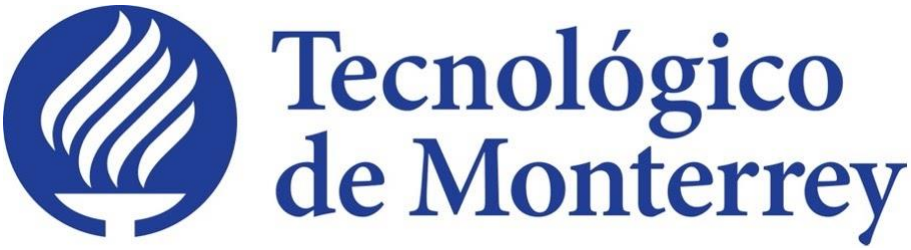


Instituto Tecnológico y de Estudios superiores de Monterrey

Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo. (Portafolio Análisis)

TC3006C.101: Inteligencia artificial avanzada para la ciencia de datos I

Alumno:

Antonio Oviedo Paredes A01752114

Docente:

Jorge Adolfo Ramírez Uresti

12 de septiembre de 2023

1. Set de datos

Características por las que se eligió el set de datos:

- Variedad de características: El set de datos cuenta con 9 columnas que representan una característica diferente, lo que le permite al árbol de decisión poder elegir entre más opciones para separar los datos, brindando un mejor ajuste y desempeño general.

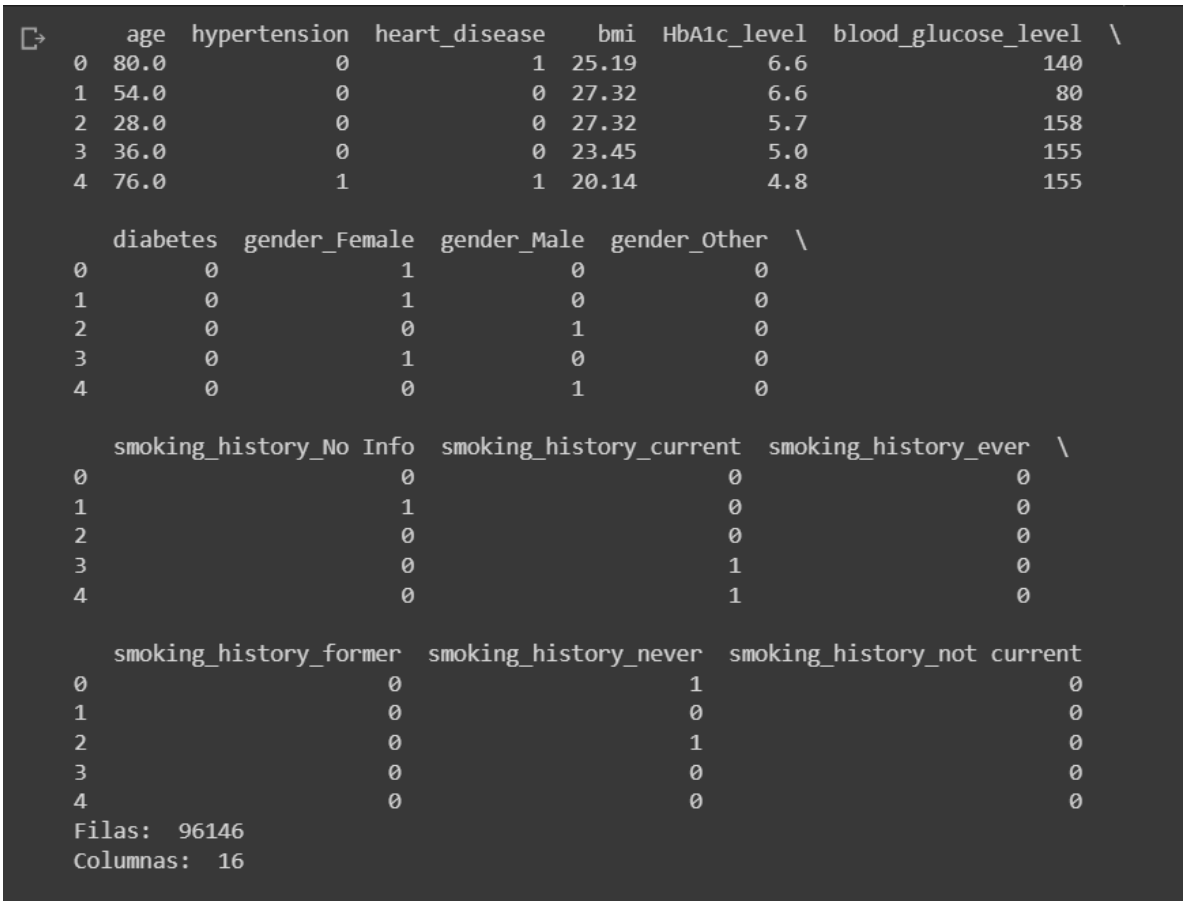
- Tamaño: Eliminando los valores faltantes, el set de datos cuenta con una longitud de 2011 registros, lo cual es suficiente para que el árbol pueda generalizar de forma correcta y obtener resultados decentes.

