LABORATORIO 1

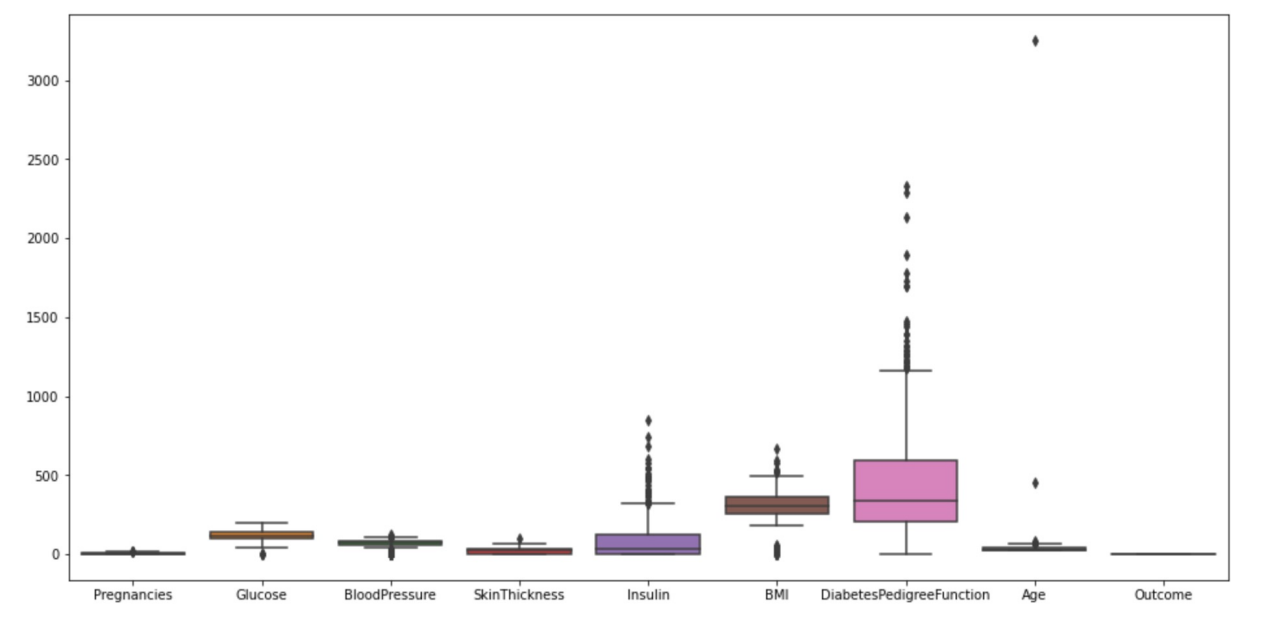
Inteligencia de Negocios

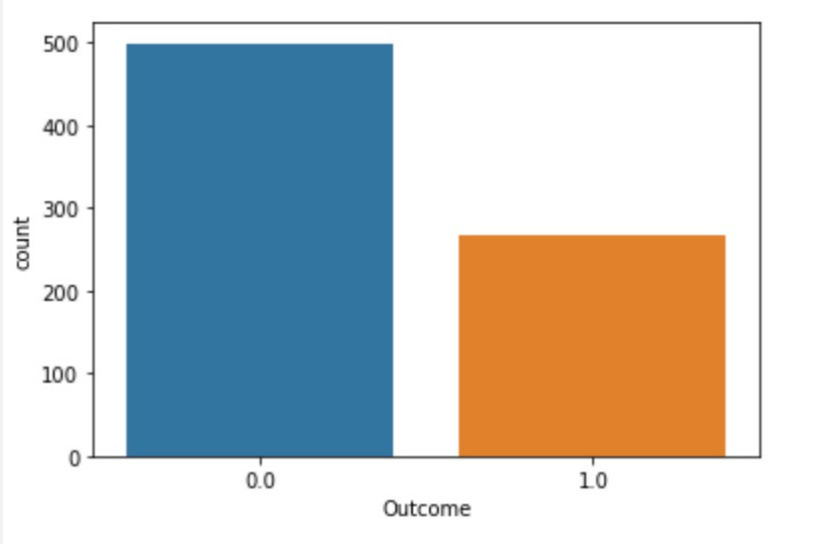
María Camila Terán — 201822000

Juan Diego Cardona — 201819447

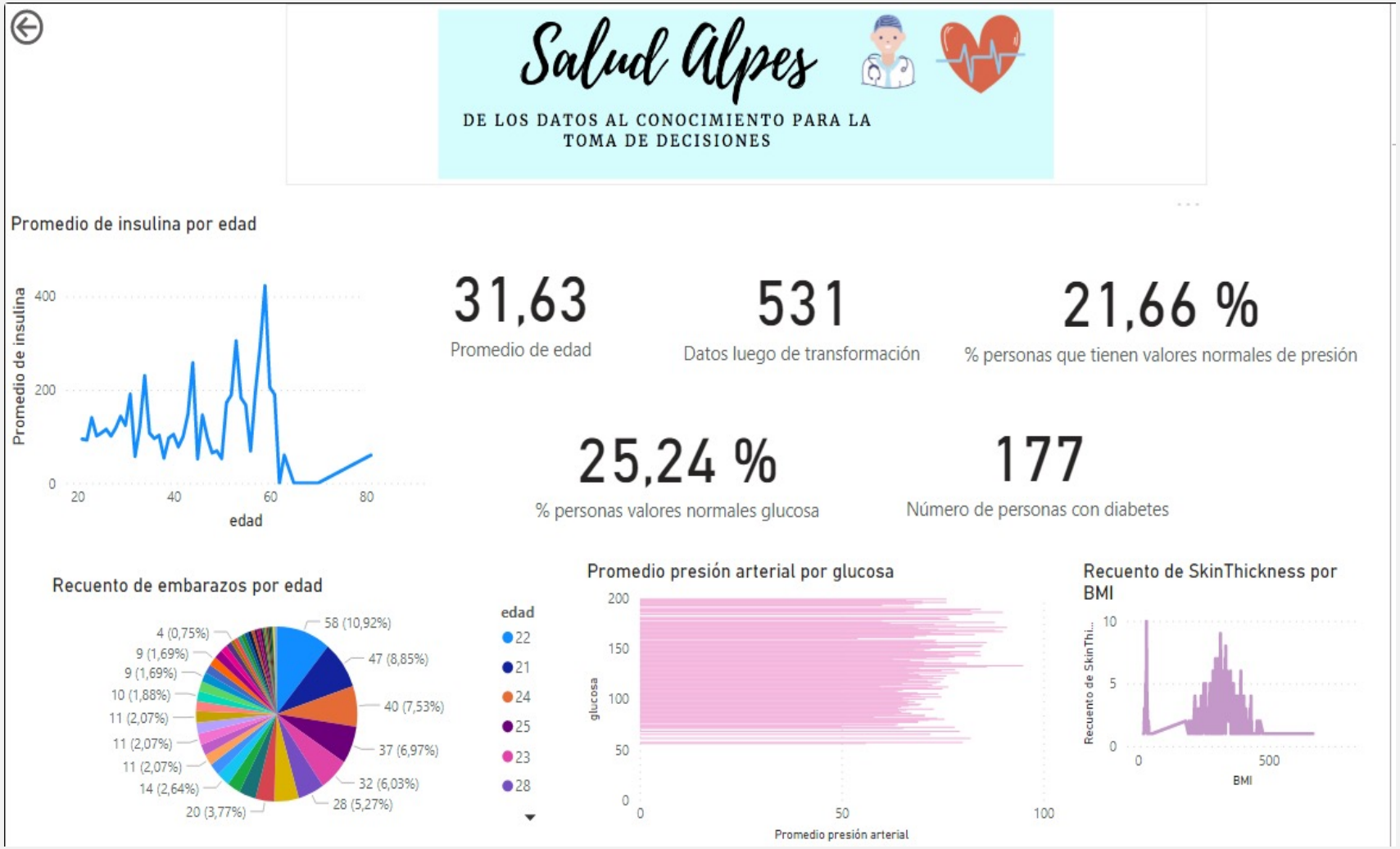
Nicolás Ortega — 201814515

ANÁLISIS Y PERFILAMIENTO DE DATOS





ANÁLISIS Y PERFILAMIENTO DE DATOS



PRE-PROCESAMIENTO Y LIMPIEZA DE DATOS

```
#Restricciones

#Edad menor a 100 y mayor a 21
datoslimpios = datoslimpios[datoslimpios.Age < 100]
datoslimpios = datoslimpios[datoslimpios.Age >= 21]

#BMI no puede ser 0
datoslimpios = datoslimpios[datoslimpios.BMI >0]

#Glucosa no puede ser 0
datoslimpios = datoslimpios[datoslimpios.Glucose >0]

#BloodPressure no puede ser 0
datoslimpios = datoslimpios[datoslimpios.BloodPressure >0]

#SkinThickness no puede ser 0
datoslimpios = datoslimpios[datoslimpios.SkinThickness> 0]

