ANÁLISIS DE DATOS

DIABETES

Delgado Gómez Karla Patricia

Mayo 2024

Datos

Sex

Indica el género del paciente

Age

Indica la categoría de edad

HighChol

Indica si el paciente ha tenido colesterol alto

HeartDiseaseorAttack

Indica si se ha padecido una enfermedad coronaria o infarto de miocardio.

CholCheck

Indica si se ha realizado un chequeo de colesterol.

BMI

Índice de masa corporal

Smoker

Indica si el paciente ha fumado al menos 100 cigarros (5 paquetes) en toda su vida

PhysActivity

Indica si se realizó actividad fisca en los últimos 30 días.

Fruits

Indica si se consume 1 fruta o más al día

Veggies

Indica si se consume 1 o más verduras al día

GenHlth

Indica en un rango de 1 a 5 cómo considera el paciente su salud

HvyAlcoholConsump

Indica si se los hombres adultos consumen >=14 tragos por semana y mujeres adultas >=7 tragos por semana

MentHlth

Indica los días de mala salud mental en una escala de 1 a 30 días

DiffWalk

Indica si se tiene problemas para caminar o subir escaleras.

Stroke

Indica si se ha padecido de un derrame cerebral

PhysHlth

Indica los días en los que se sufrió enfermedad o lesión física en una escala de 1 a 30 días

HighBP

Indica si ha sufrido de presión arterial alta.

Diabetes

Indica si se padece diabetes o no

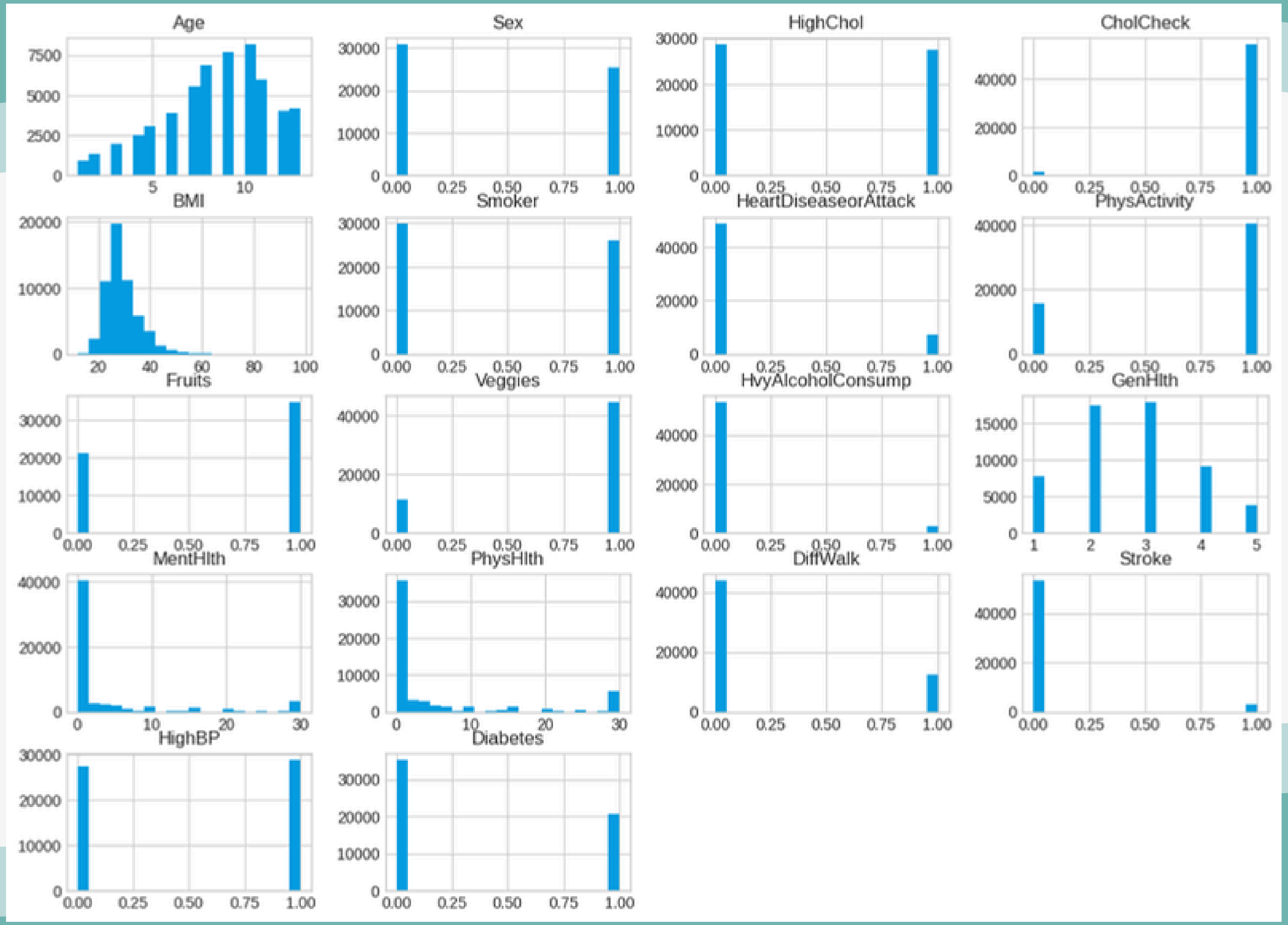
Tamaño del dataset

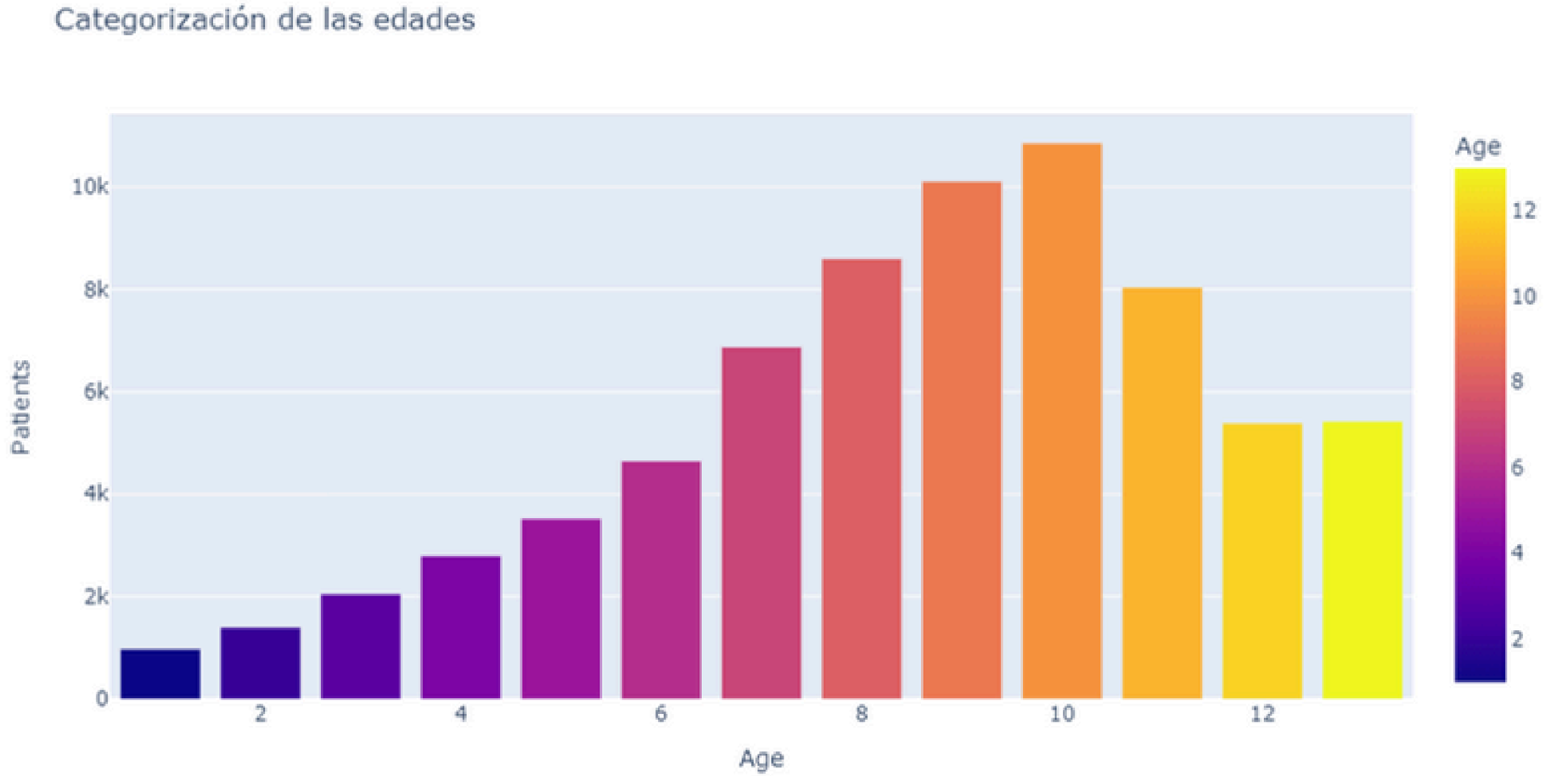
(70692, 18)

Gráficas de los datos

Gáficas generales de los datos

En esta diapositiva se presentan las gráficas de los datos que se tienen el dataset.



