Cinvestav - Departamento de Computación

Minería de datos

Presenta:
Ángel Alonso Galarza Chávez

Profesora:
Dra. Xiaoou Li

Diciembre, 2024

Contenido

Introducción

Introducción

La clasificación consiste en entrenar modelos para aprender a reconocer patrones en los datos y asignar nuevos elementos a categorías específicas [1].

Consiste en aprender de la estructura de un conjunto de datos, que se encuentran particionando en clases, el aprendizaje de estas categoría´ıas se logra construyendo un modelo de clasificación que se utilizara para estimar los identificadores de las clases que existen en un conjunto de datos[2].

Conjunto de datos

El conjunto de datos Adult-Income Dataset, contiene 15 atributos de tipo numerico y categorico.

Numerico:

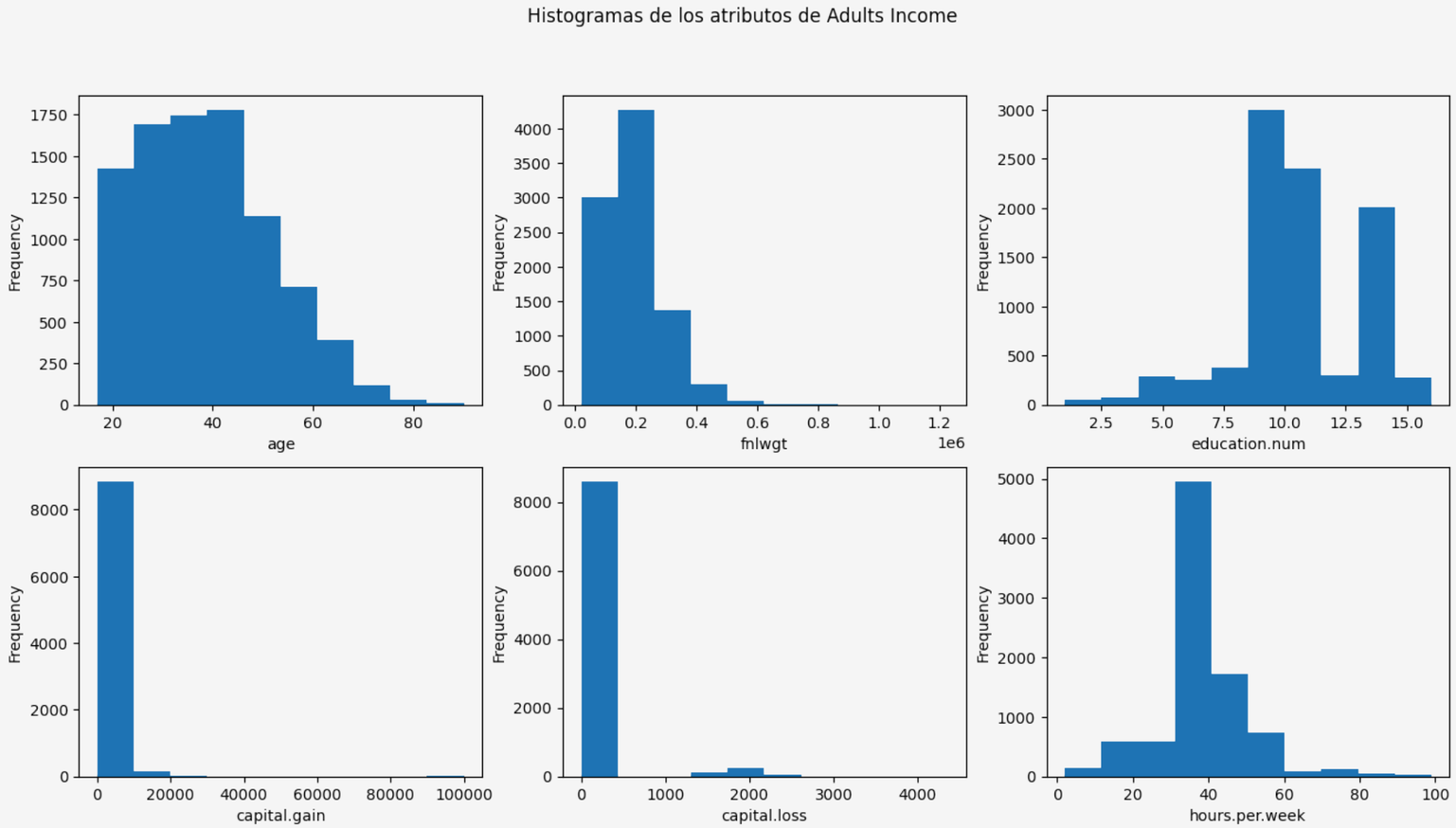
- age
- fnlwgt
- education-num
- capital-gain
- capital-loss
- hours-per-week

