Laboratorio 1

Laboratorio 1 - Clasificación

Autores

Pablo Arango	Juan Diego Trujillo	Santiago Moreno
201630495	201618006	201814353
p.arango	jd.trujillom	s.morenom

Inteligencia de Negocios

Sección 2

Departamento de Ingeniería de Sistemas y Computación

Universidad de Los Andes

Bogotá, Colombia.

Perfilamiento de datos

En primer, comenzamos perfilando el archivo que fue entregado para este laboratorio. En este encontramos la siguiente de cantidad de filas y columnas.

Número de filas	Número de columnas
768	11

Luego de ello, era de interés averiguar cuales eran los tipos de los atributos cargados en el archivo. Sin embargo, una vez cargado pudimos ver que muchos de estos valores eran de tipo *object,* probablemente porque muchos de ellos se cargaron como cadenas de caracteres. Esto se puede observar en la figura 1.

```
Hair color                object
Pregnancies               object
Glucose                   object
City                      object
BloodPressure             object
SkinThickness             object
Insulin                   object
BMI                        int64
DiabetesPedigreeFunction  object
Age                        int64
Outcome                   object
dtype: object
```

Figura 1. Tipos una vez cargado el archivo

```
Hair color                object
Pregnancies                int32
Glucose                    int32
City                      object
BloodPressure              int32
SkinThickness              int32
Insulin                    int32
BMI                        int64
DiabetesPedigreeFunction   int32
Age                        int64
Outcome                   object
dtype: object
```

Figura 2. Tipos luego de procesamiento

Luego de hacer cierto procesamiento, que se va a explicar más adelante, se obtuvieron los tipos que se pueden observar en la figura 2. En la siguiente tabla se puede observar la distribución de estos valores.

Tipo de columna	Cantidad
Numérico	9
categóricas	2

La variable que es de nuestro interés es la que está en la columna *outcome*, cuyo significado es el siguiente:

Outcome	Significado
0	No tiene diabetes
1	Tiene diabetes.

Es de interés observar que tan balanceado se encuentran los datos. Por lo tanto, se decidió graficar la distribución de valores y la figura 3 muestra estos resultados.

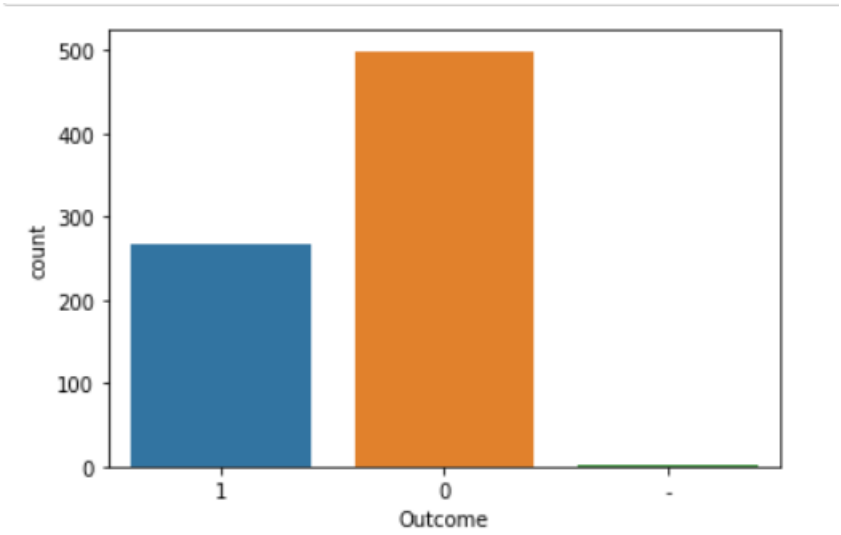

Figura 3. Distribución de valores de la variable objetivo

En efecto, los datos están desbalanceados, pues hay más personas que no tienen diabetes que las que sí tienen está enfermedad.