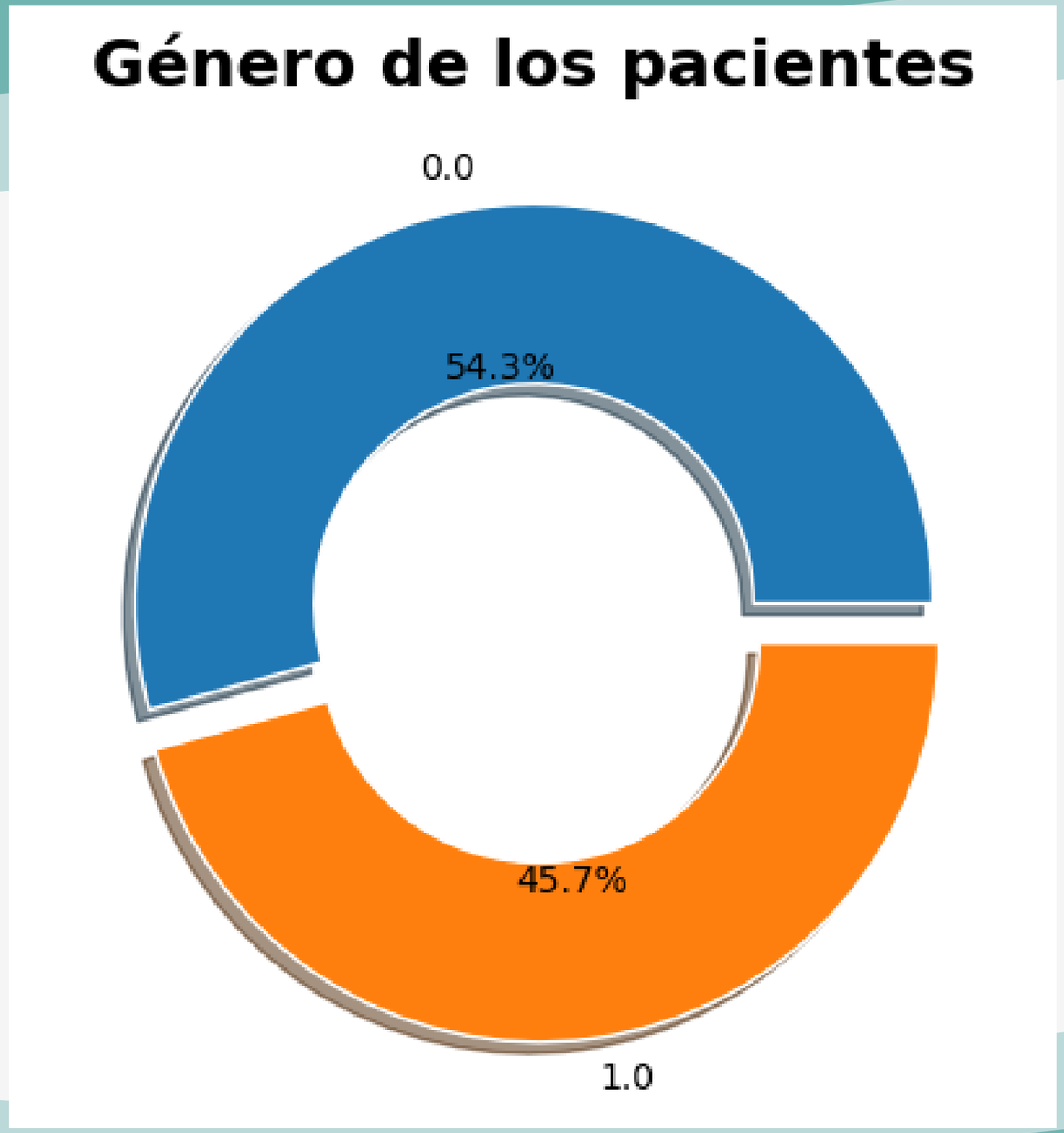


Grafica de edades

En esta diapositiva se presentan la gráfica de las personas por categorización de edad (13 categorias)

Age

Indica la categoría de edad en 13 niveles
1 = 18-24
9 = 60-64
13 = 80 o más



Gráfica de género

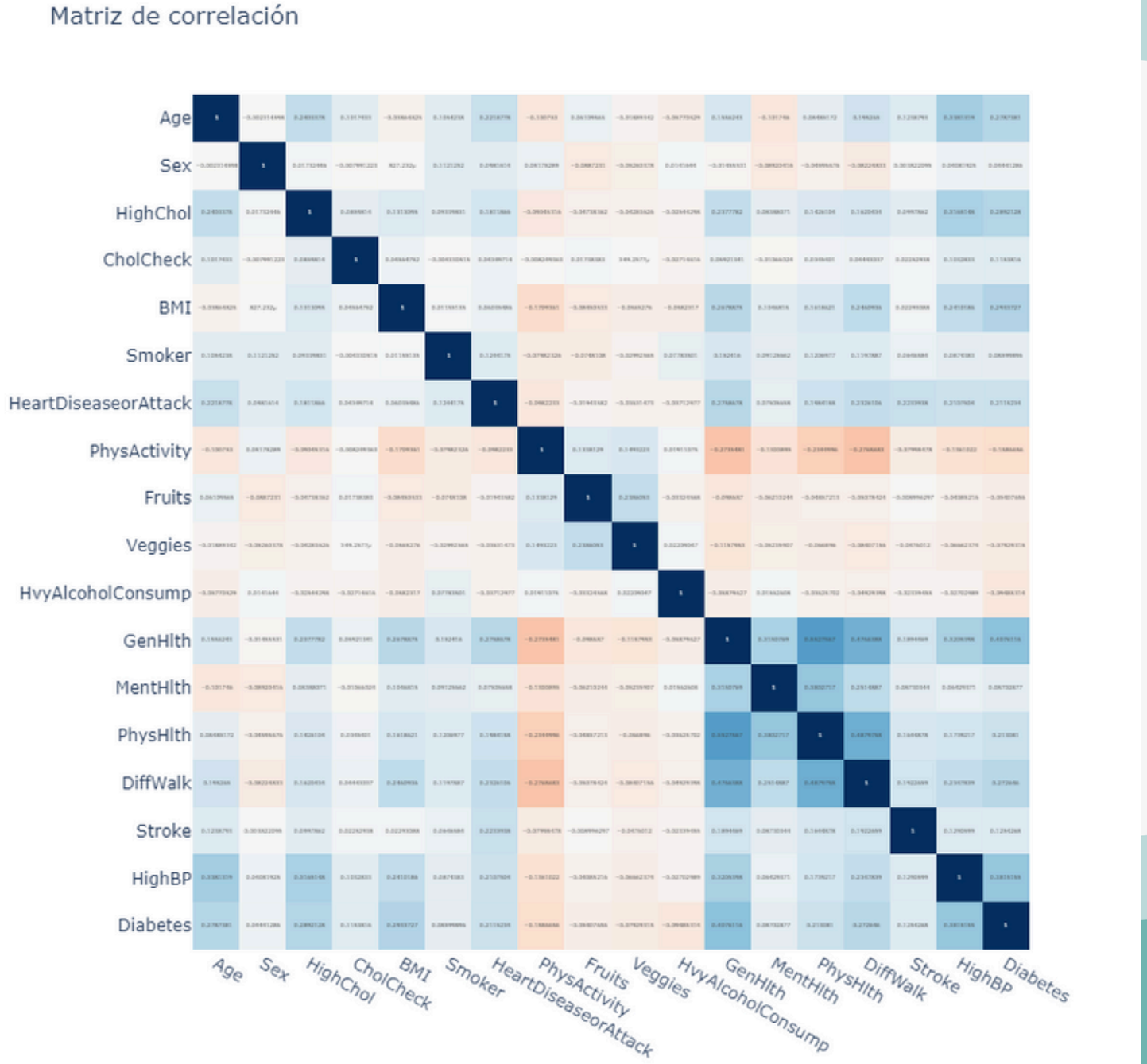
En esta diapositiva se presentan la gráfica de la cantidad de personas por sexo

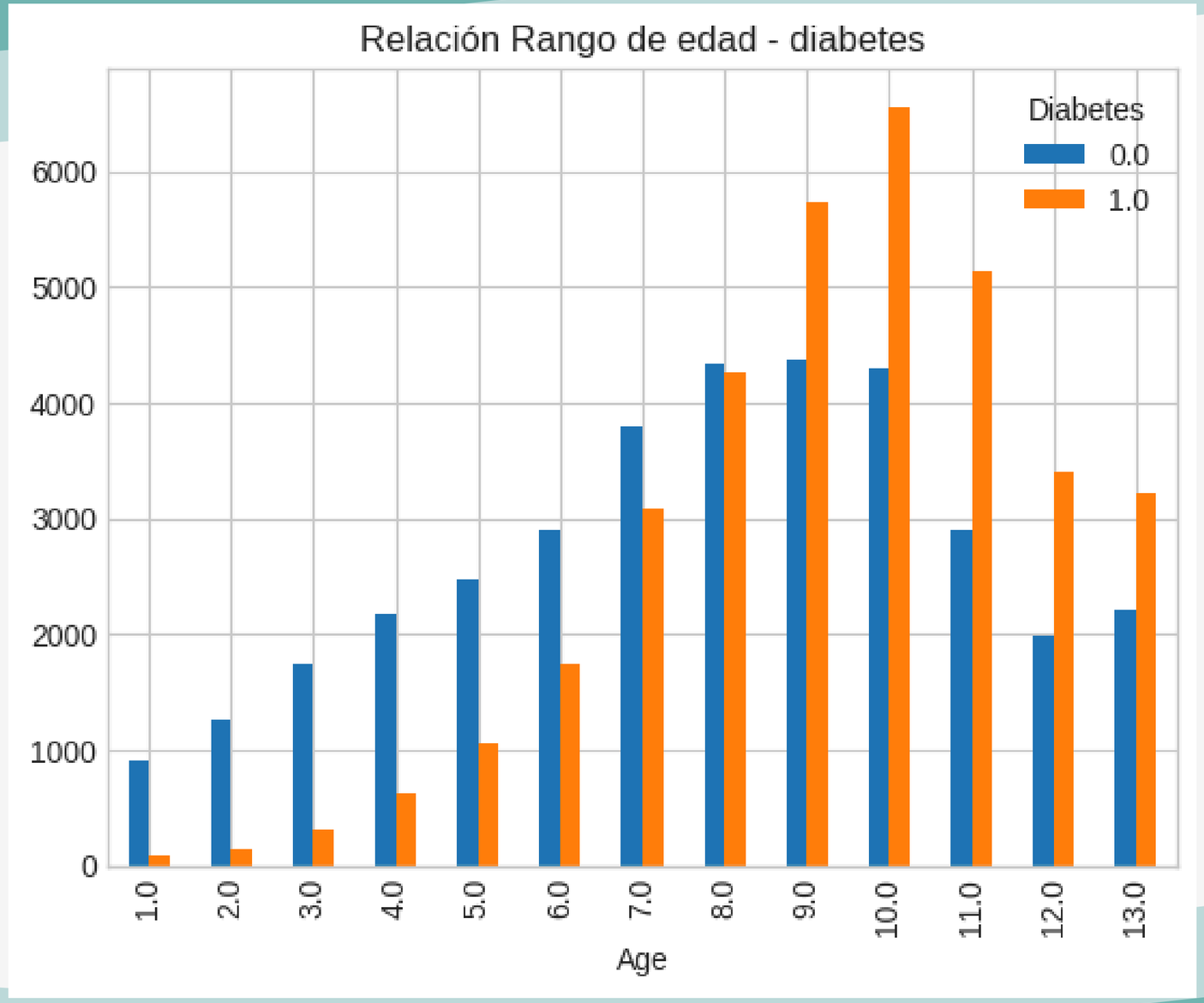
Sex

Indica el género del paciente
0 - Femenino
1 - Masculino

Matriz de correlación

En esta diapositiva se presentan la matriz de correlación que nos permite analizar la relación entre las variables.



