María Camila Terán — 201822000
Juan Diego Cardona — 201819447
Nicolás Ortega — 201814515

Inteligencia de Negocios
Laboratorio 1

Tabla de contenidos

Análisis y perfilamiento de los datos

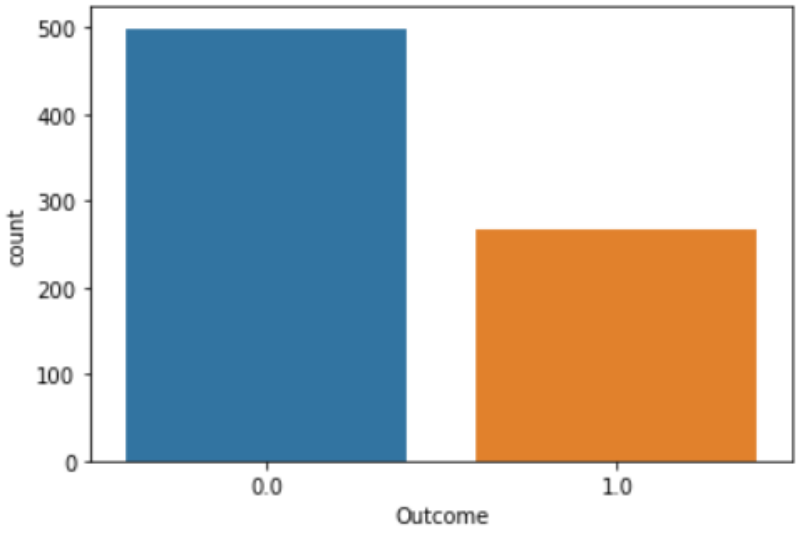
Para el análisis de datos, el primer paso fue analizarlos desde el archivo directamente. Se identificó la variable de interés "Outcome" la cual es binaria y determina si una persona tiene diabetes o no. Posteriormente se empezó a analizar el tipo de datos. Para iniciar se tiene la primera columna que hace referencia al color de pelo de una persona. Esta se llama "Hair Color" por lo que toma valores de tipo Object. De igual forma, de este tipo, se tiene la columna "City" pero al realizar el estudio, solo se tiene "Nueva York" por lo que se intuye que puede que esta no influya bastante en la variable de interés. Ahora bien, continuando con los valores numéricos, se tienen los embarazos de una persona. En esta columna hay cierta cantidad de ceros por lo que se puede concluir que, o es una mujer que no ha estado embarazada o que es hombre. En esta se observa un máximo de 17 lo cual tiene sentido de acuerdo con el diccionario proporcionado, a pesar de que en la vida cotidiana puede ser una gran cantidad. En cuanto a la columna que contiene los valores de glucosa, se observan valores por encima de 100 principalmente, aunque hay ciertos por debajo de este número por lo que no se consideraría una persona sana. No obstante, se observan algunos 0s que habrá que eliminar en la limpieza de datos pues todas las personas tienen cierto nivel de glucosa en la sangre. Similar al caso de la glucosa, está la variable de presión en la sangre la cual tiene un rango de 0 hasta 122 pero el mínimo no puede ser 0 porque el valor saludable va entre 80mm Hg hasta 120 mm Hg y aunque hay valores menores a 80, una persona debe tener este criterio. En cuanto al grosor de la piel, este también debería ser mayor a 0, y la mayoría de los valores van entre 10 y 30. Ahora bien, se tiene la columna de BMI que hace referencia al índice de masa corporal. Esta se calcula de la siguiente manera:

$$\frac{Peso\ (lb) x 703}{Altura\ (in^2)}$$

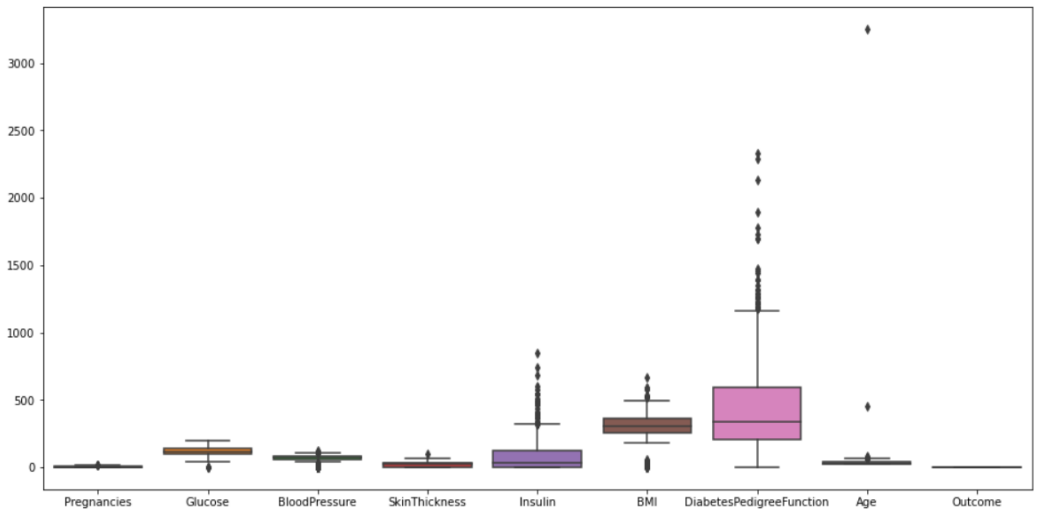
Por lo que es poco factible que sea 0. Además, la mayoría de los valores son superiores a 100. También se tiene la insulina, de valor numérico, pero está se concluyó que puede ser baja por lo que se incluye el 0. En cuanto al Diabetes Pedigree Function hace referencia a la historia clínica por lo que se aceptan los diferentes valores numéricos. En cuanto a la edad, se tiene que deben ser mayor de 21 pero hay ciertos valores atípicos como más de 3000, lo cual es imposible. Sin embargo, se realizó en Jupyter la siguiente tabla de estadísticas:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	767.000000	767.000000	767.000000	767.000000	767.000000	768.000000	767.000000	768.000000	767.000000
mean	3.839635	120.921773	69.096480	20.563233	79.903520	289.796875	432.395046	38.011719	0.349413
std	3.368429	31.984561	19.366833	15.945349	115.283105	116.757554	336.144934	117.825600	0.477096
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	251.750000	205.500000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	32.000000	309.000000	337.000000	29.000000	0.000000
75%	6.000000	140.500000	80.000000	32.000000	127.500000	359.000000	592.000000	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	671.000000	2329.000000	3256.000000	1.000000

Como se puede observar, hay cierta diferencia en los datos completos por lo que se evidencian nulos, los cuales se deben retirar en un futuro. Efectivamente se verifican los datos atípicos como que hay niveles de glucosa en 0 o una edad de 3256 años. Además, tenemos la siguiente representación visual de la variable de interés:



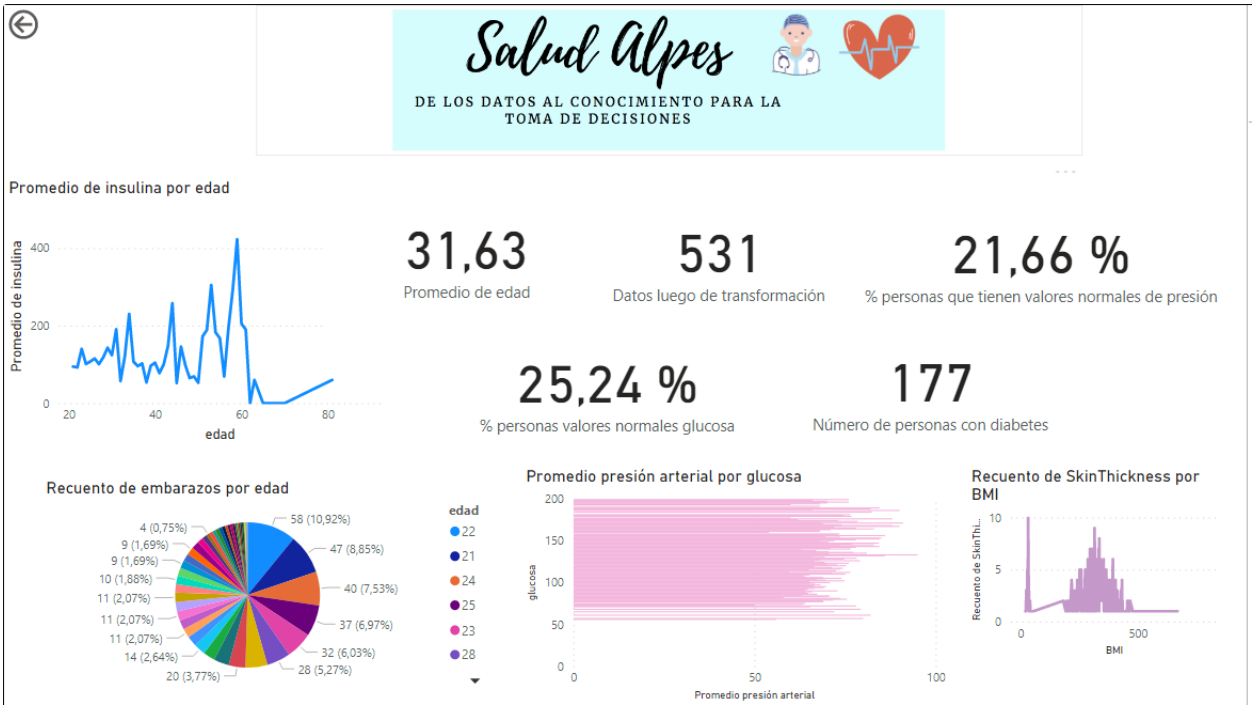
Como se puede observar la mayoría no sufren de diabetes, pero hay que realizar el estudio a profundidad para concluir la certeza de los datos. De forma visual, para hacer una perspectiva general de todas las variables, se realizó el siguiente gráfico:



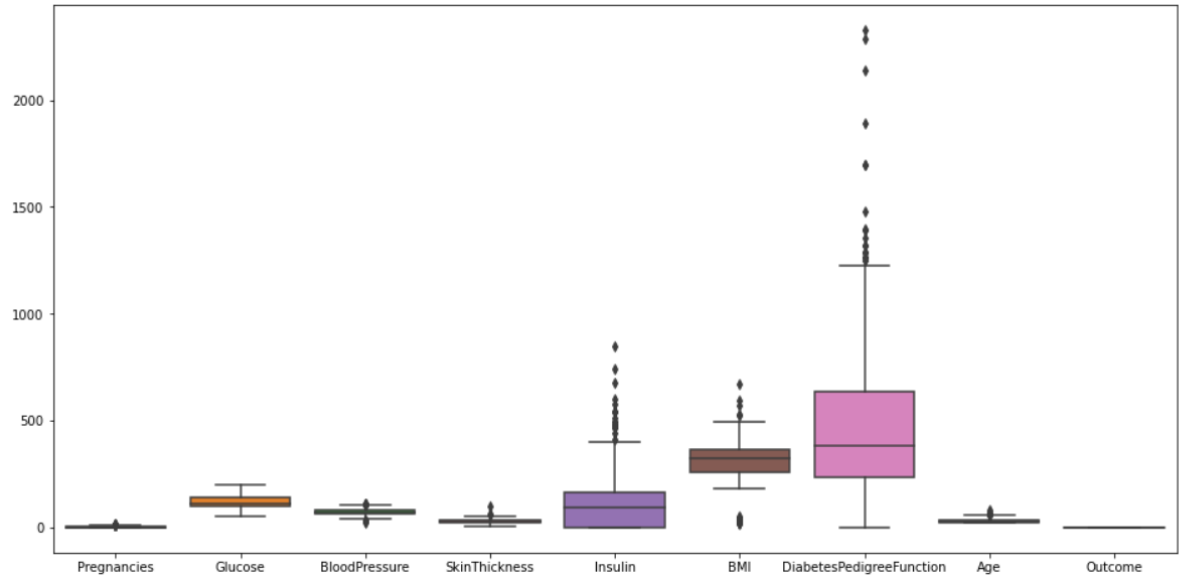
En este diagrama de cajas se puede ver que si hay diferencia entre variables por lo que unas pueden ser más significativas que otras. Las medias son bastante variadas, al igual que su dispersión. En ese sentido, se realizará la limpieza de datos.

