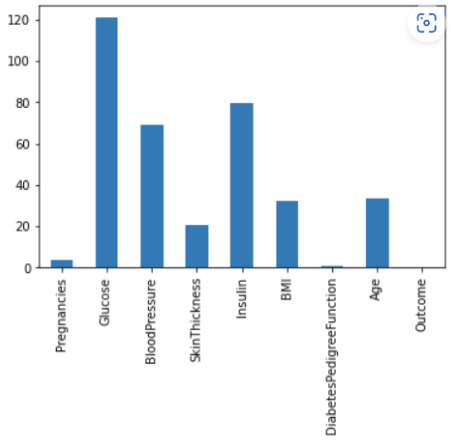
  Tipo de dato: Se obtuvieron los tipos de datos que tenía almacenado el dataframe para cada atributo y se comparó con la cantidad de registros únicos que había para cada uno para determinar si los datos son numéricos o categóricos.

```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
```

```
Pregnancies: 17
Glucose: 136
BloodPressure: 47
SkinThickness: 51
Insulin: 186
BMI: 248
DiabetesPedigreeFunction: 517
Age: 52
Outcome: 2
```

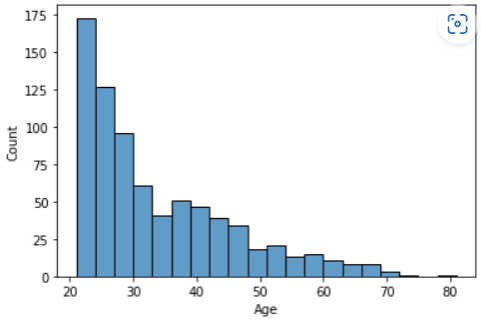
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

  Variabilidad de rango en los datos: Los datos numéricos tienen una diferencia en los rangos que abarcan, puesto que sus dimensiones son diferentes. Esto puede llegar a generar problemas cuando se quiera hacer un modelo predictivo para la clase.

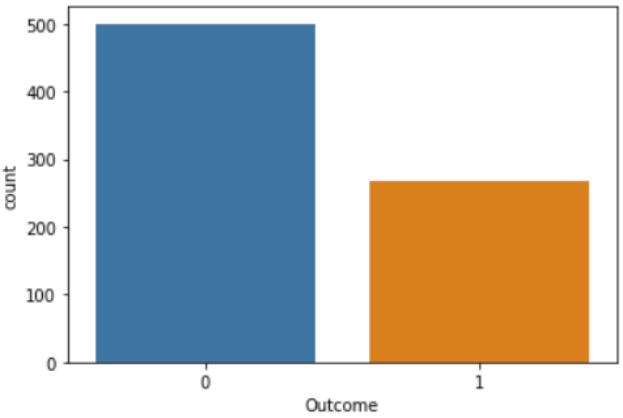


Comportamiento de algunas variables:

a) *Age*: Se realizó un histograma para ver la distribución de edades de las personas dentro del set de datos. Hay menor cantidad de personas a medida que su edad va aumentando. Esto puede ser debido a la forma o accesibilidad de éstas durante la recopilación de los datos.

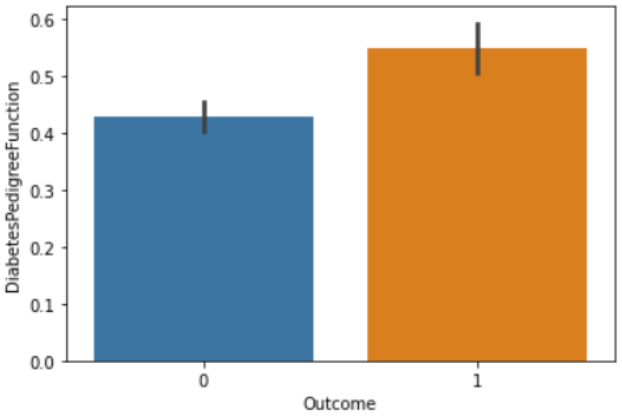


b) *Outcome*: Se contó el número de registros que tienen un diagnóstico negativo y positivo dentro del set de datos. El número de registros con diagnóstico negativo es mayor.