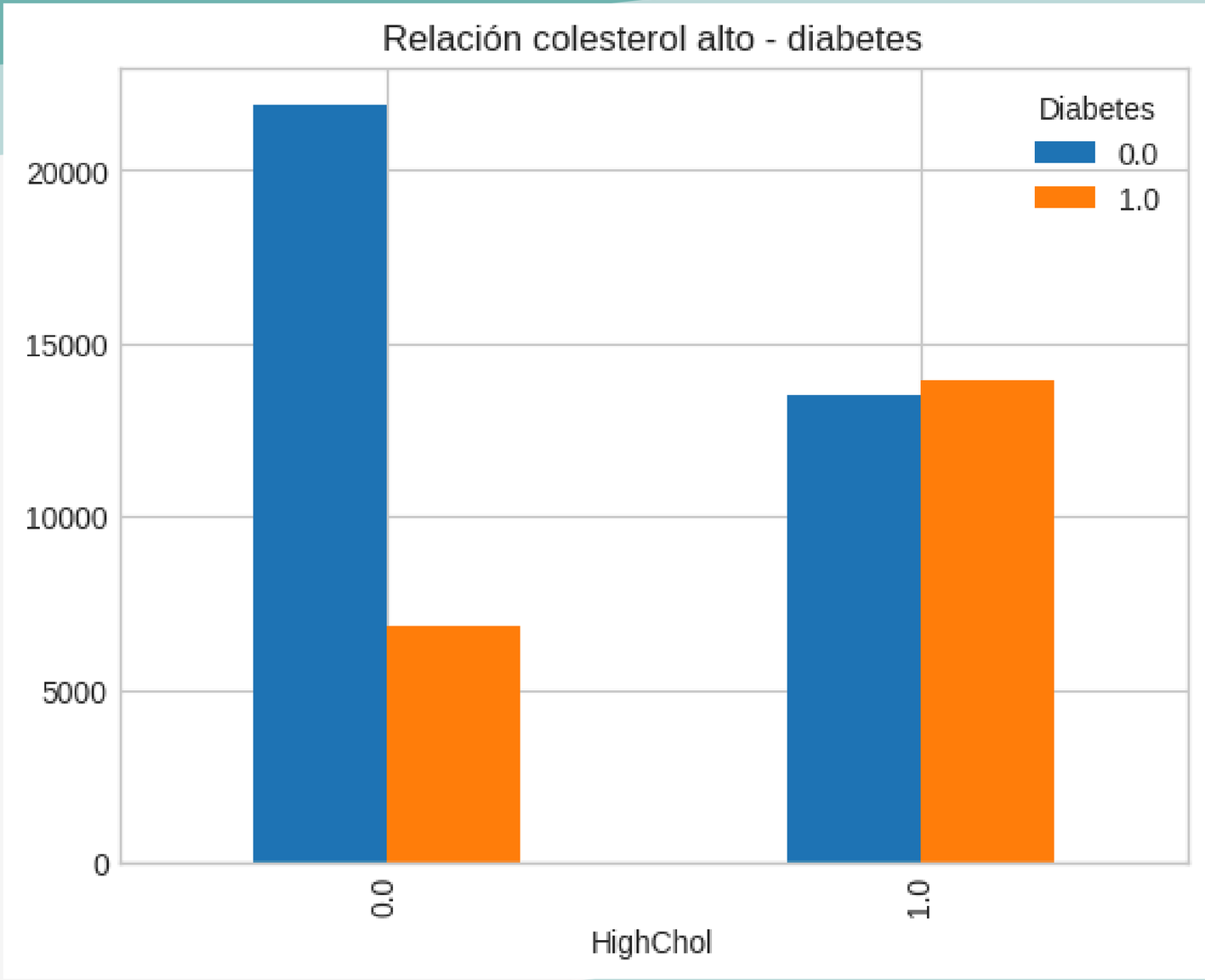


Gráfica Rango edad - Diabetes

En esta diapositiva se presentan la gráfica de la relación entre el rango de edad y si tienen diabetes o no.

Age

Indica la categoría de edad en 13 niveles
1 = 18-24
9 = 60-64
13 = 80 o más

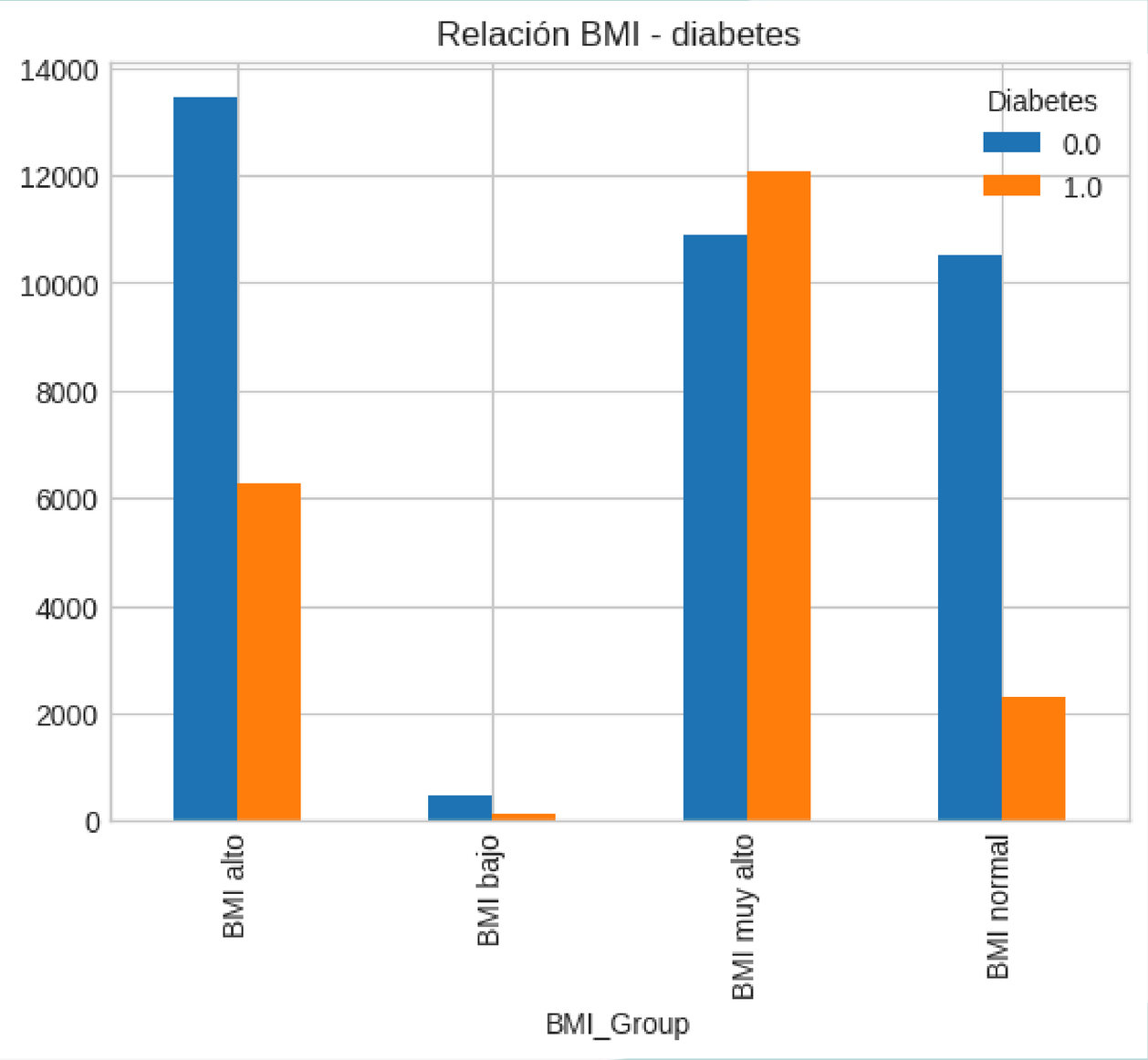


Gráfica Colesterol alto - Diabetes

En esta diapositiva se presentan la gráfica de la relación entre tener colesterol alto y si tienen diabetes o no.