Pre-procesamiento y limpieza de los datos

El primer paso para pre-procesar los datos fue revisar en PowerBi con el fin de hacer la transformación desde allí. Para ello, se utilizó la herramienta de "Transformar datos". Teniendo en cuenta el análisis realizado previamente, se deseleccionaron los ceros de Glucose, BMI, SkinThickness y BMI. Además, habían dos edades mayores a 100 que se deseleccionaron de igual manera. Ahora bien, se realizó el siguiente tablero de control. Utilizamos la función para generar nuevas medidas para obtener los porcentajes.



Con todo este primer paso de limpieza para el tablero, se realizó el mismo procedimiento en Jupyter. El primer paso fue quitar los datos que tienen nulos para que el total fuesen 767 datos en lugar de 768. Asimismo, se tuvieron ciertas restricciones teniendo en cuenta el análisis realizado previamente. La edad quedó en un rango de 21 a 100 por lo que allí se eliminaron dos filas. Ni la glucosa, ni el BMI, ni el grosor de la piel ni mucho menos el nivel de presión en la sangre pueden tener como valor 0. Además, se retiraron las columnas que hacen referencia a la ciudad pues solo tenía un valor y no iba a influenciar en si una persona tiene diabetes o no. La región en efecto puede ser significativa pero en este caso todos eran de NY. De igual manera, se eliminó el color de pelo pues era una variable categórica y no aporta información relevante al objetivo del caso de estudio. Luego de la limpieza, el diagrama de caja está de la siguiente manera:



Además, concuerda con lo establecido en el tablero de control pues ahora hay 531 datos como se ve a continuación:

531 rows × 9 columns

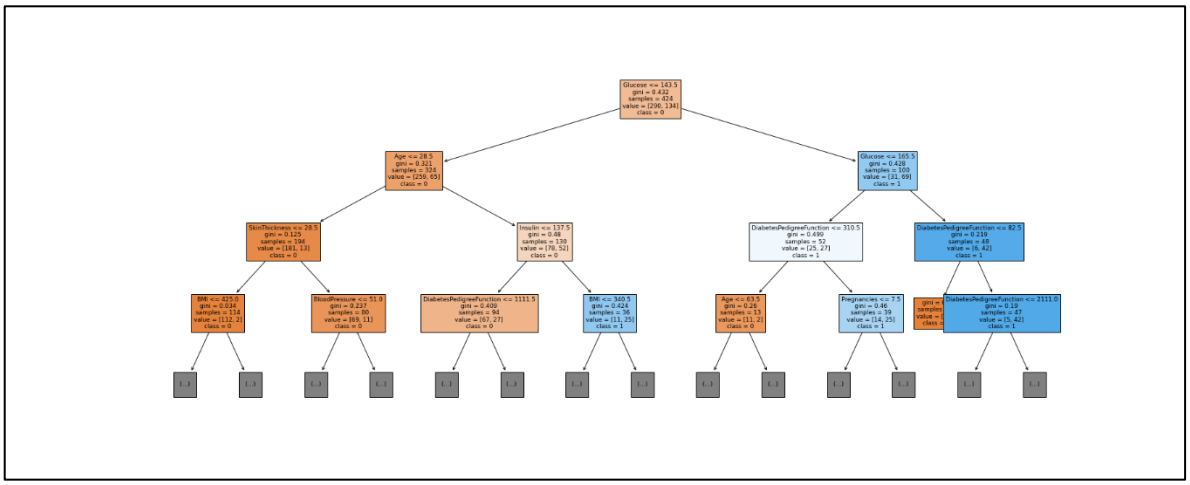
Clasificadores

Árbol de decisión (Nicolás Ortega)

Para la implementación del árbol de decisión fue necesario primero realizar el procesamiento de los datos mencionado anteriormente. Como los algoritmos de la librería utilizada solo permiten que los datos sean numéricos, se decidió no utilizar las columnas de "Ciudad" y "Color de pelo" en la construcción del modelo. Esto, en primer lugar, porque son variables categóricas, y en segundo lugar, porque no son proveen información útil que ayude a alcanzar el objetivo del caso de estudio. Por otro lado, se separó la columna de la variable objetivo (Outcome) de los datos para entrenar el modelo. Adicionalmente, se separaron los datos en un conjunto de entrenamiento del modelo y otro de conjunto de prueba.