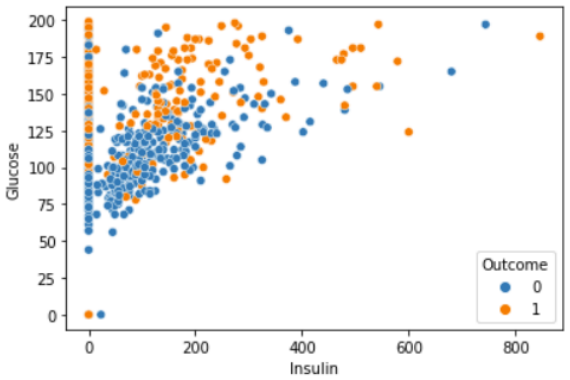


c) *DiabetesPedigreeFunction*: Se analizó la relación entre el diagnóstico de diabetes y la función de diabetes con un gráfico de barras.Muestra que el

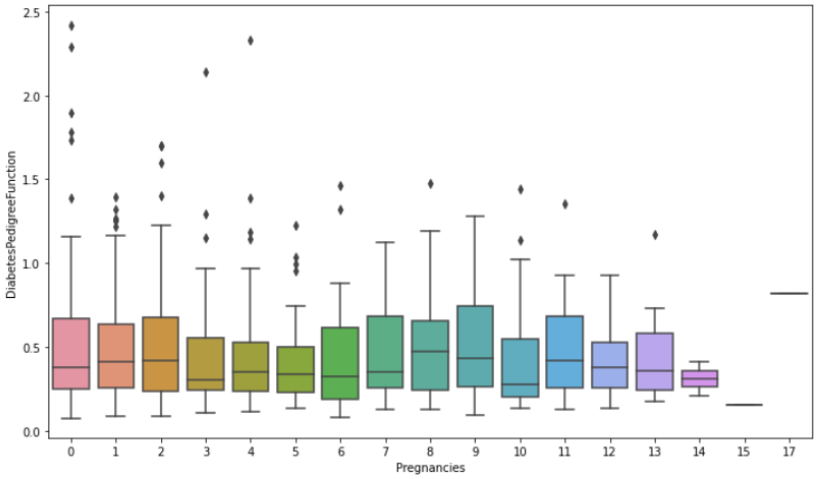
promedio de la función de diabetes es mayor para las personas con diagnóstico positivo.



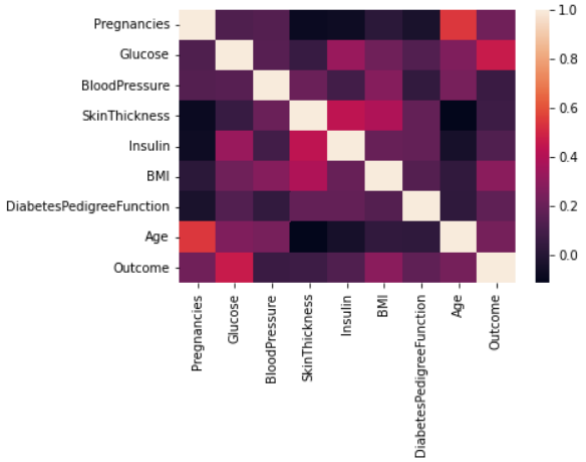
d) *Insulin and Glucose*: Se realizó un gráfico de dispersión para ver si existe una relación entre los niveles de glucosa y de insulina en la sangre de las personas. Además, se utilizó una diferenciación de colores para los casos positivos y los casos negativos. Pareciera que a mayor insulina, mayor glucosa tendrá una persona en la sangre a menos que la insulina sea 0. Se puede ver que la mayoría de los casos con diagnóstico positivo tienen un nivel de glucosa alto.