- Predicción de clases: Los árboles de decisión son muy buenos para clasificar y la columna de salida del set de datos es una clase binaria (potable, no potable).

- Fuente confiable: El set de datos fue obtenido de Kaggle, la cual es una plataforma conocida que cuenta con una gran variedad de data sets de calidad para diferentes aplicaciones.

```
[64] from typing_extensions import dataclass_transform
     import pandas as pd

     # Cargar datos
     data = pd.read_csv("diabetes.csv")
     # Eliminar duplicados y valores faltantes
     data = data.drop_duplicates()
     data = data.dropna()

     #One-hot encode the "gender" and "smoking_history" columns, creating dummy variables
     dummies = pd.get_dummies(data[["gender", "smoking_history"]])
     data = data.drop(["gender", "smoking_history"], axis=1)

     # Observations and target data
     data = pd.concat([data, dummies], axis=1)

     print(data.head())
     print("Filas: ", data.shape[0])
     print("Columnas: ", data.shape[1])
```

Figura 1.1 Código para importar set de datos



	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	\
0	80.0	0	1	25.19	6.6	140	
1	54.0	0	0	27.32	6.6	80	
2	28.0	0	0	27.32	5.7	158	
3	36.0	0	0	23.45	5.0	155	
4	76.0	1	1	20.14	4.8	155	

	diabetes	gender_Female	gender_Male	gender_Other	\
0	0	1	0	0	
1	0	1	0	0	
2	0	0	1	0	
3	0	1	0	0	
4	0	0	1	0	

	smoking_history_No Info	smoking_history_current	smoking_history_ever	\
0	0	0	0	
1	1	0	0	
2	0	0	0	
3	0	1	0	
4	0	1	0	

	smoking_history_former	smoking_history_never	smoking_history_not current
0	0	1	0
1	0	0	0
2	0	1	0
3	0	0	0
4	0	0	0

Filas: 96146
Columnas: 16

Figura 1.2 Visualización del set de datos

2. Separación y evaluación del modelo.

Para separar los datos en entrenamiento y validación se utilizó la proporción de 80% entrenamiento y 20% pruebas. Esta proporción es una de las más utilizadas, ya que está comprobado que, para sets de datos medianos, el modelo generaliza de mejor manera y se garantiza un equilibrio adecuado entre la evaluación y el entrenamiento del modelo.

Esta selección de datos se realiza de forma aleatoria para evitar sesgos al momento de elegir los datos, por lo que los índices que se muestran en las cabeceras de los datos no están en algún orden particular.

Al elegir una proporción de 80 - 20 y tener un set de datos de 2011 filas, se espera que la longitud 1608 renglones de entrenamiento y 403 renglones de testeo.

En la gráfica (figura 2.1) se puede observar que se respeta la proporción de 80 - 20, mientras que se muestra el número de elementos que pertenecen a cada clase. Mientras cada clase tenga una cantidad similar de datos, el modelo está mejor balanceado. Balancear el número de clases puede ser una herramienta para mejorar el desempeño del modelo.