Sin diabetes	Con diabetes
499	268

De igual forma, era importante detallar algunos estadísticos sobre los datos numéricos. La figura 4 los muestra.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000
mean	3.839635	121.717080	72.382008	20.563233	79.903520	289.670143	432.395046	33.260756	0.349413
std	3.368429	30.445723	12.103830	15.945349	115.283105	116.780873	336.144934	11.746998	0.477096
min	0.000000	44.000000	24.000000	0.000000	0.000000	0.000000	1.000000	21.000000	0.000000
25%	1.000000	99.500000	64.000000	0.000000	0.000000	251.500000	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	81.000000	1.000000

Figura 4. Estadísticos importantes sobre datos numéricos.

Preprocesamiento

Antes de que cada uno de los integrantes del grupo empezara a desarrollar cada uno de los modelos, se decidió hacer un procesamiento general de los datos. Se hicieron múltiples cambios para mantener la consistencia con el diccionario entregado y para permitir la creación de los modelos de clasificación. Entre los cambios realizados a las columnas fueron:

- En caso de encontrar una instancia de uno de los atributos que no es consistente con el resto, en todas las columnas menos “Outcome” o una columna categórica como “Ciudad”, se eliminaban de forma temporal del dataframe, se encontraba el valor promedio de los elementos de la columna y luego se reemplaza el valor inconsistente con el valor promedio. Esto se realizó, por ejemplo, en la columna “BloodPressure” y la columna “Age”. En el caso de la primera, se encontraron valores en 0, lo cual va en contra a su definición en el diccionario. En la segunda, se encontraron edades de 450 y más de 3 mil años entre los datos, lo cual no es un dato posible, al menos refiriéndose a las edades de humanos, por lo que se cambió su valor por el promedio.

- La fila que contenía un “-” como el valor de la columna “Outcome” fue removida. Esto se hizo porque no era consistente con el resto de los datos de la columna y de que, dado que la columna “Outcome” es una label calculada, no podemos decidir su valor a partir del promedio de los datos presentes en la columna. Por esto se decidió eliminarla.

Figura 5. Tablero de control que incluye los niveles de glucosa

La figura 5 muestra uno de los tableros de control creados que incluía la glucosa promedio, la glucosa promedio por edad, la edad promedio y el número de datos ingresados.

Figura 6. Tablero de control que incluye las cantidades de insulina

De manera similar, se realizó un segundo tablero de control que mostraba la cantidad de insulina por edad y se describe en la figura 6.

Árbol de decisión: Santiago Moreno

En cuanto a las columnas del dataframe, se decidió no utilizar las siguientes en el modelo:

- *"Hair color"* y *"Cities",* dado que son atributos no numéricos y que no consideramos como relevantes en el cálculo de la Diabetes.
- *"Outcome"*, dado que esta será utilizada como nuestra utilizada como nuestra variable objetivo, se separó del conjunto de datos utilizados en el modelaje.

En el caso de "Pregnancies" y "SkinThickness" se eliminaron después de realizar pruebas y comparando los resultados del modelo, por lo que se justificará más adelante en el documento.

KNN: Pablo Arango

De forma similar a los arboles de decisión, se decidió no usar las columnas *Hair Color, Cities* y *Outcome* por las razones antes dadas.

Para el resto de las variables numéricas, se hizo una limpieza donde se reemplazaron datos que no tenían sentido, como valores de 0 en columnas donde tenían que ser mayores a 0 o valores de "-" que implicaban un dato mal registrado. Además de eso, en la columna de "Age" se eliminaron datos como 450 y 3256, los cuales eran valores ilogicos. "Pregnanncies" y "SkinThickness" si se utilizaron en este caso.

Regresión logística: Juan Diego Trujillo

De forma similar al modelo de los arboles de decisión y KNN, se omiten las columnas cuyos valores son categóricos.

Dado que al momento de desarrollar este modelo ya se habían mostrado los resultados de KNN y arboles de decisión, se decidió evaluar si las columnas "Pregnancies" y "SkinThickness" eran relevantes para ejecutar el modelo. Al hacer varias evaluaciones, se concluyó que para cada una de las métricas, los valores fluctuaba en menos del 1% en cada uno de ellas, por lo que se decidió no tenerlas en cuenta para que ejecutar los modelos.