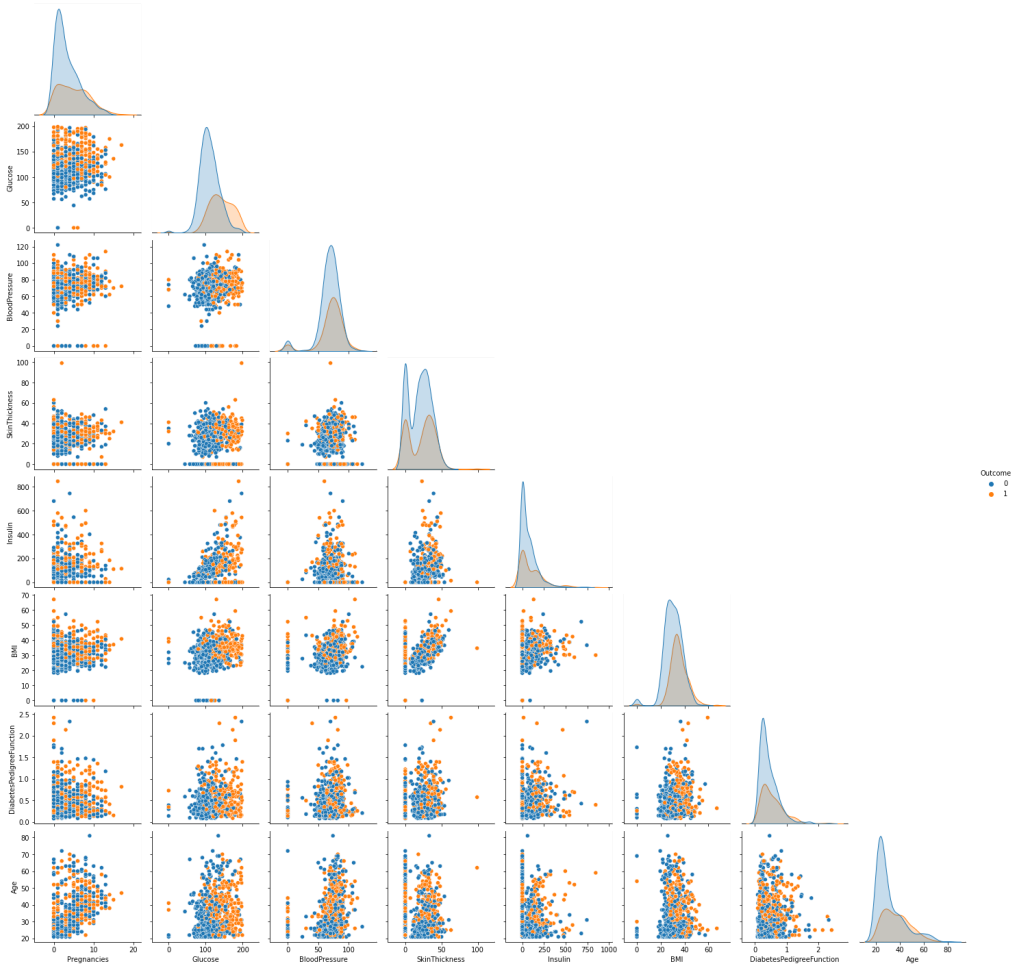
e) *Pregnancies & DiabetesPedigreeFunction*: Se realizó un gráfico de cajas para ver la relación que hay entre la función de diabetes según el número de embarazos que hayan tenido las personas. Pareciera que no hay una tendencia ya que todos permanecen debajo de 1, y en todos los casos hay valores fuera de rango hacia arriba.



f) *Correlaciones*: Por medio de la matriz de correlaciones de los datos, se puede conocer el tipo de relación independiente que tiene cada atributo con los demás e incluso con la clase. Pareciera que con la clase hay muy poca relación exceptuando *Glucose* y *BMI*.



g) *Pairplot*: Se utilizó un gráfico que nos permite visualizar las relaciones entre todas las variables según el diagnóstico de la persona para ver cuáles describen mejor la diferencia entre los casos positivos y los negativos. A primera vista, parece ser que con la glucosa se nota la diferencia.