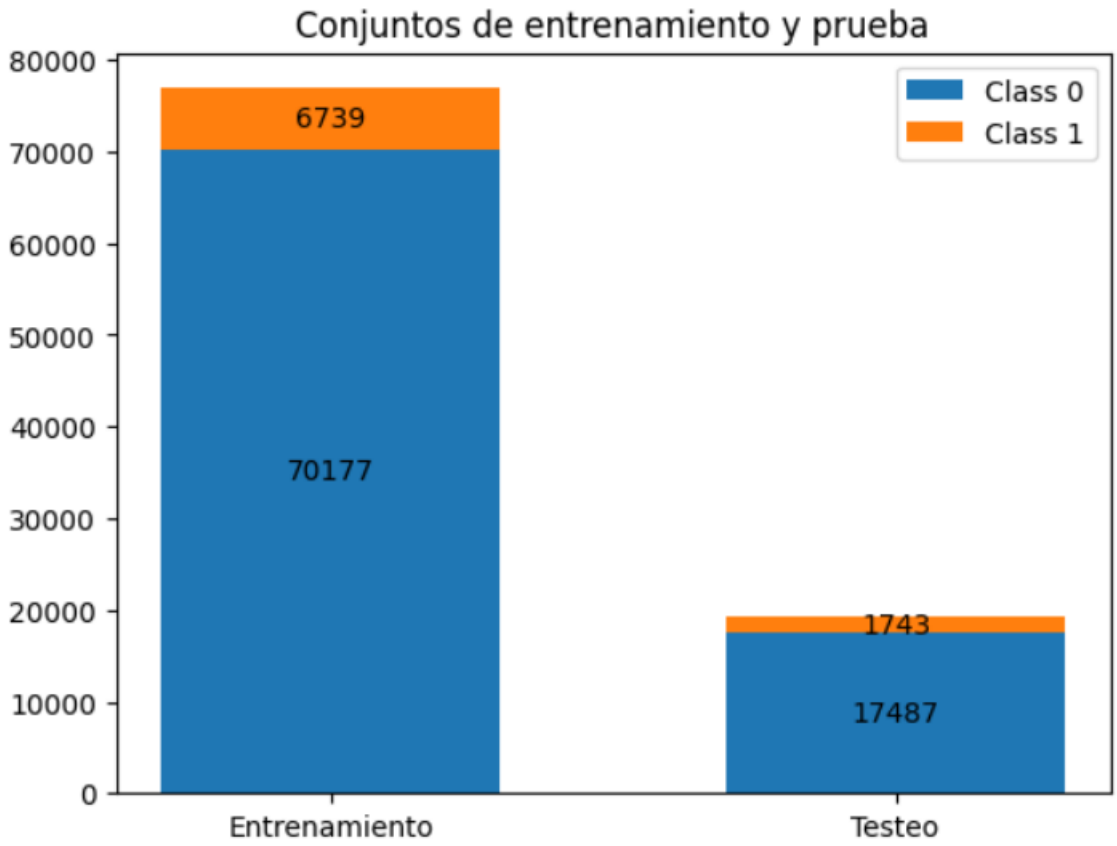
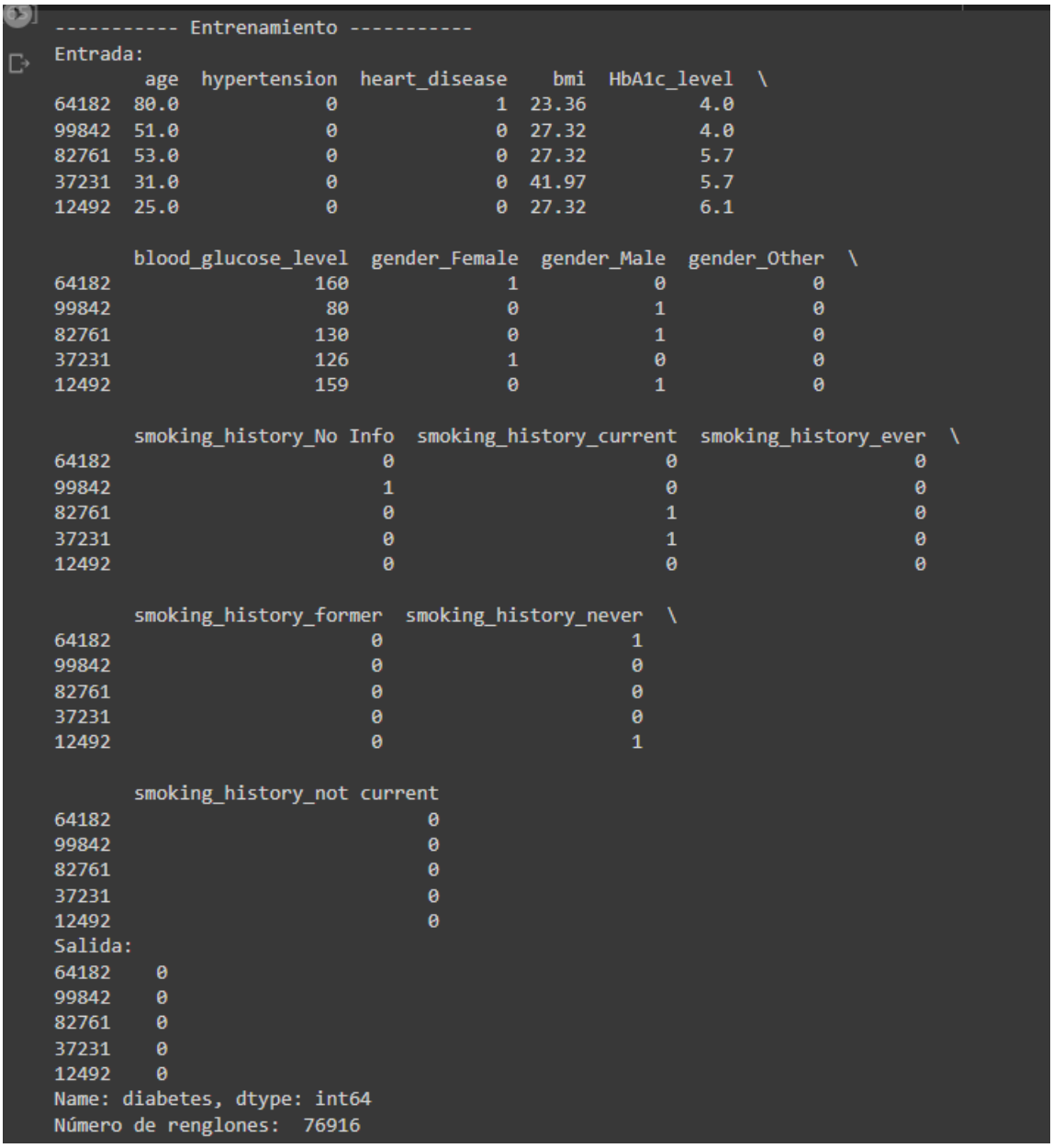
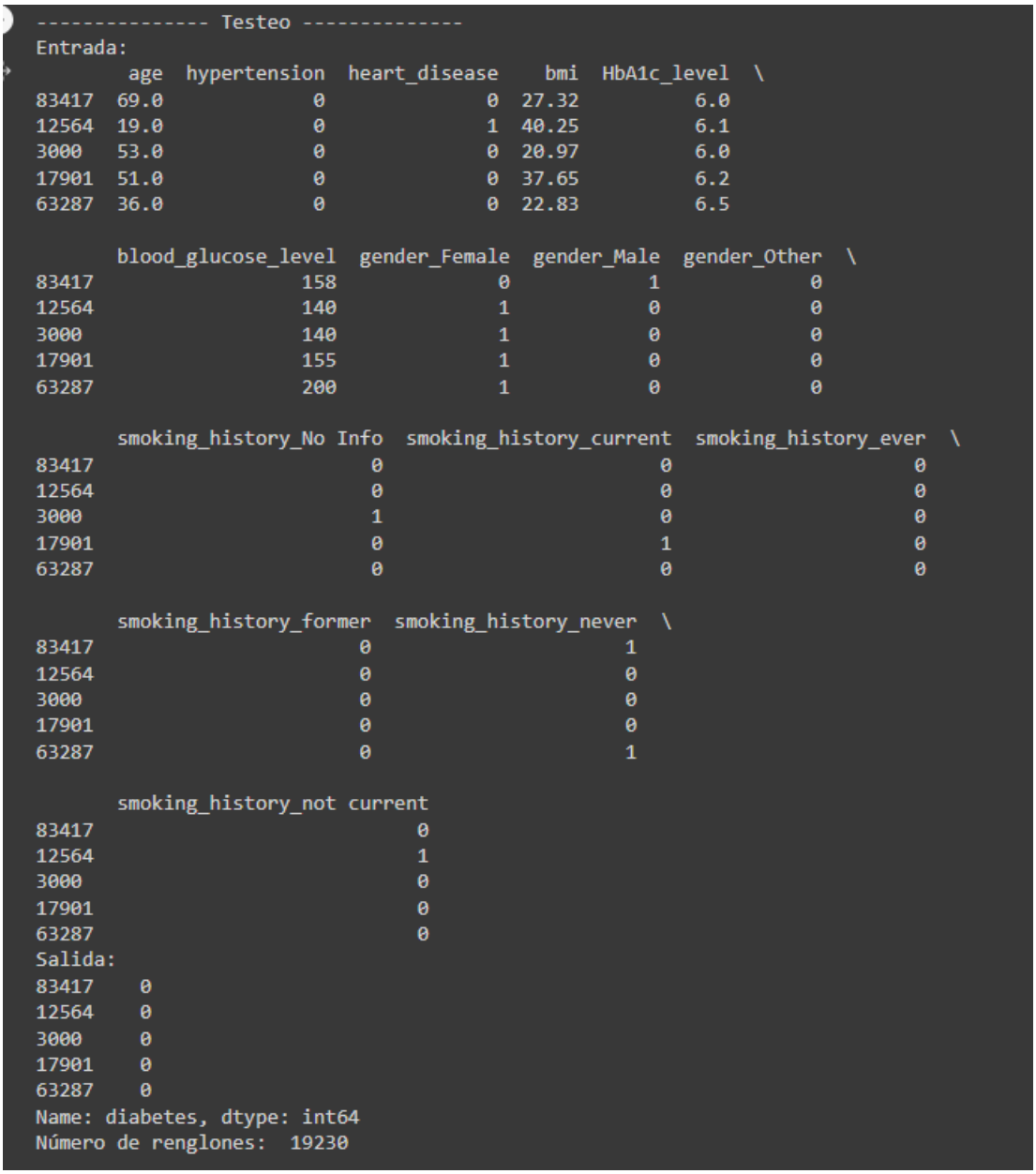