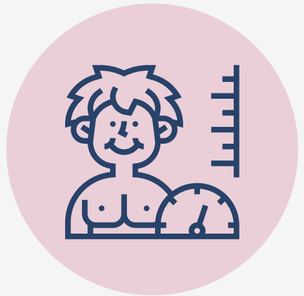
HighChol

Indica si el paciente ha tenido colesterol alto
0 - No
1- Sí



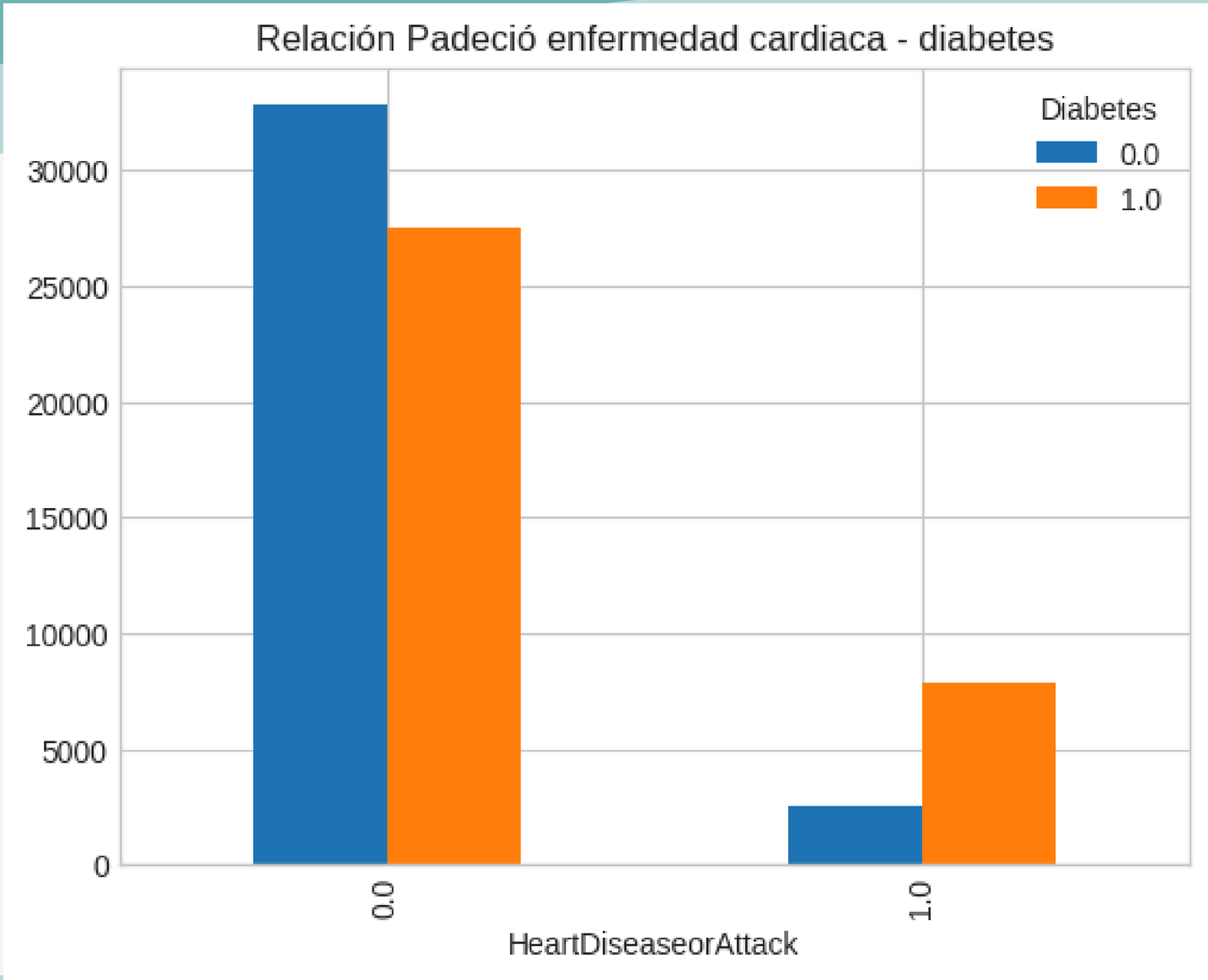
Gráfica BMI - Diabetes

En esta diapositiva se presentan la gráfica de la relación entre el índice de masa corporal y si tienen diabetes o no.



BMI

Índice de masa corporal
BMI < 18.5 : BMI bajo
BMI < 25: BMI normal
BMI < 30: BMI alto
BMI >= 30: BMI muy alto

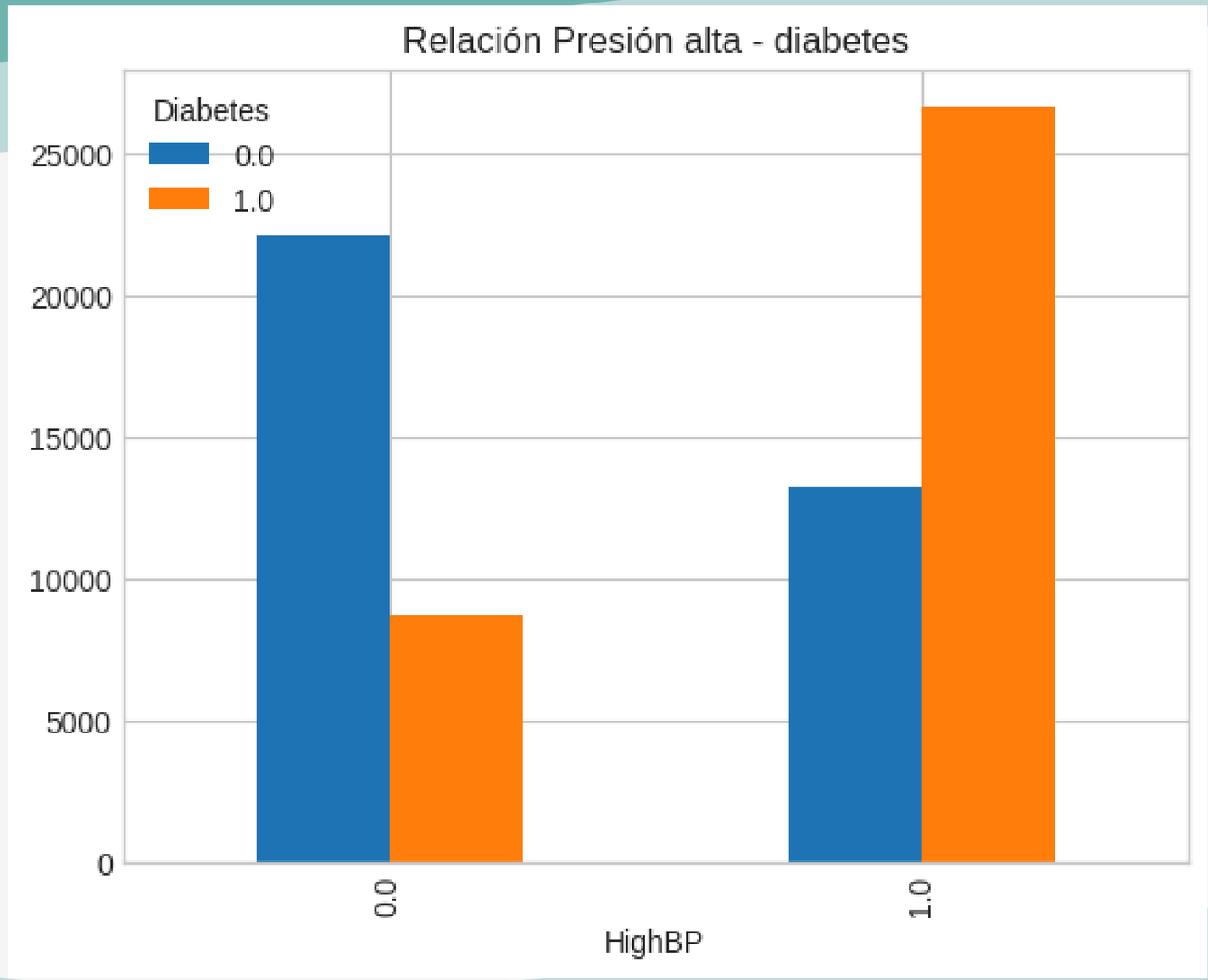


Gráfica Enfermedad cardiaca - Diabetes

En esta diapositiva se presentan la gráfica de la relación entre si se pacedició una enfermedad carfiaca o infarto de miocardio y si tienen diabetes o no.

HeartDiseaseorAttack

Indica si se ha padecido una enfermedad coronaria o infarto de miocardio.
0 - No
1 - Sí



Gráfica Presión alta - Diabetes

En esta diapositiva se presentan la gráfica de la relación entre si se ha tenido presión alta y si tienen diabetes o no.