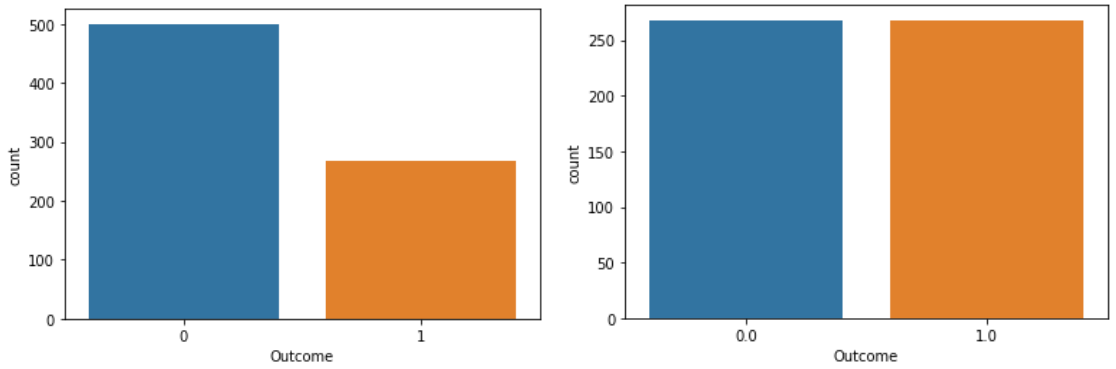


- Limpieza y Tratamiento

  Posteriormente, se realizó una limpieza y tratamiento de los datos por medio de una normalización de datos con el objetivo de poder analizar mejor la información. Para esto, se creó un arreglo con los valores del dataset. Luego, se agregó una variable con los valores del rango (0,1) utilizando **MinMaxScaler()** para poder hacer la normalización. Una vez teniendo esta variable, se realizó una transformación del arreglo de valores del dataset en valores mínimos y máximos, es decir, en ceros y unos. Después, los resultados se muestran en una tabla utilizando el método **describe()** que devuelve la información estadística como el promedio, desviación estándar, valor mínimo, entre otros; por cada columna del dataset.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	0.226180	0.607510	0.566438	0.207439	0.094326	0.476790	0.168179	0.204015	0.348958
std	0.198210	0.160666	0.158654	0.161134	0.136222	0.117499	0.141473	0.196004	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.058824	0.497487	0.508197	0.000000	0.000000	0.406855	0.070773	0.050000	0.000000
50%	0.176471	0.587940	0.590164	0.232323	0.036052	0.476900	0.125747	0.133333	0.000000
75%	0.352941	0.704774	0.655738	0.323232	0.150414	0.545455	0.234095	0.333333	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Para tener certeza de los modelos, también se preparó una regularización del número de registros para la clase positiva y para la clase negativa, sin embargo, al reducir el set de datos los modelos ya no desempeñaron bien, así que se mantuvo este paso en teoría.



Antes → Después

- Selección de Variables

  Para realizar los modelos había que determinar cuales variables eran las más adecuadas para utilizar, ya que, es probable que muchas de ellas no dieran suficiente información como para tener que incluirlas. Durante la exploración se observó que se podía incluir *Glucose*. Pero, para explorar si usar las otras es adecuado, entonces se optó por hacer modelos con todas, algunas pocas y con sólo una para ver cuál sería mejor.

- Selección de Modelos
  a. *Red Neuronal MLP*: Para el primer modelo, se comenzó con una validación cruzada usando 20 vías con una estrategia de bootstrap que es para obtener de forma iterativa el área bajo la curva y la precisión del modelo. Enseguida, se procedió a utilizar un clasificador de redes neuronales (MLP) utilizando el modelo de regresión logística (lbfgs) y una capa oculta de cinco neuronas. Después, al haber realizado una función óptima con el 30% de prueba y el 70% de entrenamiento para ajustar el modelo. Teniendo esto, se pudo realizar el cálculo del área bajo la curva, que en este caso se tienen los 20 valores para medir el desempeño del modelo; finalmente se calculó la probabilidad de precisión, de igual manera con 20 valores para medir el desempeño del modelo. Con este modelo, se obtuvieron los siguientes resultados:

```
Prediccion:  [1. 0. 1. 1. 1. 0. 0. 0. 1. 0.]
Datos Reales:  [1. 0. 0. 1. 1. 0. 0. 0. 0. 0.]
Area bajo la curva: 0.8283 sd: 0.0303
Accurracy: 0.7701 sd: 0.0292
```

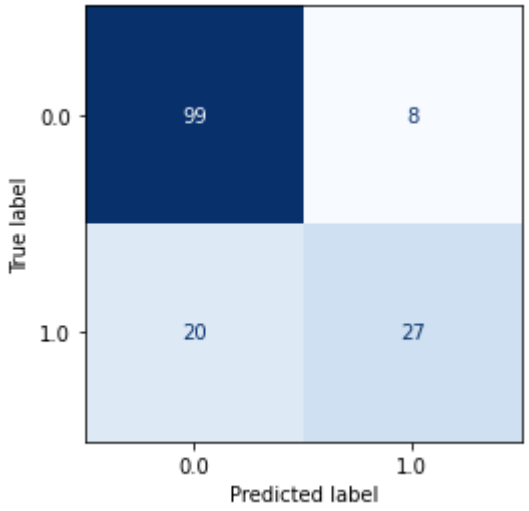
b. *Modelos de Clasificación*: Para el segundo modelo, se realizaron cinco pruebas con parámetros base o default. Para esto se crearon dos variables con el objetivo de crear un conjunto de entrenamiento y prueba.
   i. Regresión Logística
   ii. Árbol de Decisiones
   iii. K-Nearest Neighbors (KNN)
   iv. Random Forest
   v. K-Means

```
------ Logistic Regression ------
Test accuracy = 81.82 %

------ Decision Tree ------
Test accuracy = 79.87 %

------ KNN ------
Test accuracy = 77.92 %

------ Random Forest ------
Test accuracy = 81.17 %

------ K Means ------
Test accuracy = 29.87 %
```