Figura 2.1 Separación de datos en conjuntos de entrenamiento y prueba



```
---------- Entrenamiento ----------
Entrada:
        age  hypertension  heart_disease    bmi  HbA1c_level  \
64182  80.0             0              1  23.36          4.0
99842  51.0             0              0  27.32          4.0
82761  53.0             0              0  27.32          5.7
37231  31.0             0              0  41.97          5.7
12492  25.0             0              0  27.32          6.1

       blood_glucose_level  gender_Female  gender_Male  gender_Other  \
64182                  160              1            0             0
99842                   80              0            1             0
82761                  130              0            1             0
37231                  126              1            0             0
12492                  159              0            1             0

       smoking_history_No Info  smoking_history_current  smoking_history_ever  \
64182                        0                        0                     0
99842                        1                        0                     0
82761                        0                        1                     0
37231                        0                        1                     0
12492                        0                        0                     0

       smoking_history_former  smoking_history_never  \
64182                       0                      1
99842                       0                      0
82761                       0                      0
37231                       0                      0
12492                       0                      1

       smoking_history_not current
64182                            0
99842                            0
82761                            0
37231                            0
12492                            0
Salida:
64182    0
99842    0
82761    0
37231    0
12492    0
Name: diabetes, dtype: int64
Número de renglones:  76916
```