Categorico:

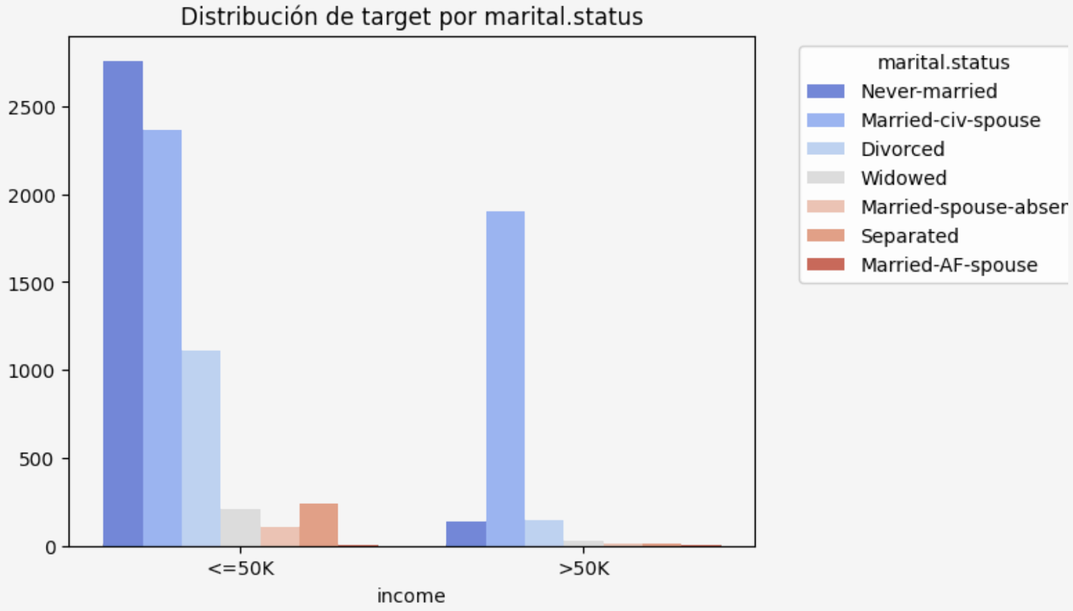
- workclass
- education
- marital-status
- occupation
- relationship
- race, sex
- native-country
- income