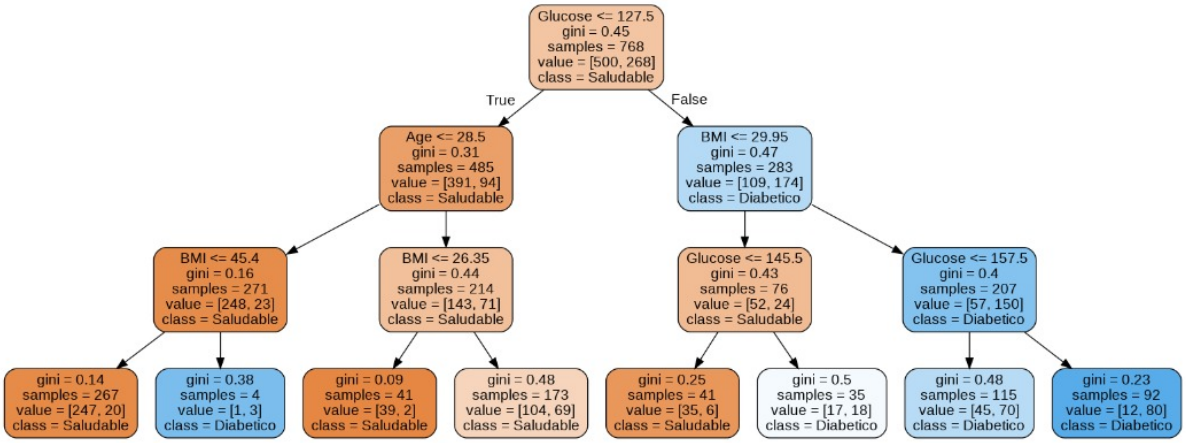
c. *Regresión Logística Unidimensional*: se trató también de analizar el comportamiento del modelo; pero no era lo suficientemente complejo como para poder estimar mejor que los modelos de clasificación anteriores, ya que obtuvo una precisión de 76%.

d. *Regresión Logística:* Al obtener la regresión logística como el modelo con mayor accuracy se decidió como el mejor modelo por lo que se volvió a trabajar utilizando los parámetros **multi_class='ovr', solver='lbfgs'** sin embargo se obtuvo la misma accuracy que con los parámetros defaults por lo que se decidió esta como resultado final y se obtuvo su matriz de confusión.

```
The accuracy of Logistic Regression is : 81.82 %
```



e. *Árbol de decisión:* por último como extra se realizó un árbol de decisión para poder apreciar cuáles son los casos en los que las diferentes variables van llevando a que el resultado al diagnóstico de diabetes sea positivo o negativo.



Recomendaciones de negocio

Después de haber analizado el set de datos y también de haber hecho pruebas con los diferentes modelos, se consiguió alcanzar uno de regresión logística. Este modelo tenía una precisión del 81.82%. Para un modelo esto es bastante bueno y efectivo, sin embargo, al llevarlo a caso práctico, entonces se estaría sesgando a un número considerable de personas al interpretar de manera errónea una inexistencia de diabetes cuando realmente sí la padecen; esto es un problema crítico, pues al no controlar la diabetes o estar inconsciente de su existencia puede ocasionar graves daños a la salud, incluso llevando al paciente a una muerte prematura. Esto sin contar los problemas fuera del paciente, ya que, un diagnóstico erróneo ocasionado por la desinformación se puede interpretar como

negligencia médica; teniendo consecuencias más severas cuando el objetivo era ayudar a la visibilización de esta enfermedad.