HighBP

Indica si ha sufrido de presión arterial alta.
0 - No
1 - Sí

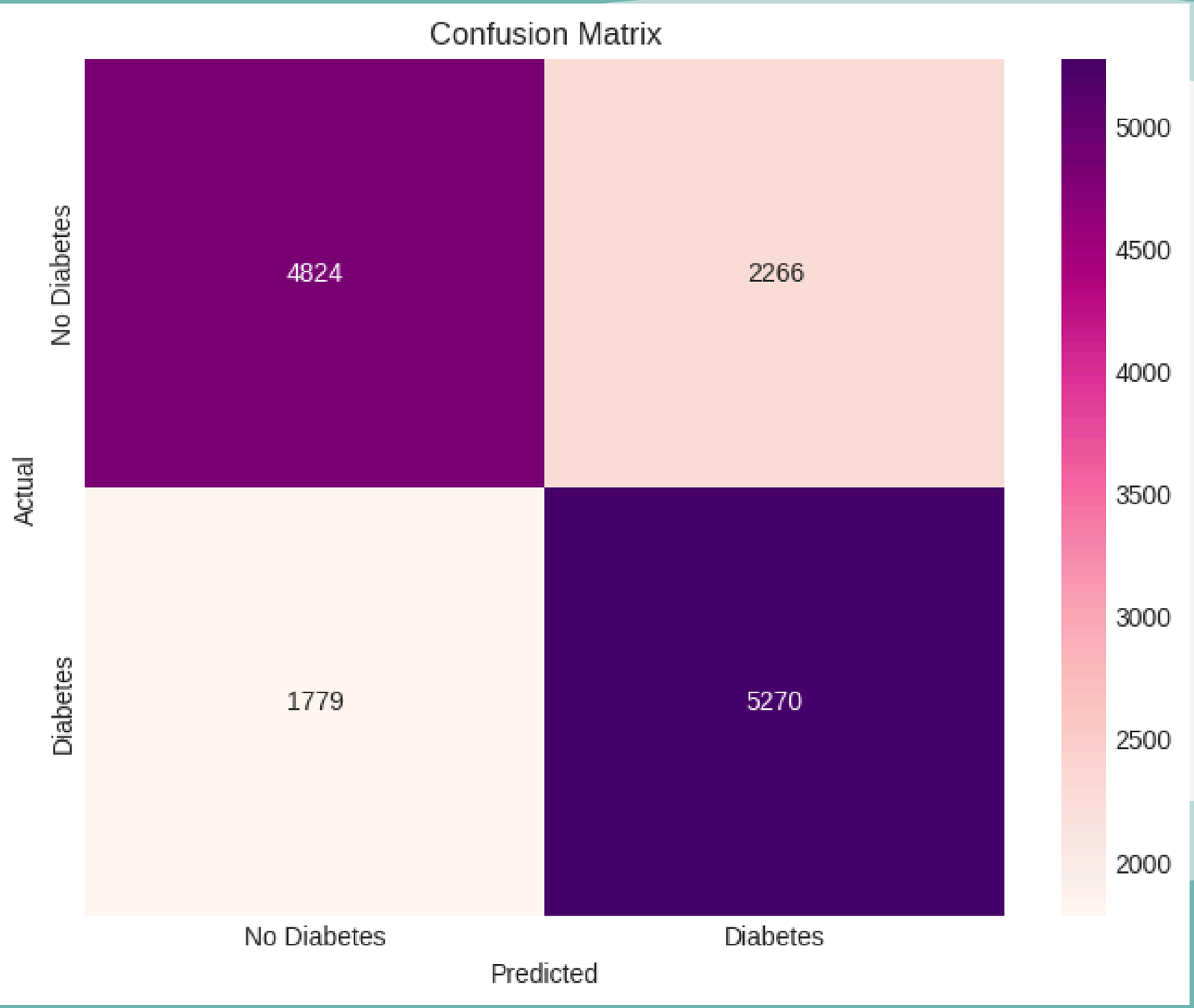
MODELOS

K-Nearest Neighbors (KNN)

	precision	recall	f1-score	support
0.0	0.73	0.68	0.70	7090
1.0	0.70	0.75	0.72	7049
accuracy			0.71	14139
macro avg	0.71	0.71	0.71	14139
weighted avg	0.71	0.71	0.71	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo K-Nearest Neighbors (KNN) y el dato real que se obtuvo.

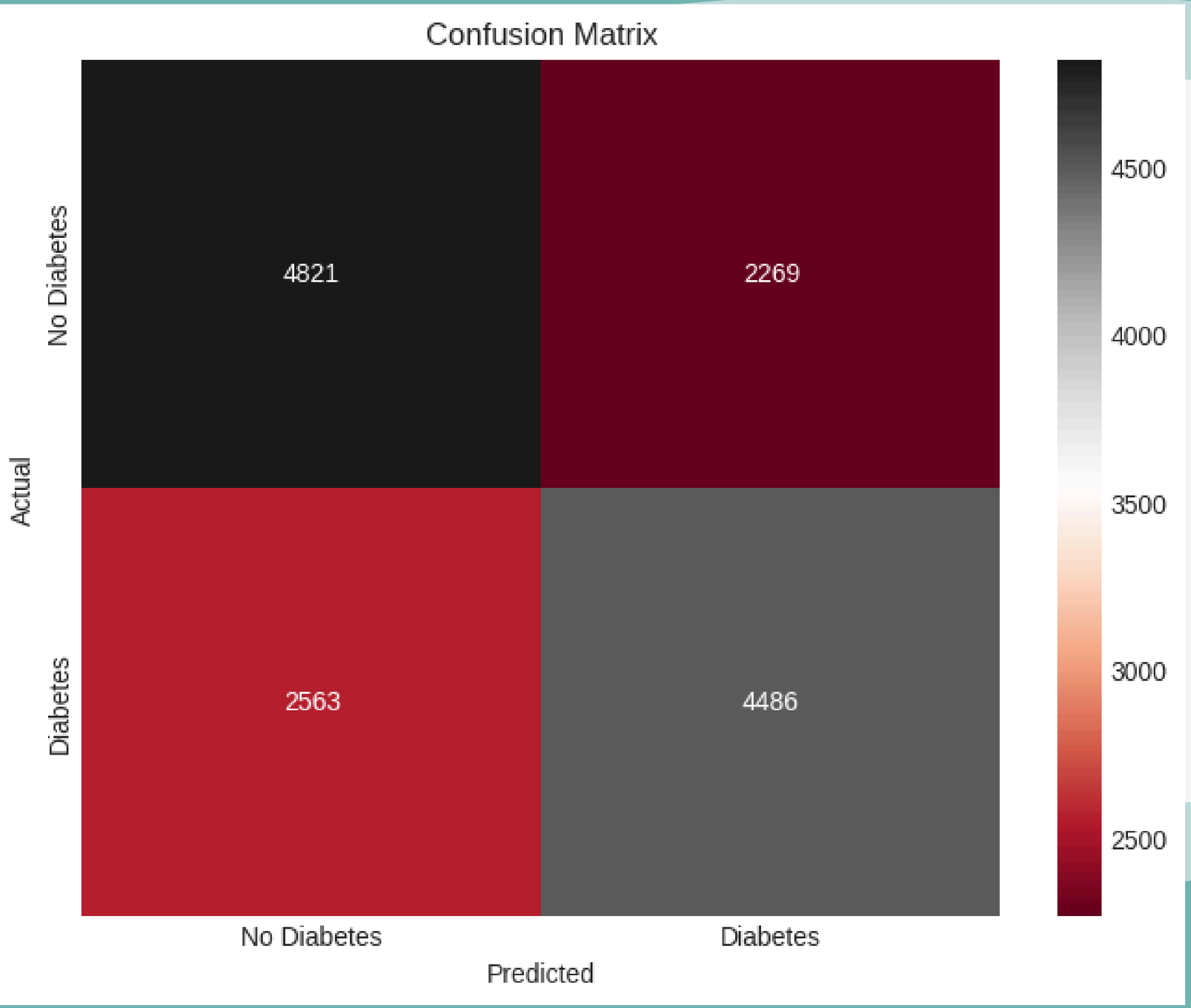


Decision Trees

	precision	recall	f1-score	support
0.0	0.65	0.68	0.67	7090
1.0	0.66	0.64	0.65	7049
accuracy			0.66	14139
macro avg	0.66	0.66	0.66	14139
weighted avg	0.66	0.66	0.66	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo Decision Trees y el dato real que se obtuvo.

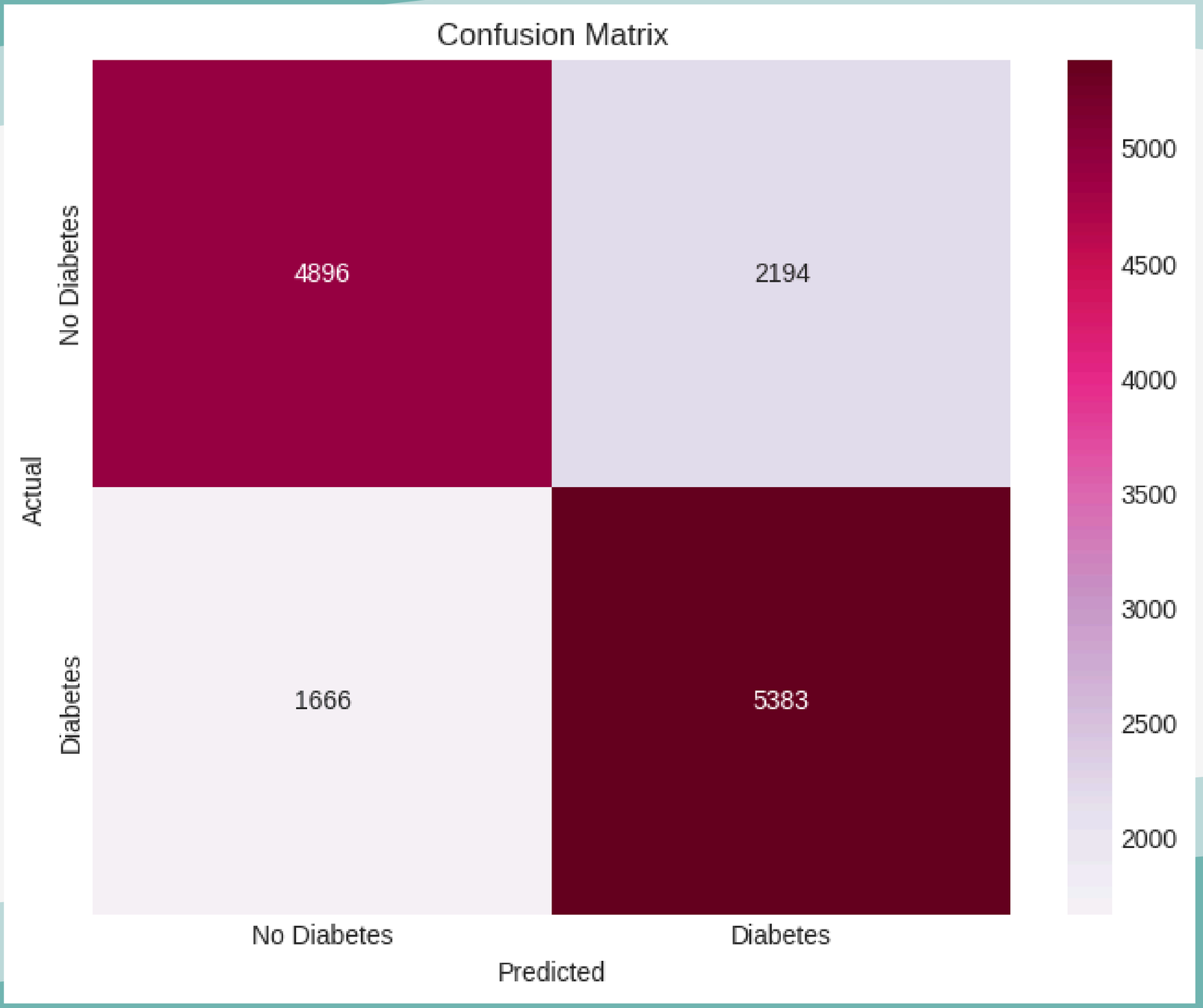


Random Forest

	precision	recall	f1-score	support
0.0	0.75	0.69	0.72	7090
1.0	0.71	0.76	0.74	7049
accuracy			0.73	14139
macro avg	0.73	0.73	0.73	14139
weighted avg	0.73	0.73	0.73	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo Random Forest y el dato real que se obtuvo.