Figura 2.2 Visualización de datos de entrenamiento



```
--------------- Testeo --------------
Entrada:
        age  hypertension  heart_disease    bmi  HbA1c_level  \
83417  69.0             0              0  27.32          6.0
12564  19.0             0              1  40.25          6.1
3000   53.0             0              0  20.97          6.0
17901  51.0             0              0  37.65          6.2
63287  36.0             0              0  22.83          6.5

       blood_glucose_level  gender_Female  gender_Male  gender_Other  \
83417                  158              0            1             0
12564                  140              1            0             0
3000                   140              1            0             0
17901                  155              1            0             0
63287                  200              1            0             0

       smoking_history_No Info  smoking_history_current  smoking_history_ever  \
83417                        0                        0                     0
12564                        0                        0                     0
3000                         1                        0                     0
17901                        0                        1                     0
63287                        0                        0                     0

       smoking_history_former  smoking_history_never  \
83417                       0                      1
12564                       0                      0
3000                        0                      0
17901                       0                      0
63287                       0                      1

       smoking_history_not current
83417                            0
12564                            1
3000                             0
17901                            0
63287                            0
Salida:
83417    0
12564    0
3000     0
17901    0
63287    0
Name: diabetes, dtype: int64
Número de renglones:  19230
```

Figura 2.3 Visualización de datos de testeo

3. Sesgo y varianza

Para diagnosticar el grado de sesgo y varianza se utilizaron 3 acercamientos:

- Gráfica de la curva de aprendizaje: La gráfica (figura 3.1) muestra la relación entre el número de muestras en el entrenamiento y la exactitud en los datos de entrenamiento y testeo. La forma de la gráfica nos muestra que a medida que el número de datos de entrenamiento aumenta, hay un mejor balance en el aprendizaje del modelo ya sea para los datos de entrenamiento o prueba.
  En este caso, el espacio entre las líneas indica que hay un poco de varianza, esto quiere decir que no hay buena generalización.