A pesar de ello, se pueden obtener varias recomendaciones debido a la exploración hecha de los datos que no están encaminados a únicamente realizar un diagnóstico exitoso y seguro, sino que a visibilizar y elevar el nivel de conciencia o mantener más informadas a las personas cuyos hábitos alimenticios y su estilo de vida sea más propenso a poder padecer diabetes.

Por ejemplo, personas que no han sido diagnosticadas que tengan elevados niveles de glucosa en la sangre o índices de masa corporal altos deberían de hacerse pruebas o tomar cuidados preventivos ya que los valores mencionados anteriormente muestran que tendrán una mayor tendencia a ser diabéticos.

Conclusiones

Los resultados de este proyecto fueron positivos en cuanto al gran porcentaje de precisión de los modelos trabajados, sin embargo, no fue completamente satisfactorio ya que al tratarse de un diagnóstico de diabetes profesional un resultado equívoco negativo puede llegar a ser un punto de no retorno en la salud del paciente. Se alcanzó una precisión total de 81.82% utilizando un modelo de regresión logística. Aunque esto indica una gran posibilidad de eficacia, se considera que, nuevamente tratándose del campo médico, este modelo no es suficiente para declararlo como una medida estrictamente confiable al momento de evaluar la probabilidad de diabetes en las personas. Esto se sustenta con los resultados obtenidos en la **Figura 1**.

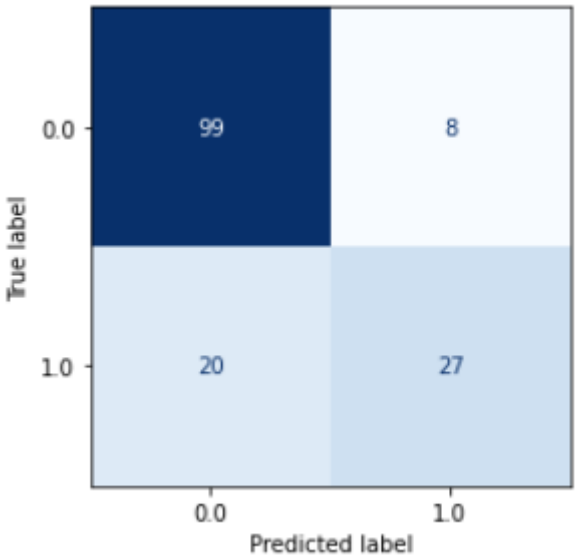


Figura 1. *Predicción del diagnóstico y presencia de diabetes*

Como se puede ver en la imagen, hay un total de 20 casos en los que la predicción de un resultado negativo para diabetes fue erróneo, siendo esto una falla en nuestro modelo por el número de registros en el dataset y convirtiéndose en uno de los puntos a mejorar. Por otro

lado, cabe destacar que fueron un total de 126 las predicciones que sí fueron acertadas, en comparación a las 28 que no lo fueron.

Los datos manejados se componen de 768 registros. Podría considerarse necesario implementar un número más elevado de registros para abordar un modelo de este tipo, no obstante, el dataset explorado fue bastante eficiente ya que no había presencia de datos nulos o faltantes y en general fue una fuente de información muy completa. Esto facilitó encontrar la relación entre las variables y saber cuáles influyen con mayor sustancia para determinar si hay presencia de diabetes. Por ejemplo, la columna de glucosa terminó siendo la variable más significativa para el diagnóstico de esta enfermedad. Además, se notó que otros factores como la edad contribuyen en menor medida al diagnóstico pero aun así era recurrente el hecho de que sí la persona tenía menos de 28 años era altamente probable que no tuviera diabetes.

En conclusión, este proyecto permitió explorar varios modelos para encontrar la solución a nuestra problemática, siendo la regresión logística la técnica más eficiente y con mayor precisión para dictaminar si es probable que una persona tenga diabetes con base en factores como la edad, BMI, insulina en la sangre, número de embarazos, glucosa, presión sanguínea, entre otros. Un modelo especializado como este, accesible al público en general, permitiría un chequeo médico profesional a tiempo para evitar problemas de salud a largo plazo. Una gran área de oportunidad para este proyecto podría ser la implementación de este análisis en una aplicación para dispositivos inteligentes que le solicitara al usuario determinados valores para calcular un porcentaje de probabilidad de diabetes aproximado. De esta manera, se impactaría la salud de muchas vidas y se podrían mejorar conductas y hábitos para alcanzar un estilo de vida más saludable.