#Quitar color de pelo y ciudad
datoslimpios = datoslimpios.drop(['HairColor'], axis=1)
datoslimpios = datoslimpios.drop(['City'], axis=1)

datoslimpios
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.0	35.0	0.0	336	627.0	50	1.0
1	1.0	85.0	66.0	29.0	0.0	266	351.0	31	0.0
3	1.0	89.0	66.0	23.0	94.0	281	167.0	21	0.0
4	0.0	137.0	40.0	35.0	168.0	431	2288.0	33	1.0
6	3.0	78.0	50.0	32.0	88.0	31	248.0	26	1.0
...	...	...	...	...	...	...	...	...	...
761	9.0	170.0	74.0	31.0	0.0	44	403.0	43	1.0
763	10.0	101.0	76.0	48.0	180.0	329	171.0	63	0.0
764	2.0	122.0	70.0	27.0	0.0	368	34.0	27	0.0
765	5.0	121.0	72.0	23.0	112.0	262	245.0	30	0.0
767	1.0	93.0	70.0	31.0	0.0	304	315.0	23	0.0

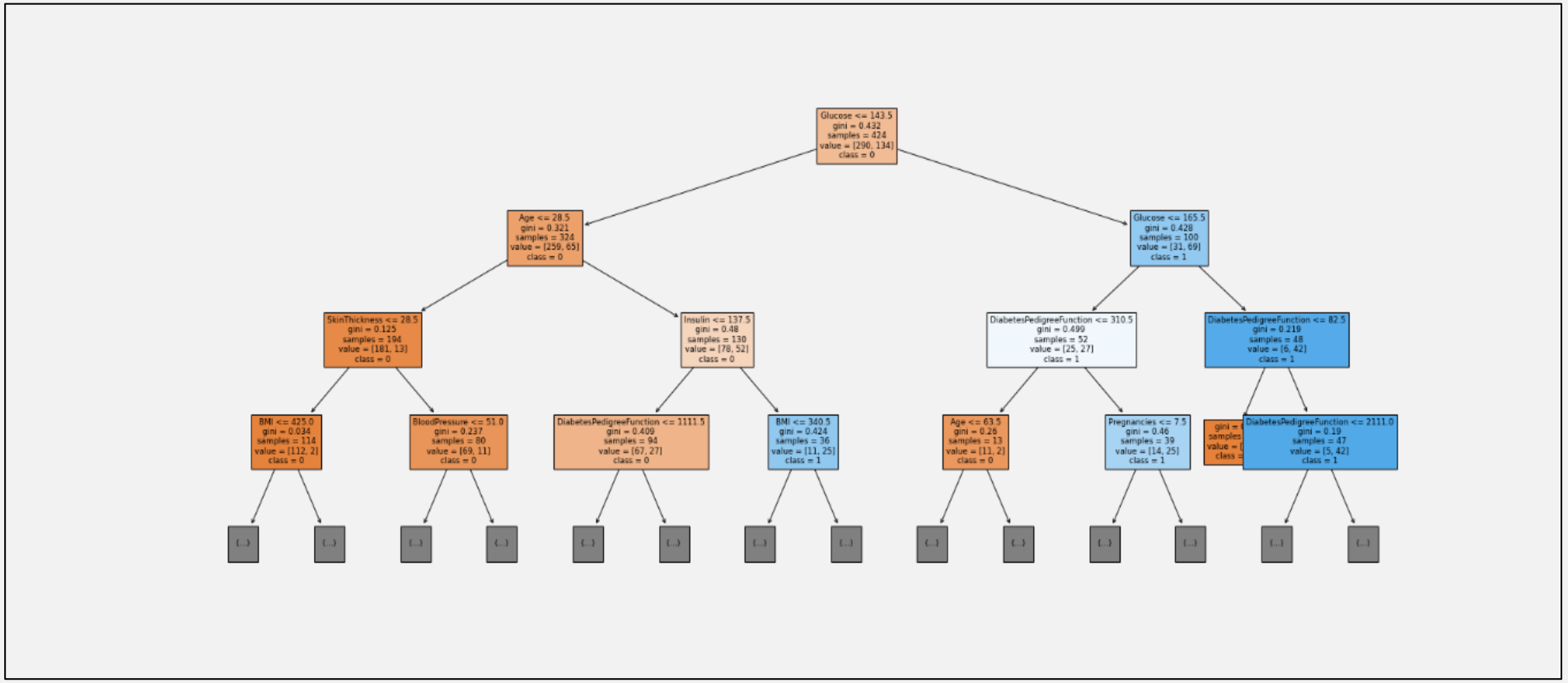
CLASIFICADORES

Árbol de decisión

K- Nearest Neighbours

Regresión Logística

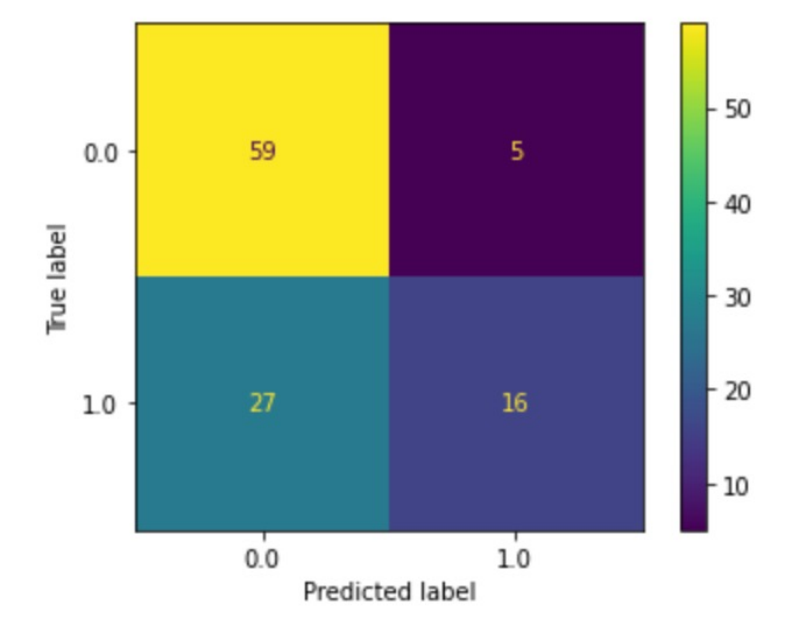
ÁRBOL DE DECISIÓN