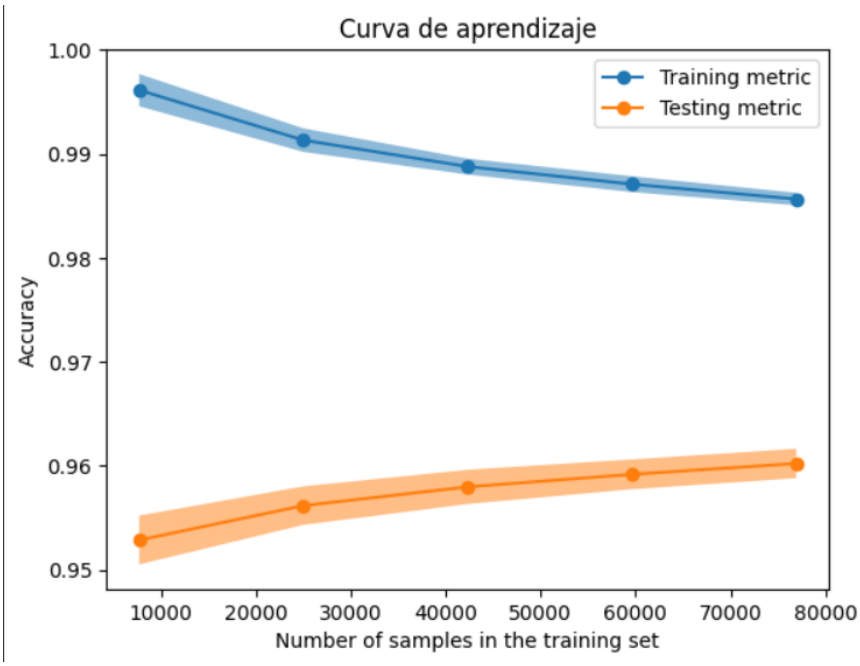


Figura 3.1 Curva de aprendizaje del modelo

- Comparación de aprendizaje en conjuntos de prueba y entrenamiento: Se pueden confirmar los datos de la gráfica anterior (figura 3.1) calculando el porcentaje de respuestas correctas e incorrectas en los datos de entrenamiento y testeo. En las gráficas de pie (figura 3.2) se puede observar que el modelo generaliza de mejor manera para los datos de entrenamiento, mientras que los datos de prueba tienen un mayor error.

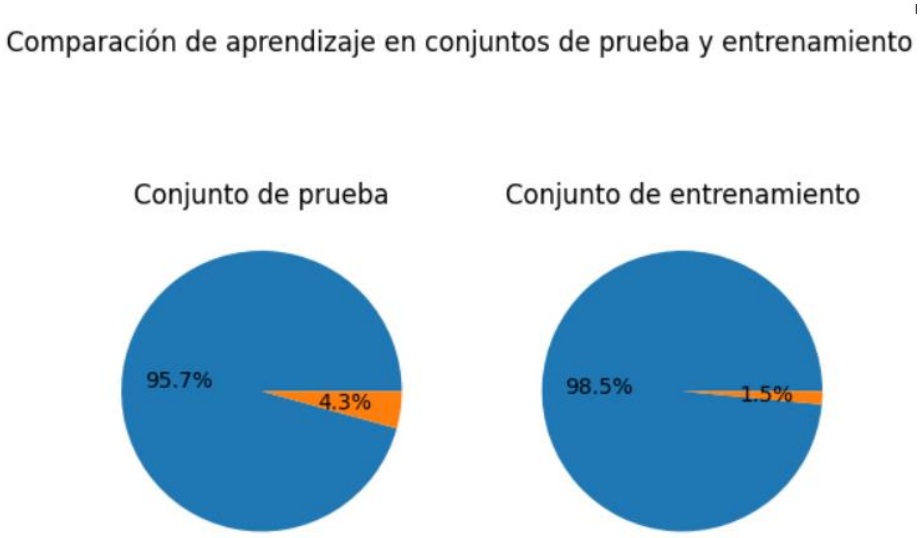


Figura 3.2 Comparación de aprendizaje en conjuntos de prueba y entrenamiento

- Sesgo y varianza promedio en 200 corridas: Si se calcula la varianza y sesgo del modelo 200 veces y obtenemos el promedio se puede tener una idea del comportamiento del modelo. En este caso la varianza y el sesgo son bajos, lo que indica que el modelo está valanceado

```
----- Sesgo y varianza promedio de 200 entrenamientos -----
Sesgo promedio: 0.032
Varianza promedio: 0.023
```

Figura 3.3 Salida de código para calcular varianza y sesgo promedio

Tomando en cuanta la información anterior se puede concluir que el modelo cuenta con muy poca varianza y sesgo, ya que las gráficas como los cálculos demuestran que el modelo generaliza bien para los datos de entrenamiento como de prueba.