En primera instancia se corrió el algoritmo para la construcción del árbol sin especificar hiperparámetros, pero para encontrar el mejor modelo luego fue necesario aplicar el algoritmo de K-Fold Cross Validation. Una vez realizado esto, se obtuvo el mejor modelo de árbol de clasificación con los siguientes hiperparámetros:

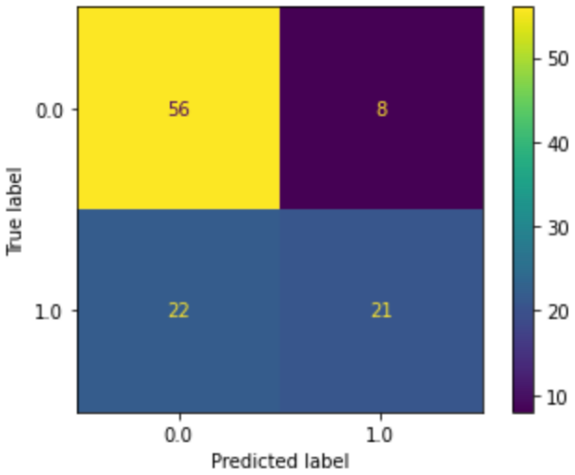
- criterion: Gini
- max_depth: 4
- min_samples_split: 2



Las métricas asociadas a este modelo fueron:

- **Exactitud**: 0.86
- **Recall**: 0.7014925373134329
- **Precisión**: 0.8173913043478261
- **Puntuación F1**: 0.7550200803212851

Se obtuvo también la matriz de confusión:



Se puede ver que hay unos pocos falsos positivos, lo que significa que el modelo es preciso, sin embargo, hay una cantidad considerable de falsos negativos, que el contexto del negocio es un poco preocupante puesto que significaría dar un diagnóstico errado a un paciente con diabetes. Este modelo puede ser útil para predecir si un paciente sí tiene diabetes con un alto nivel de confianza, pero cuando la respuesta es negativa, sería mejor realizar más exámenes y tener cuidado para no dar un diagnóstico equivocado.

Finalmente, se realizó un análisis de la importancia de cada variable sobre la decisiones de clasificación con las que se construyó el árbol. Se obtuvo que la variable más importante fue el nivel de Glucosa:

	Atributo	Importancia
0	Glucose	0.446536
1	Age	0.195729
2	DiabetesPedigreeFunction	0.132512
3	Insulin	0.089814
4	BMI	0.061063
5	BloodPressure	0.032868
6	Pregnancies	0.027403
7	SkinThickness	0.014075

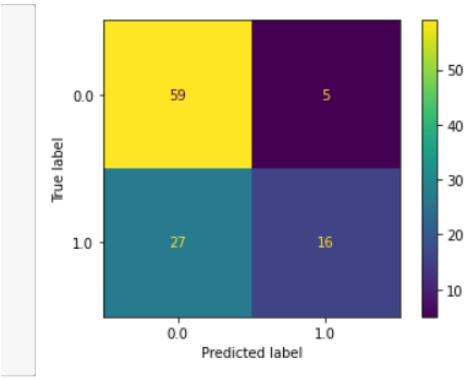
K-nearest neighbours (Maria Camila Terán)

El KNN es un método de clasificación de datos de forma supervisada. A diferencia del árbol de decisión, este se basa en el "espacio" en donde se encuentran diversos datos. Por ejemplo, un dato se asigna a una clase dependiendo de los K elementos más cercanos, de allí el nombre. Para esto se suele utilizar la distancia euclidiana, la cual, se ve reflejada en la siguiente ecuación:

$$D(x,y) = \sqrt[2]{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Ahora bien, en cuanto a la implementación, lo primero es seleccionar la variable objetivo, en este caso es "Outcome". Posteriormente, es necesario dividir los datos en entrenamiento y prueba. Cabe resaltar que los datos de entrenamiento son aquellos datos previamente clasificados y observados con intervención humana. Para continuar, se realizó un primer intento con k=3. Se decidió utilizar k=3 ya que es mejor utilizar un número impar para que no se tenga el problema de que hay dos vecinos igual de cercanos. Además, se eligen 3 vecinos porque se desea eficiencia y de esta manera puede clasificar de forma más rápida.