Implementación y Resultados:

Árbol de decisión

La implementación se llevó a cabo primero generando un modelo sin hiper-parámetros, el cual se analizó y se usó como base, para luego manipular los hiper-parámetros utilizando *K-Fold Cross Validation* y encontrarse así los mejores parámetros.

Hiper-parametro	**Valor**
Criterion	*Entropy*
Max_depth	10
Min_samples_split	5

Estos valores mejoraron los resultados en gran manera, en comparación al primer árbol generado. Tal modelo fue el mejor de los ejecutados con árboles de decisión ya que resultó en una precisión del 77% y una exactitud del 81%. Además, los verdaderos positivos fueron 94 y los verdaderos falsos fueron 27.

En la siguiente tabla se presentan los resultados obtenidos en las métricas de evaluación del modelo.

Exactitud	Recall	Precision	F1
81%	59%	71%	64%

La figura 7 presenta la matriz de confusión obtenida para este modelo de árboles de decisión.

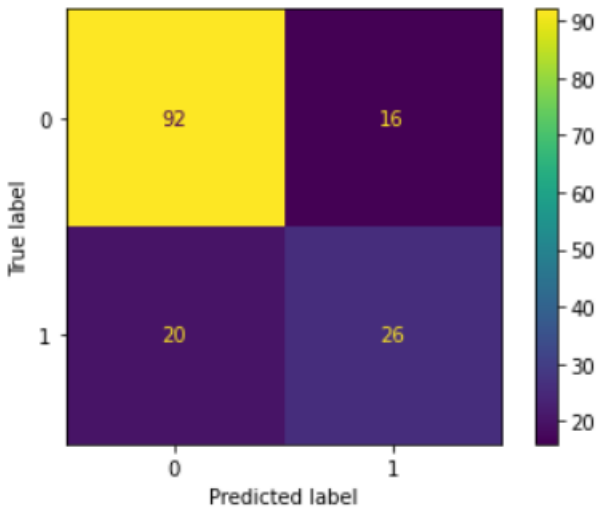

Figura 7. Matriz de confusión para el modelo de arboles

KNN

Para la implementación del KNN primero se construyó el modelo con los datos corregidos. Se escogió arbitrariamente el valor de K=4 para poder realizar una primera visualización del modelo. Tal ejercicio arrojó una precisión de 44% y una exactitud de 68%, además de 94 verdaderos positivos y 11 verdaderos negativos. Después se procedió a normalizar los datos para ver cómo se comportaba el modelo. La normalización de los datos, utilizando el mismo número de vecinos y la distancia Manhattan, dio un resultado mejor al que sin normalizar los datos, con una exactitud del 77% y una precisión del 65%

Posteriormente se hizo el procedimiento de los hiper-parametros para identificar el mejor valor de parámetros para los modelos. En cuanto a los datos sin ser normalizados, los mejores parámetros fueron un número de 7 vecinos y la distancia de minkowski. Tal modelo, para los datos no normalizados, arrojó una exactitud del 79%, una precisión del 88%, 89 verdaderos positivos y 17 verdaderos falsos. Finalmente, se hizo el proceso de hiper-parametros pero con los valores de los datos normalizados. Tal proceso reflejó que el numero óptimo de vecinos es k=4 y la distancia Minkowski. Utilizando ese modelo, la exactitud fue del 80%, una precisión del 65%, 97 verdaderos positivos y 21 verdaderos falsos. Por tal motivo, la mejor versión del modelo fue con k=4, la distancia Minkowski y con los datos normalizados.

En la siguiente tabla se presentan los resultados obtenidos en las métricas de evaluación del modelo.

Exactitud	Recall	Precision	F1
77%	45%	65%	53%

La figura 8 presenta la matriz de confusión obtenida para este modelo KNN.