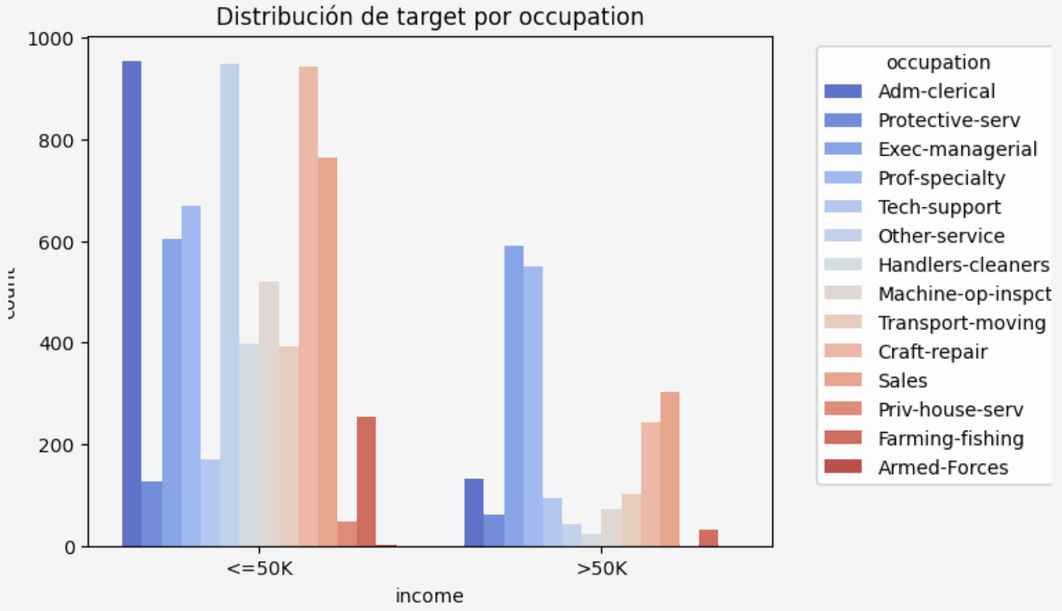
Conjunto de datos

Histogramas de los atributos categoricos.