Asimismo, se realiza la matriz de confusión, con la que se obtuvieron los siguientes resultados:



Efectivamente hay pocos falsos positivos, pero hay 27 falsos negativos lo que indica error tipo 2 relacionado con la sensibilidad. Esto es preocupante pues esta indicando que personas que no tienen diabetes si la padecen. Sin embargo, la mayoría de los datos están en verdaderos positivos y verdaderos negativos, pero hay que mirar cómo hacer mejoras.

Luego se obtuvo el reporte de clasificación, en el que se obtuvieron los siguientes resultados:

	precision	recall	f1-score	support
0.0	0.69	0.92	0.79	64
1.0	0.76	0.37	0.50	43
accuracy			0.70	107
macro avg	0.72	0.65	0.64	107
weighted avg	0.72	0.70	0.67	107

Sin embargo, se consideró que el modelo podía tener mejoras dado los resultados obtenidos en la matriz, por lo que, luego de normalizar los datos, se obtuvieron los siguientes resultados con los datos de entrenamiento:

- **Exactitud:** 0.86
- **Recall:** 0.6940298507462687
- **Precisión:** 0.8378378378378378
- **Puntuación F1:** 0.759183673469387

Elección Libre: Regresión logística (Juan Diego Cardona)

Para el tercer clasificador, elegimos el modelo de regresión logística. Este modelo es un algoritmo de clasificación. Es similar a una regresión lineal pero adicionalmente nos permite modelar la relación de una variable continua con otras variables independientes por medio de un ajuste en la ecuación lineal.

La definición típica para este tipo de regresiones es:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i$$

A partir de esta ecuación, los elementos más importantes para interpretar son β_0 y β_j.

β_0 es el intercepto y corresponde al valor esperado de la variable respuesta cuando los demás son 0. Para el caso de estudio particular nuestra variable respuesta es la columna “Outcome” la cual nos determina si la persona tiene o no tiene diabetes.

Por otro lado, los β_j son los coeficientes que indican el promedio esperado de la variable respuesta cuando incrementa una unidad de las demás variables. Para nuestro caso estas son BloodPressure, pregnancias, glucose, skinThickness, Insulin, BMI, DiabetesPedigree Function y Age.

En primera instancia se corrió el algoritmo sin especificar hiperparámetros, pero para encontrar el mejor modelo, luego fue necesario aplicar el algoritmo de solver 'lbfgs' el cual es un método de optimización quasi-newtoniana con un gran numero de parámetros u complejidad. Realizando este algoritmo también se tuvo que ampliar el máximo de iteraciones a 1000.

Una vez realizado el modelo, se obtuvieron los siguientes resultados acerca del intercepto, los coeficientes y la precisión:

```
Intercept: [-8.35202895]
Coeficiente: [('Pregnancies', 0.0917974261313487), ('Glucose', 0.03872156718053289), ('BloodPressure', -0.01072564647035161),
('SkinThickness', 0.041945963713283387), ('Insulin', -0.0011031177126530143), ('BMI', 0.016294168432222045), ('DiabetesPedigre
eFunction', 0.0015176320018447478), ('Age', 0.02643433001133034)]
Accuracy: 0.7871939736346516
```

A continuación se pueden ver las métricas del modelo obtenido:

	precision	recall	f1-score	support
0.0	0.79	0.89	0.84	87
1.0	0.72	0.57	0.63	46
accuracy			0.77	133
macro avg	0.76	0.73	0.74	133
weighted avg	0.77	0.77	0.77	133

A continuación, se procedió a hacer predicciones probabilísticas para determinar la probabilidad predicha de pertenecer a cada una de las dos clases (Outcome positivo o negativo) según cada uno de los parámetros. La siguiente tabla muestra dichos resultados:

	0.0	1.0
0	0.833012	0.166988
1	0.473739	0.526261
2	0.955230	0.044770
3	0.842753	0.157247
4	0.587482	0.412518
5	0.822638	0.177362
6	0.907326	0.092674
7	0.499198	0.500802

Con esta información, se puede determinar que las variables Glucosa, Insulina y la Edad son las variables pues tienen mayor relación o probabilidad de generar diabetes.