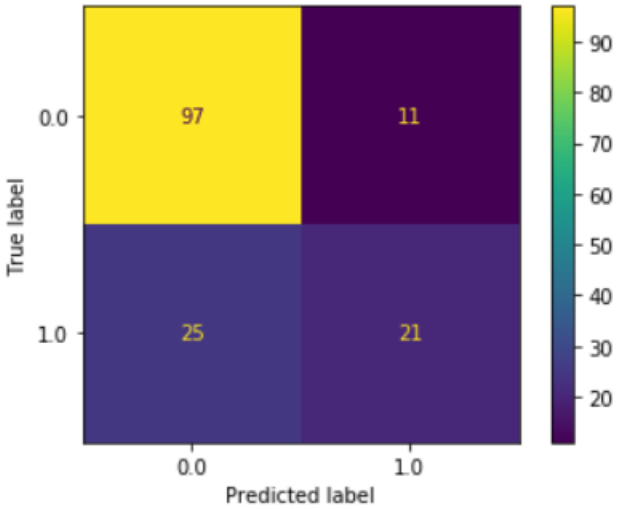

Figura 8. Matriz de confusión para el modelo de arboles

Regresión logística.

Finalmente, para la implementación de este modelo se usaron los datos corregidos, tal como se hizo en los otros. Para una primera visualización del modelo, se usaron los hiperparametros por defecto de utiliza la librería. Luego, se utilizó *K-Fold Cross Validation* para identificar los mejor hiperparametros y esto fue lo que se encontró.

Hiper-parametro	Valor
C	*1*
Penalty	L1
Solver	liblinear

En la siguiente tabla se presentan los resultados obtenidos en las métricas de evaluación del modelo.

Exactitud	Recall	Precision	F1
83%	56%	81%	66%

La figura 9 presenta la matriz de confusión obtenida para este modelo de regresión logística.

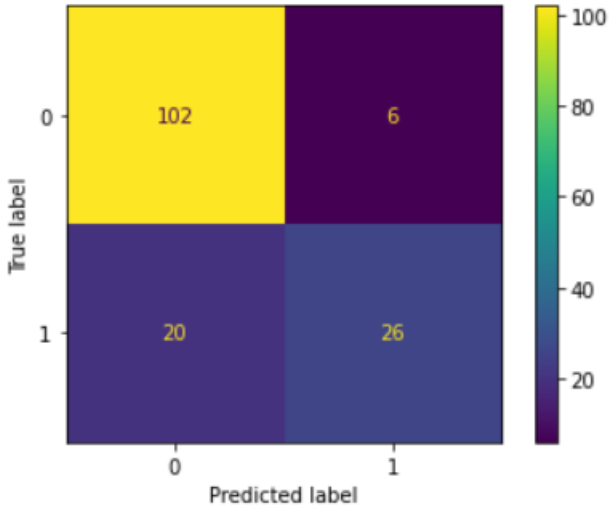

Figura 9. Matriz de confusión para el modelo de regresión logística.

Comparación de Modelos

Con base a los resultados obtenidos por los distintos modelos que se implementaron, se pudo llegar a la conclusión de que el modelo que arrojó los mejores resultados fue el de regresión logística, al menos en cuanto a la exactitud, precisión y F1. La siguiente tabla presenta dicha comparación.

Modelo	**KNN**	**Árbol Decisión**	**Regresión Logística**
Exactitud	77%	81%	**83%**
Recall	45%	**59%**	56%
Precisión	65%	71%	**71%**
F1	53%	64%	**66%**

Sin embargo, es importante hacer notar que el modelo de árbol de decisión tiene el mayor *recall*, lo cual es relevante porque en este problema deseamos minimizar los falsos negativos; es decir, aquellas personas que predijimos que no son diabéticas, pero en realidad si lo son. En el caso contrario no es tan grave, pues si decimos que alguien es diabético y en realidad no lo es, basta con que la saludAlpes haga un chequeo y se de cuenta que este paciente no es diabético. No es diabético y no lo predije, puede que esa persona tenga problemas de salud graves al no recibir un tratamiento a tiempo.

No obstante, la diferencia entre los valores de recall para el modelo de árbol y de regresión es solo de tres puntos porcentuales, mientras que el modelo de regresión tiene mejores métricas de desempeño en todas las demás. **En conclusión, nuestro grupo le recomendaría a SaludAlpes escoger el modelo de Regresión Logística.**