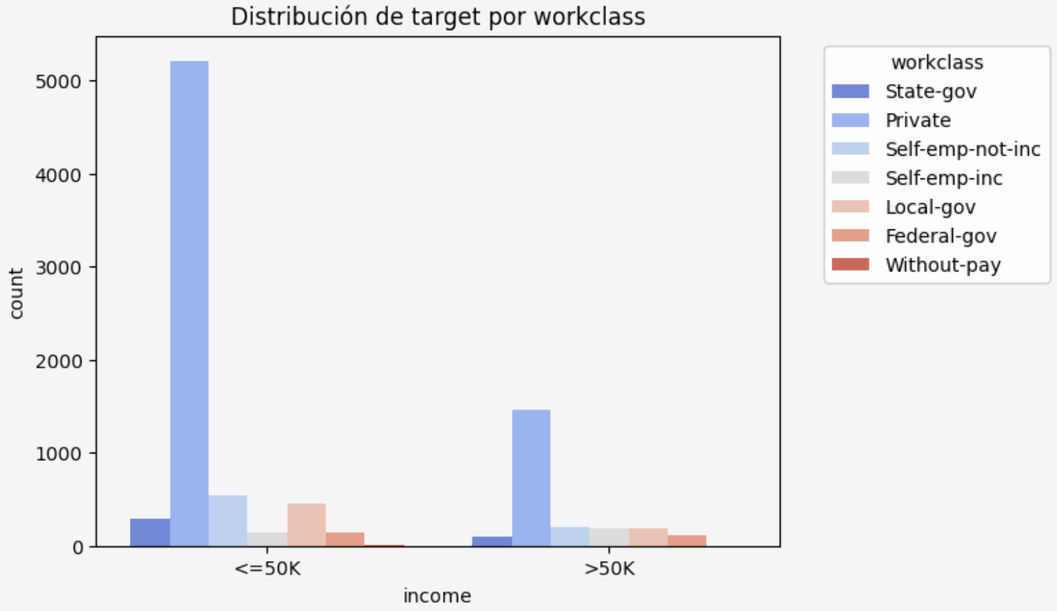


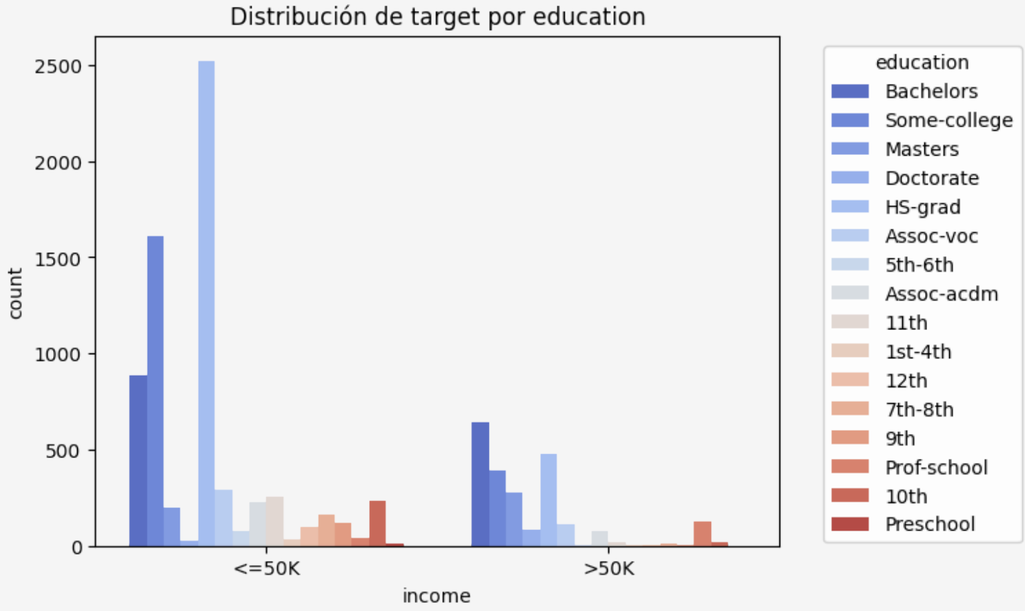
Conjunto de datos



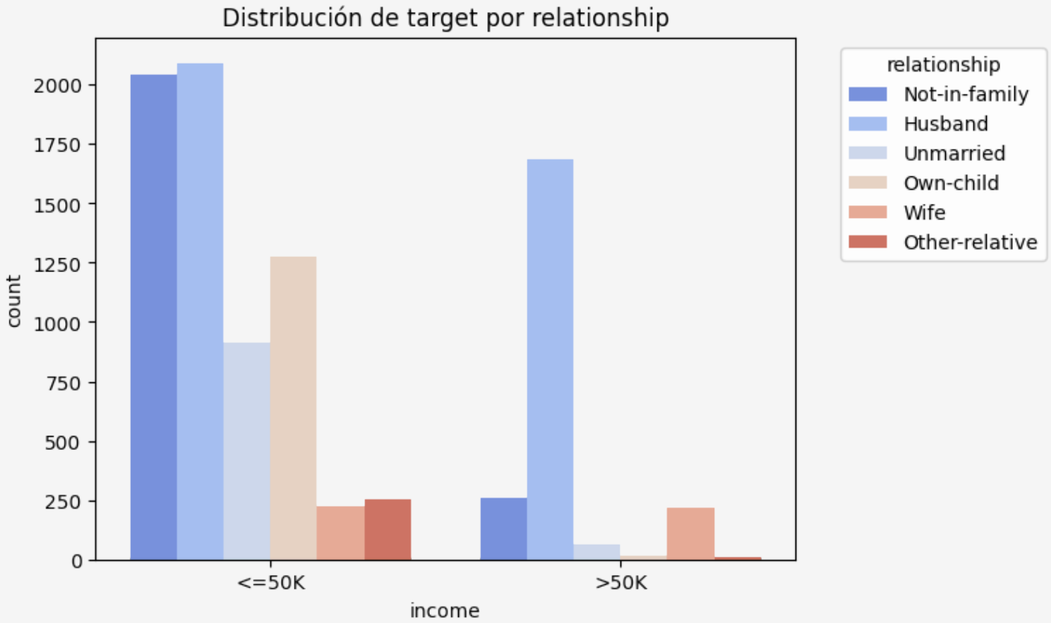