Adicionalmente a esto, procedemos a realizar un modelo de regresión logística más completo con base a una matriz de predictores:

```
Optimization terminated successfully.
         Current function value: 0.442070
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                Outcome   No. Observations:                  398
Model:                          Logit   Df Residuals:                      389
Method:                           MLE   Df Model:                            8
Date:                Sat, 04 Sep 2021   Pseudo R-squ.:                  0.3023
Time:                        17:48:52   Log-Likelihood:                -175.94
converged:                       True   LL-Null:                       -252.16
Covariance Type:            nonrobust   LLR p-value:                 6.084e-29
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          -8.3515      1.040     -8.030      0.000     -10.390      -6.313
Pregnancies     0.0920      0.050      1.832      0.067      -0.006       0.190
Glucose         0.0387      0.005      7.080      0.000       0.028       0.049
BloodPressure  -0.0107      0.012     -0.918      0.359      -0.034       0.012
SkinThickness   0.0420      0.015      2.811      0.005       0.013       0.071
Insulin        -0.0011      0.001     -0.948      0.343      -0.003       0.001
BMI             0.0016      0.001      1.160      0.246      -0.001       0.004
DiabetesPedigreeFunction     0.0015      0.000      3.736      0.000       0.001       0.002
Age             0.0264      0.016      1.657      0.097      -0.005       0.058
==============================================================================
```

Este modelo puede evidenciar los coeficientes de cada variable, su desviación estándar y otros valores como la distribución Z e intervalos de confianza. Todos estos valores refuerzan los hiperparámetros seleccionados.

Finalmente se procedió a realizar los intervalos de confianza del modelo para cada una de las variables:

	2.5%	97.5%
const	-10.390024	-6.313075
Pregnancies	-0.006428	0.190482
Glucose	0.028003	0.049441
BloodPressure	-0.033633	0.012177
SkinThickness	0.012705	0.071200
Insulin	-0.003383	0.001176
BMI	-0.001123	0.004381
DiabetesPedigreeFunction	0.000721	0.002314
Age	-0.004824	0.057614

Análisis de resultados, Comparación de los modelos y recomendación

De todos los modelos realizados se calcularon las métricas de desempeño de exactitud, precisión, sensibilidad y el score F1. A continuación, encontramos el resumen de los valores obtenidos en cada modelo:

Clasificador	Exactitud	Sensibilidad	Precisión	F1 score
Árbol	0.86	0.7	0.82	0.75
KNN	0.86	0.69	0.83	0.75
R. Logística	0.77	0.73	0.77	0.77

Los modelos de Árbol de decisión y KNN son más exactos que el de Regresión Logística. Esto significa no presentan demasiados falsos positivos ni negativos, en otras palabras, dan un diagnóstico más acertado al momento de predecir. El modelo con mayor precisión es el de KNN, aunque no difiere mucho de los otros dos. El que presenta el mayor F1 score es el de Regresión Logística, que como el caso anterior, no es un desempeño muy diferente al de los otros modelos. En cuanto a sensibilidad, el que obtuvo mayor valor fue el también el de Regresión Logística, lo que nos lleva a pensar que es mejor a la hora de predecir los diagnósticos negativos.

Se puede observar que los modelos en general tienen un desempeño similar, las métricas nos muestran que no hay una diferencia demasiado significativa entre ellos. En general todos funcionan adecuadamente y pueden ser útiles para predecir el diagnóstico de los pacientes, sin embargo, como ya se ha mencionado antes, toca tener especial cuidado con los resultados falsos positivos. La idea es que el modelo que se vaya a utilizar se integral y funcione como una herramienta confiable que apoye a la empresa de SaludAlpes. Por esto, si tuviéramos que dar una recomendación de un modelo a SaludAlpes, recomendaríamos utilizar el **Árbol de Decisión**, ya que presenta buenas métricas generales (tiene el mejor promedio de métricas): tiene un alto nivel de exactitud, también tiene un buen nivel de precisión, y su confiabilidad de predecir resultados negativos es aceptable, no difiere significativamente del modelo que lo hace mejor, que es el de Regresión.