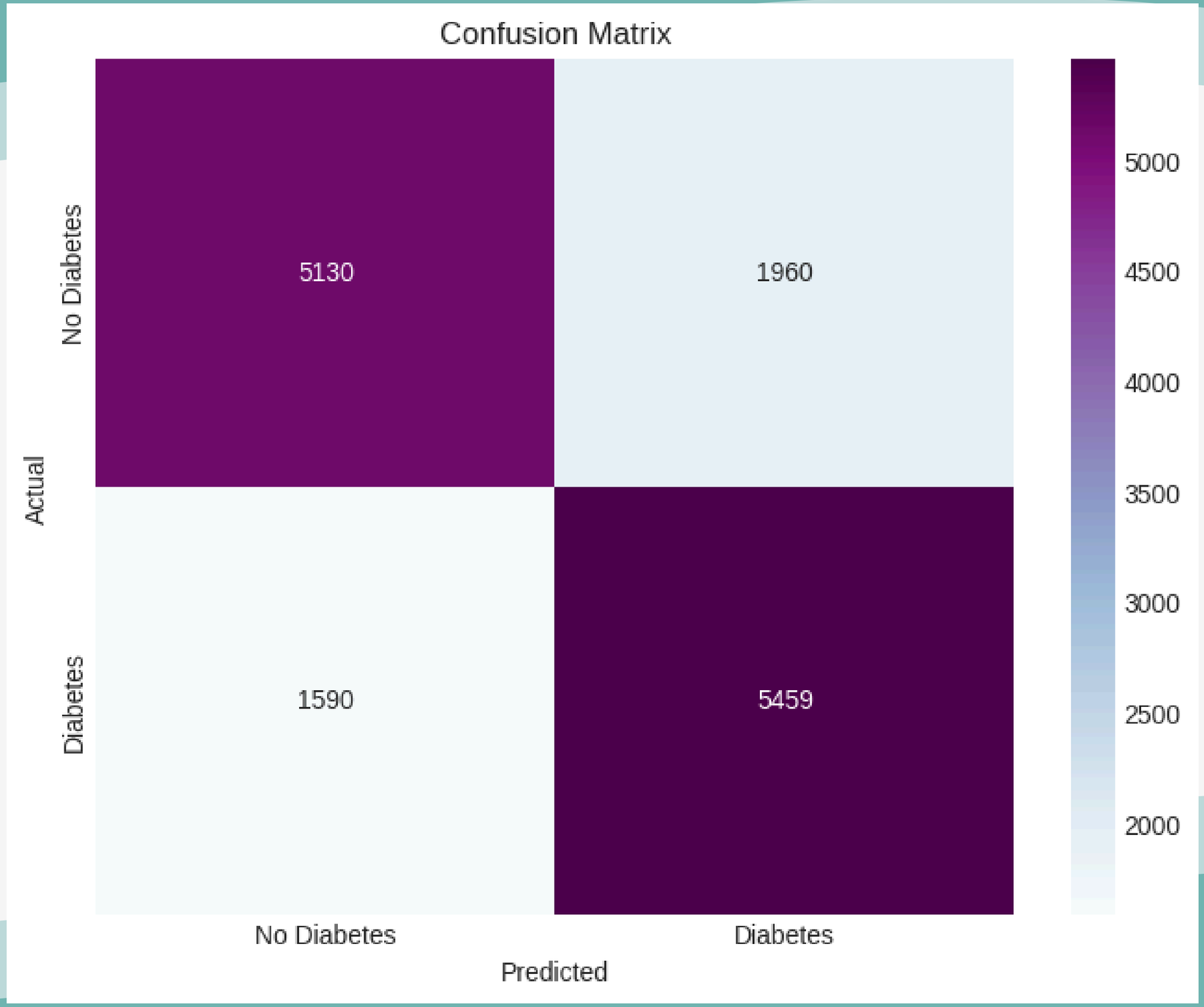
Logistic Regression

Logistic Regression Classification Report:

	precision	recall	f1-score	support
0.0	0.76	0.72	0.74	7090
1.0	0.74	0.77	0.75	7049
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo Logistic Regression y el dato real que se obtuvo.



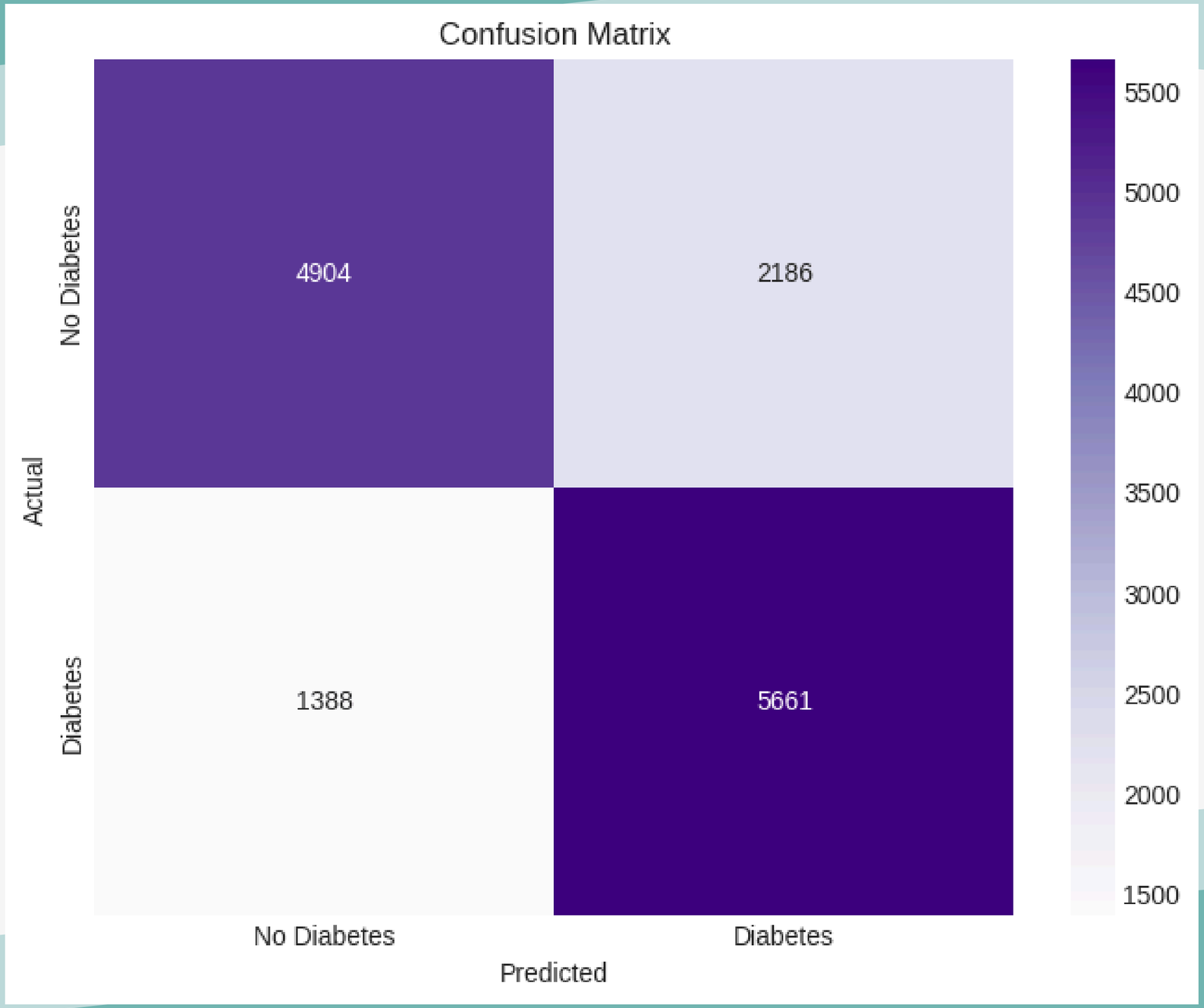
Support Vector Machines (SVM)

SVM Classification Report:

	precision	recall	f1-score	support
0.0	0.78	0.69	0.73	7090
1.0	0.72	0.80	0.76	7049
accuracy			0.75	14139
macro avg	0.75	0.75	0.75	14139
weighted avg	0.75	0.75	0.75	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo Support Vector Machines (SVM) y el dato real que se obtuvo.