Aunque el modelo es bueno, pero no perfecto, se pueden aplicar diferentes herramientas para perfeccionar las predicciones.

```python
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, precision_score, f1_score, confusion_matrix
from mlxtend.evaluate import bias_variance_decomp
import matplotlib.pyplot as plt
from sklearn.model_selection import LearningCurveDisplay, ShuffleSplit


# Árbol de decision
tree = DecisionTreeClassifier(
    criterion="entropy",
    max_depth=20
)

# Curva de aprendizaje
fig, ax = plt.subplots()
ax.set_title("Curva de aprendizaje")
common_params = {
    "X": X,
    "y": y,
    "train_sizes": np.linspace(0.1, 1.0, 5),
    "cv": ShuffleSplit(n_splits=50, test_size=0.2, random_state=0),
    "score_type": "both",
    "n_jobs": 4,
    "line_kw": {"marker": "o"},
    "std_display_style": "fill_between",
    "score_name": "Accuracy",
}
LearningCurveDisplay.from_estimator(tree, **common_params, ax=ax)
```

```python
# Exactitud en conjuntos de entrenamiento y testeo
tree.fit(X_train, y_train)

y_test_pred = tree.predict(X_test)
y_train_pred = tree.predict(X_train)

test_accuracy = accuracy_score(y_test, y_test_pred)
train_accuracy = accuracy_score(y_train, y_train_pred)

y_test_correct = (y_test == y_test_pred).sum()
y_train_correct = (y_train == y_train_pred).sum()

labels = ["Predicción correcta", "Prediccion erronea"]

fig, axs = plt.subplots(1, 2)
fig.suptitle("Comparación de aprendizaje en conjuntos de prueba y entrenamiento")
axs[0].pie([y_test_correct, len(y_test) - y_test_correct], autopct='%1.1f%%')
axs[0].set_title("Conjunto de prueba")
axs[1].pie([y_train_correct, len(y_train) - y_train_correct], autopct='%1.1f%%')
axs[1].set_title("Conjunto de entrenamiento")

# Obtener el sesgo y varianza promedio de 200 corrridas iguales
avg_expected_loss, avg_bias, avg_var = bias_variance_decomp(
    tree, X_train.to_numpy(), y_train.to_numpy(), X_test.to_numpy(), y_test.to_numpy(),
    loss='0-1_loss',
    num_rounds=200,
    random_seed=123)

print("----- Sesgo y varianza promedio de 200 entrenamientos -----")
print('Sesgo promedio: %.3f' % avg_bias)
print('Varianza promedio: %.3f' % avg_var)
```

Figura 3.4 Código de análisis de varianza y sesgo

4. Nivel de ajuste del modelo

Gracias al análisis de varianza y sesgo, se puede concluir que el modelo bueno, la varianza como el sesgo son bajos, lo que significa que el modelo está balanceado.

Aunque el modelo es bueno, no es perfecto. El árbol de decisión cuenta con un sesgo promedio de 0.023 y una variancia promedio de 0.032 (figura 3.3), esto se traduce en una ligera inclinación a un modelo con overfitting.

Utilizando las gráficas de pie (figura 3.2) se puede confirmar que el modelo cuenta con un ligero overfiting, ya que se observa claramente que el porcentaje de error con los datos de entrenamiento es menor al porcentaje de error de los datos de prueba. Esto quiere decir que el modelo está aprendiendo los detalles de los datos de entrenamiento y al momento de realizar predicciones nuevas el modelo no generaliza de forma correcta.

Por otra parte, la gráfica de curva de aprendizaje (figura 3.1) muestra un espacio entre las líneas de exactitud para los datos de entrenamiento y test. El espacio indica que hay variancia, mientras que el hecho de las líneas se comporte de manera asintótica a un valor de exactitud aproximado de 0.975 indica que el sesgo es menor. Una vez más se puede concluir que una mayor varianza y menor sesgo se traduce en un modelo con overfitting.