	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

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

K-NEAREST NEIGHBOURS



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	\
0	0.352941	0.643357	0.558140	0.304348	0.000000	0.485407	
1	0.058824	0.202797	0.488372	0.239130	0.000000	0.377880	
3	0.058824	0.230769	0.488372	0.173913	0.111111	0.400922	
4	0.000000	0.566434	0.186047	0.304348	0.198582	0.631336	
6	0.176471	0.153846	0.302326	0.271739	0.104019	0.016897	
..	...	...	...	...	...	...	
761	0.529412	0.797203	0.581395	0.260870	0.000000	0.036866	
763	0.588235	0.314685	0.604651	0.445652	0.212766	0.474654	
764	0.117647	0.461538	0.534884	0.217391	0.000000	0.534562	
765	0.294118	0.454545	0.558140	0.173913	0.132388	0.371736	
767	0.058824	0.258741	0.534884	0.260870	0.000000	0.436252	

	DiabetesPedigreeFunction	Age	Outcome
0	0.268900	0.483333	1.0
1	0.150344	0.166667	0.0
3	0.071306	0.000000	0.0
4	0.982388	0.200000	1.0
6	0.106100	0.083333	1.0
..	...	...	...
761	0.172680	0.366667	1.0
763	0.073024	0.700000	0.0
764	0.014175	0.100000	0.0
765	0.104811	0.150000	0.0
767	0.134880	0.033333	0.0

	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

REGRESIÓN LOGÍSTICA

```
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:                        19:01:59   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
================================================================================
```

	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

	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

ANÁLISIS DE RESULTADOS, COMPARACIÓN DE MODELOS Y RECOMENDACIÓN

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

Nuestra recomendación de un modelo a SaludAlpes es utilizar el Árbol de Decisión.