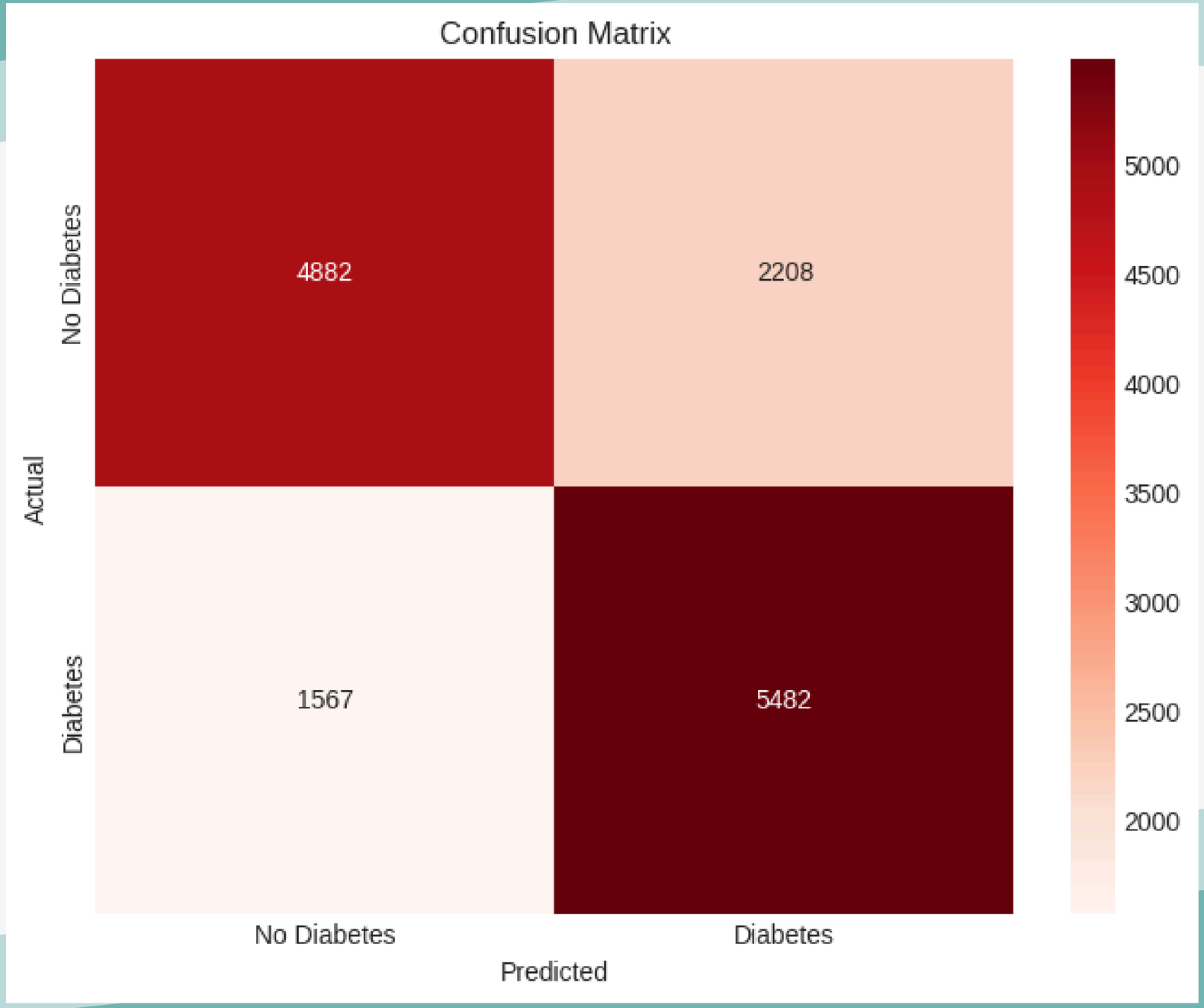
Naive Bayes

Naive Bayes Classification Report:

	precision	recall	f1-score	support
0.0	0.76	0.69	0.72	7090
1.0	0.71	0.78	0.74	7049
accuracy			0.73	14139
macro avg	0.73	0.73	0.73	14139
weighted avg	0.74	0.73	0.73	14139

Matriz de confusión

En esta diapositiva se presenta la matriz de confusión en donde se puede observar la predicción del modelo Naive Bayes y el dato real que se obtuvo.



Comparación de modelos

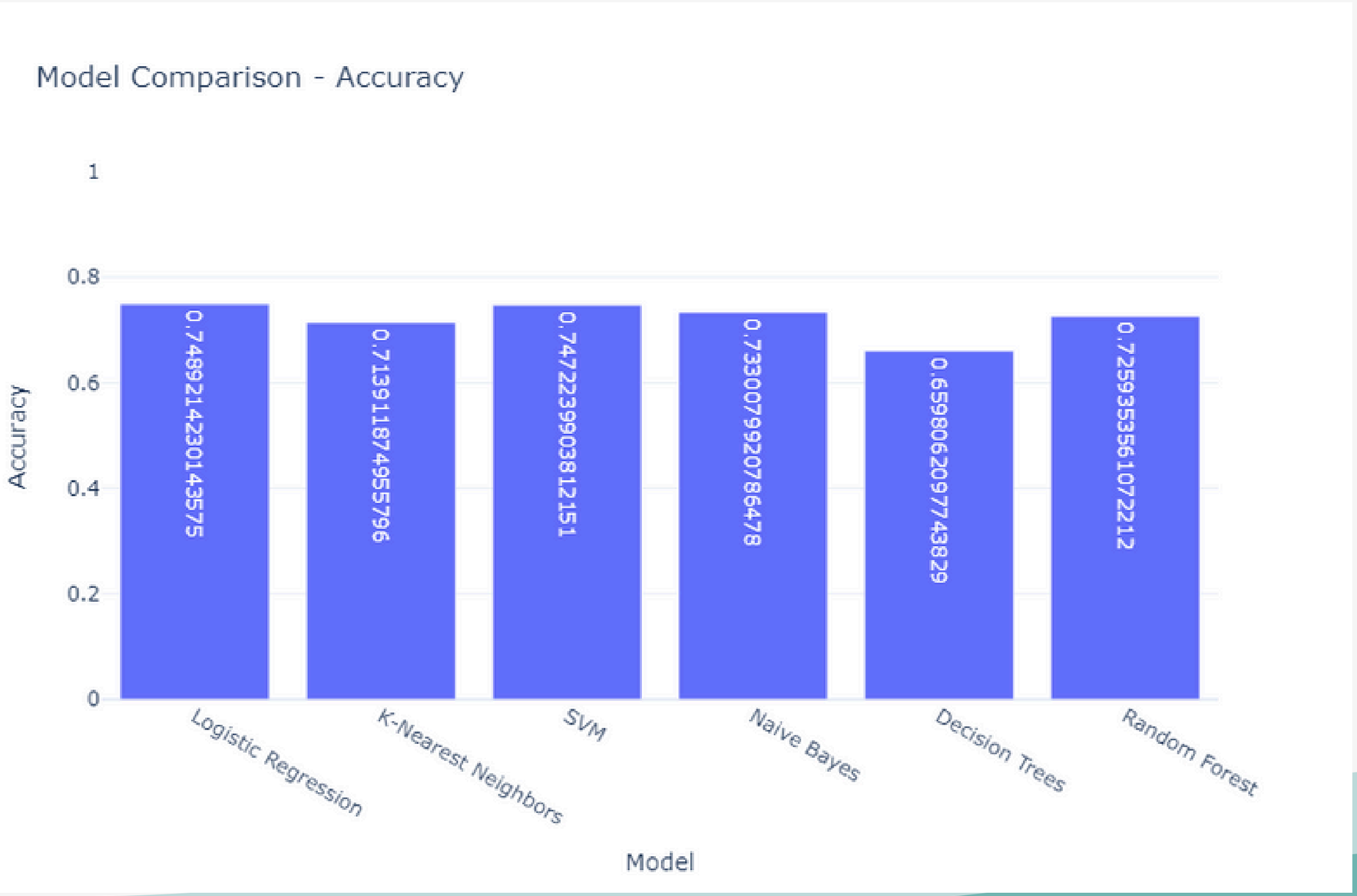
	Model	Accuracy	Precision	Recall	F1-Score
0	Logistic Regression	0.748921	0.735813	0.774436	0.754631
1	K-Nearest Neighbors	0.713912	0.699310	0.747624	0.722660
2	SVM	0.747224	0.721422	0.803093	0.760070
3	Naive Bayes	0.733008	0.712874	0.777699	0.743877
4	Decision Trees	0.658250	0.664101	0.636402	0.649957
5	Random Forest	0.726996	0.710439	0.763654	0.736086