5. Técnicas de regularización o ajuste de parámetros

Para mejorar el desempeño del modelo y disminuir el sesgo y la varianza se aplicaron las siguientes técnicas:

1. Ajuste de parámetros: Se identificó que el árbol tiene un cierto nivel de overfitting, por lo que una forma de atacar este problema en los árboles de decisión es limitando el crecimiento del árbol. Antes del ajuste de parámetros, el árbol tenía una profundidad máxima de 20, lo que fomenta la memorización de los datos de entrenamiento. Se encontró que actualizar el valor a una profundidad máxima de 10 da mejores resultados, ya que el crecimiento del árbol se ve limitado y el modelo puede generalizar de mejor manera para datos nuevos.

El resultado de implementar estas técnicas el siguiente:

En las siguientes gráficas (figura 5.1) se aprecia una mejora en las métricas del modelo. Tanto la variancia como el sesgo se han reducido, la gráfica de curva de aprendizaje muestra una varianza menor y las gráficas de pie denotan un mejor balance en los porcentajes de error en las predicciones.

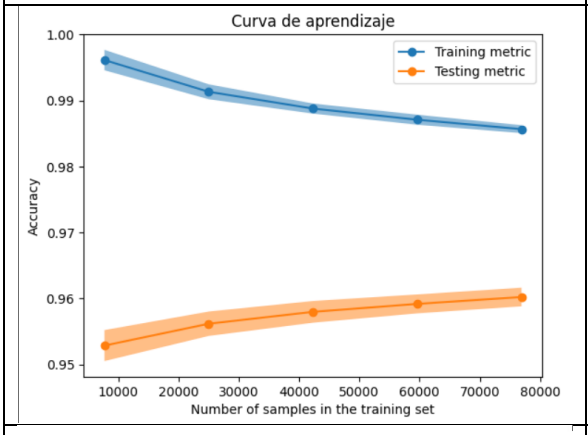
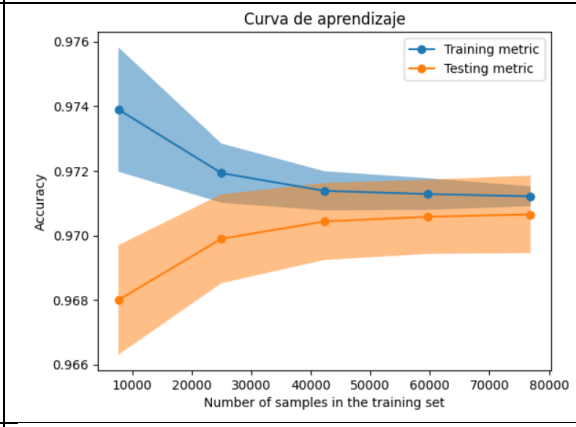
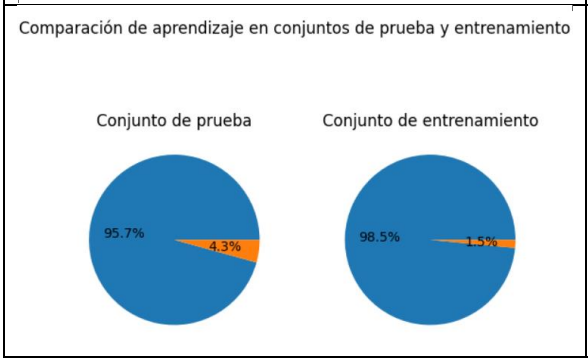
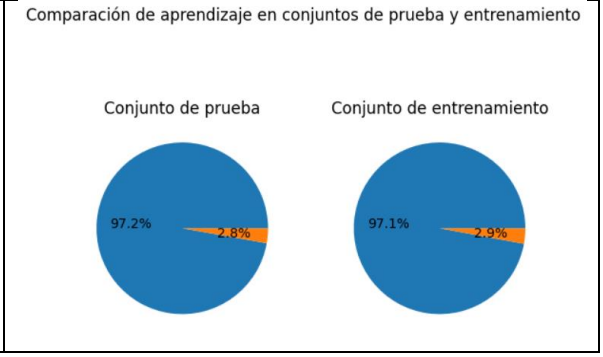
Resultados anteriores	Resultados nuevos
	
	

Figura 5.1 Comparación de resultados antes y después de técnicas de regularización o ajustes de parámetros