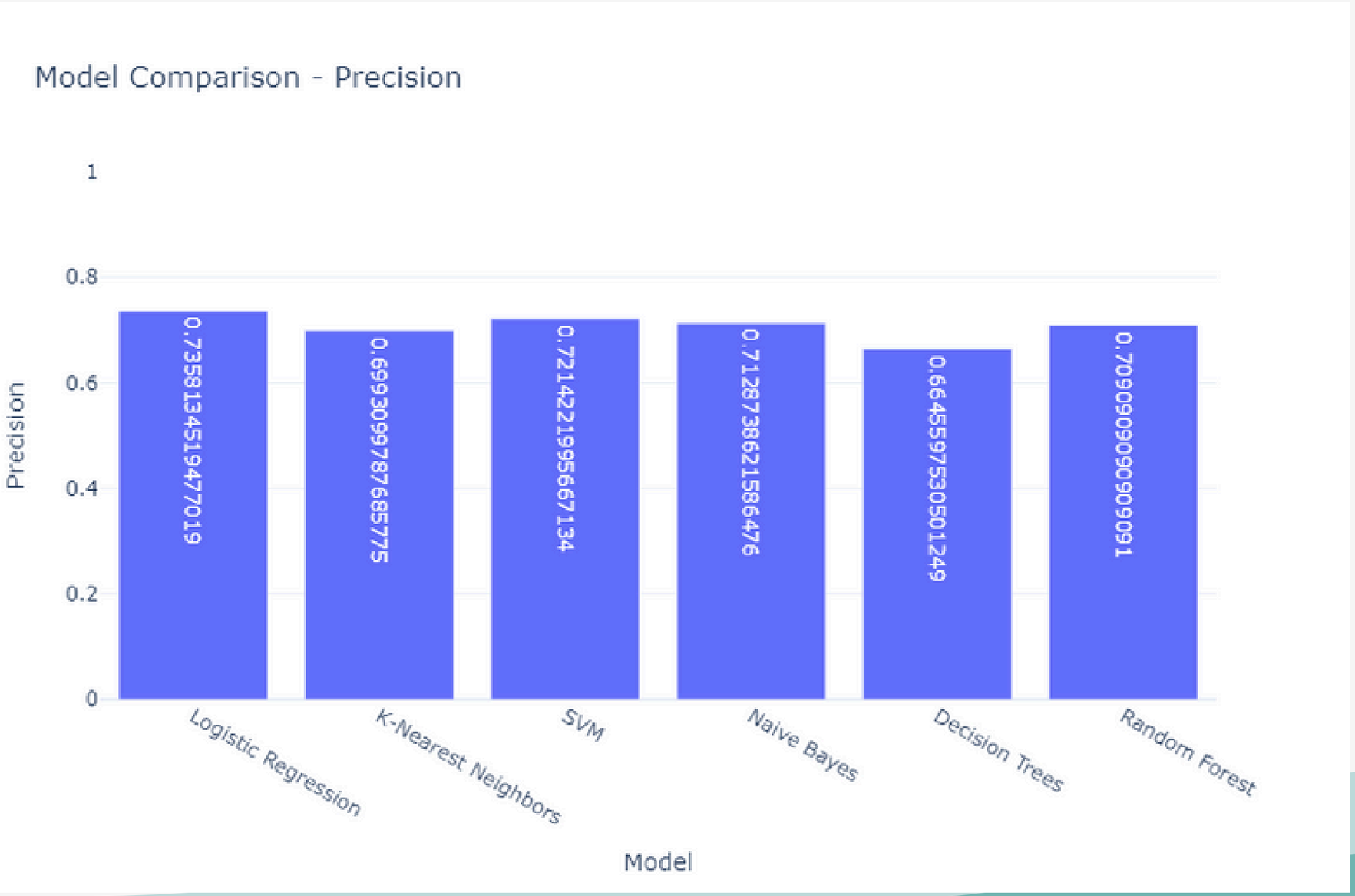
Modelos que más se ajustan

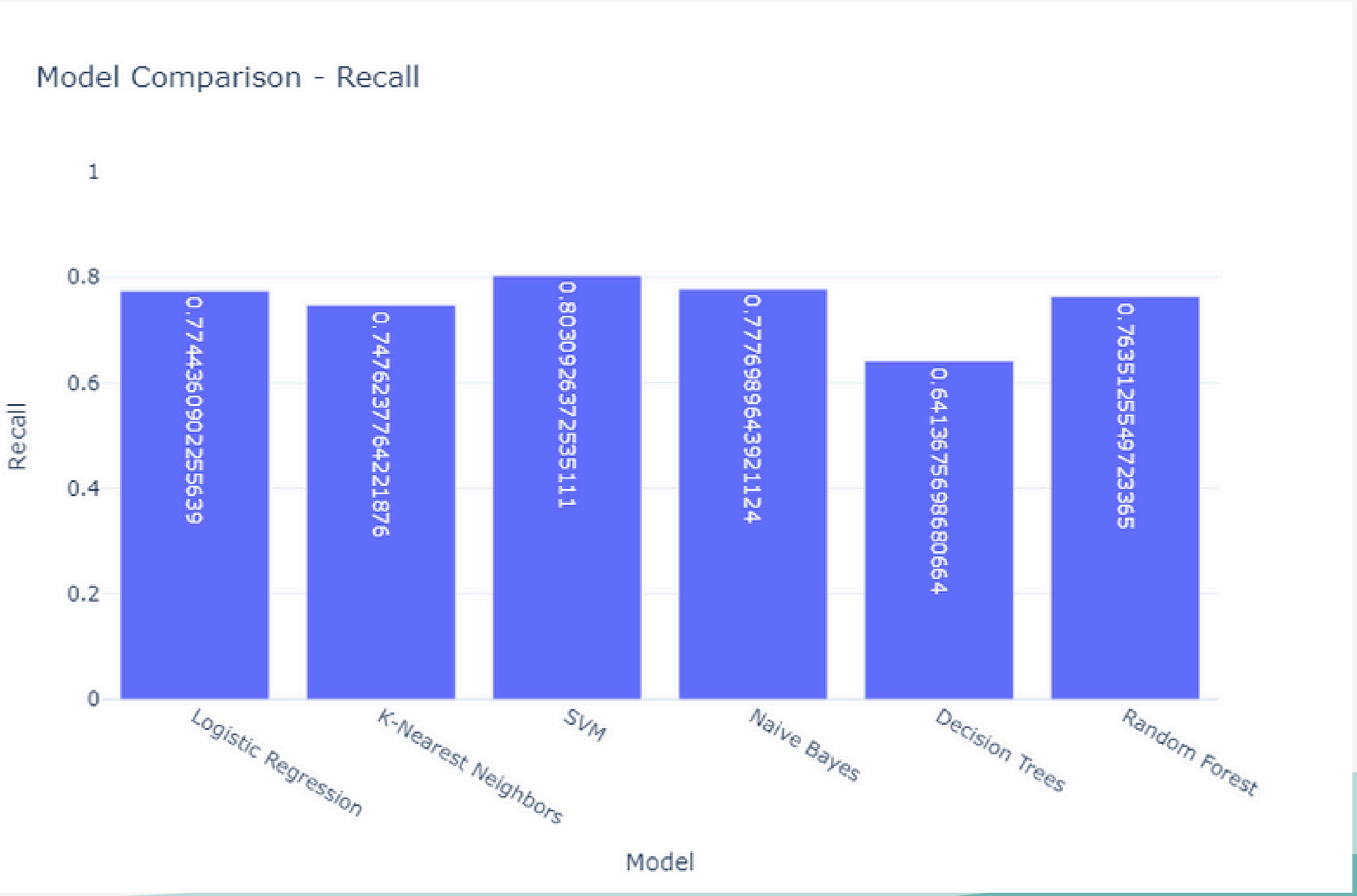
Observamos que los son los que obtuvieron los mejores valores.

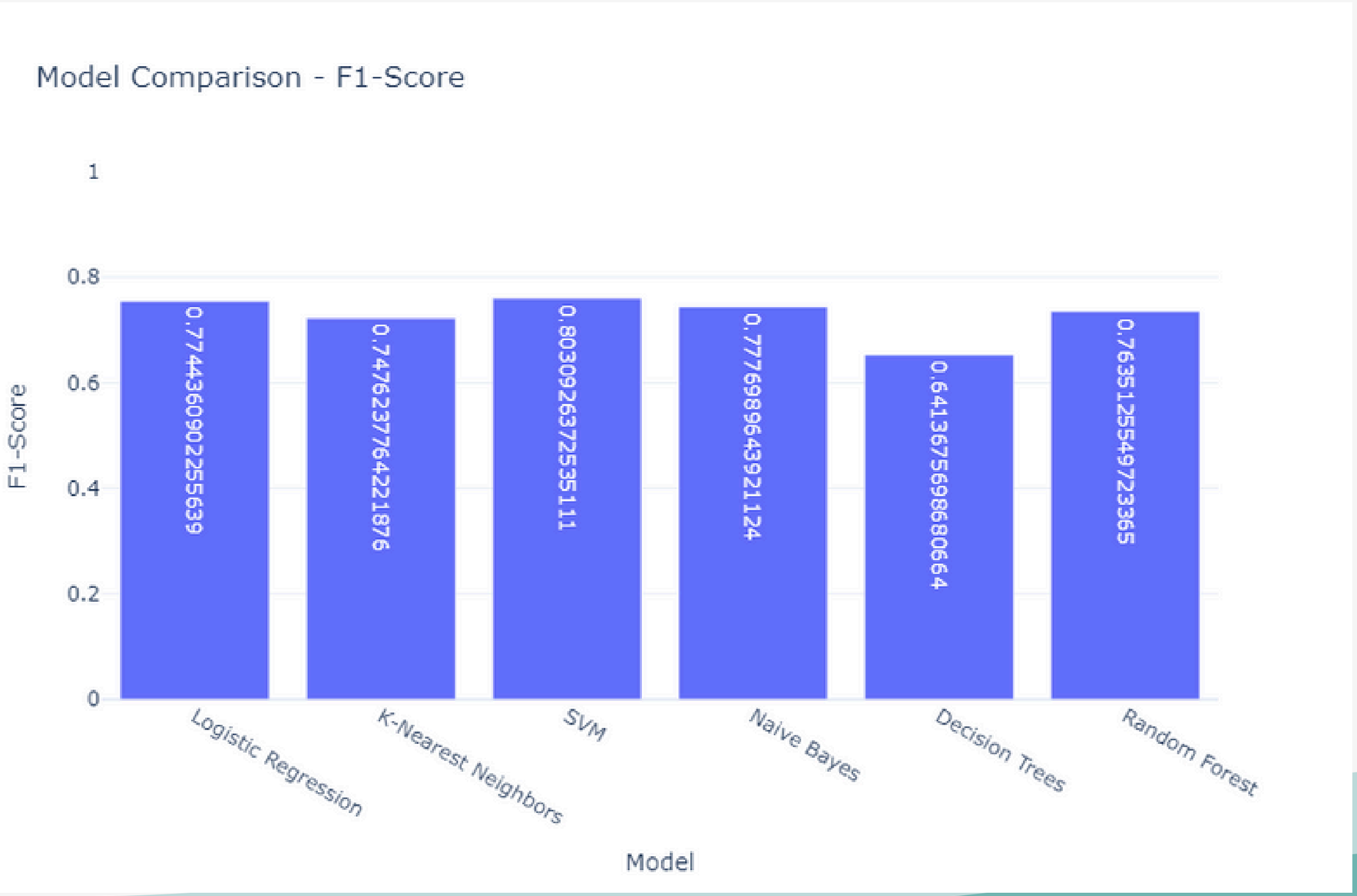
Modelo que menos se ajusta

Observamos que los valores no son muy favorabes


Model Comparison - Accuracy
Accuracy
1
0.8
0.6
0.4
0.2
0
0.7489214230143575
0.7139118749557796
0.7472239903812151
0.7330079920786478
0.6598062097743829
0.7259353561072212
Logistic Regression
K-Nearest Neighbors
SVM
Naive Bayes
Decision Trees
Random Forest
Model


Model Comparison - Precision
Precision
1
0.8
0.6
0.4
0.2
0
0.7358134519477019
0.6993099787685775
0.7214221955667134
0.7128738621586476
0.6645597530501249
0.7090909090909091
Logistic Regression
K-Nearest Neighbors
SVM
Naive Bayes
Decision Trees
Random Forest
Model


Model Comparison - Recall
Recall
1
0.8
0.6
0.4
0.2
0
0.7744360902255639
0.7476237764221876
0.8030926372535111
0.7769896439211124
0.6413675698680664
0.7635125549723365
Logistic Regression
K-Nearest Neighbors
SVM
Naive Bayes
Decision Trees
Random Forest
Model


Model Comparison - F1-Score
F1-Score
1
0.8
0.6
0.4
0.2
0
0.7744360902255639
0.7476237764221876
0.8030926372535111
0.7769896439211 24
0.6413675698680664
0.7635125549723365
Logistic Regression
K-Nearest Neighbors
SVM
Naive Bayes
Decision Trees
Random Forest
Model