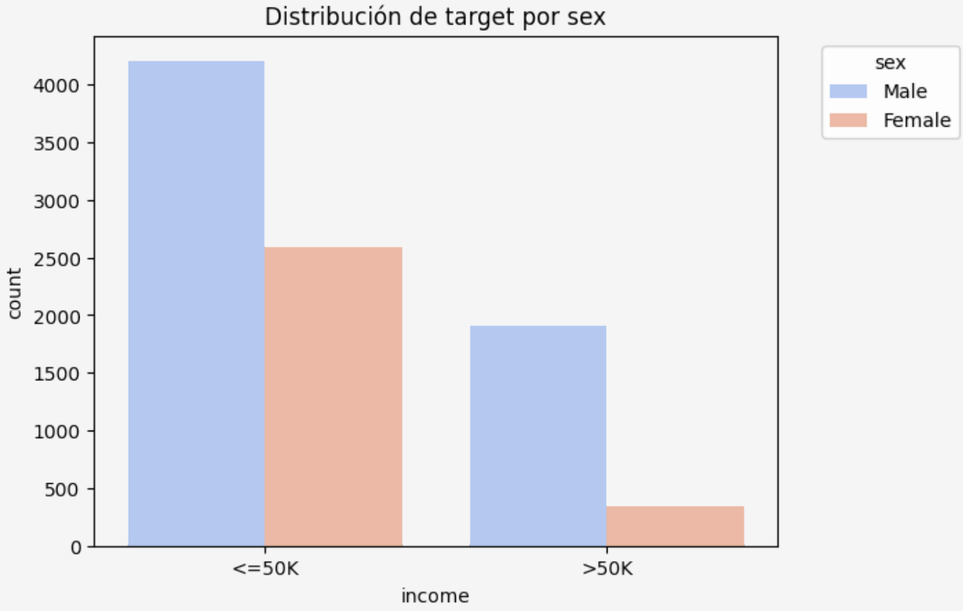
Conjunto de datos



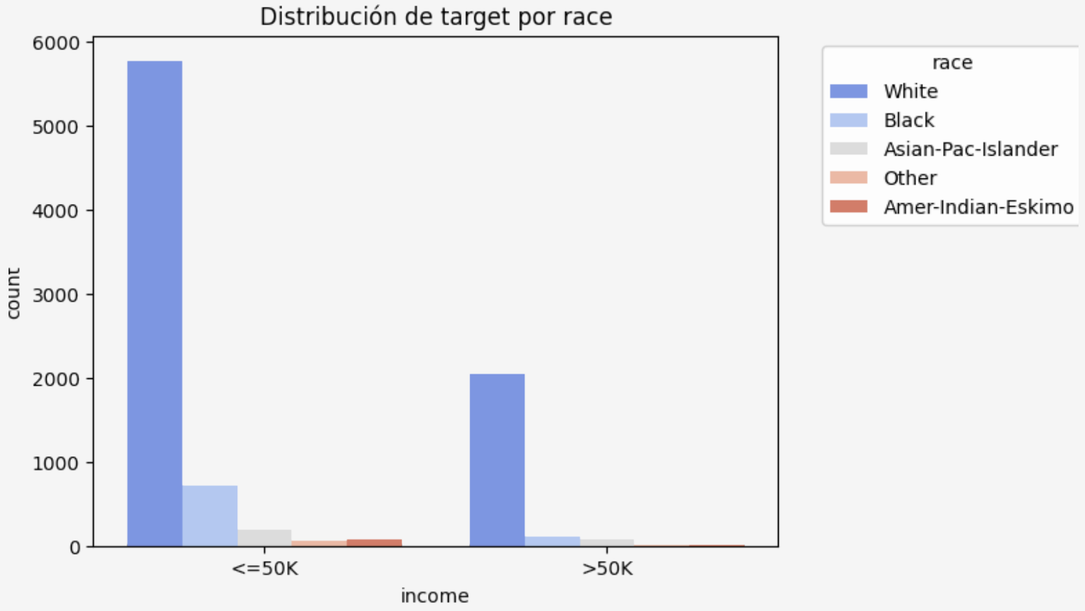


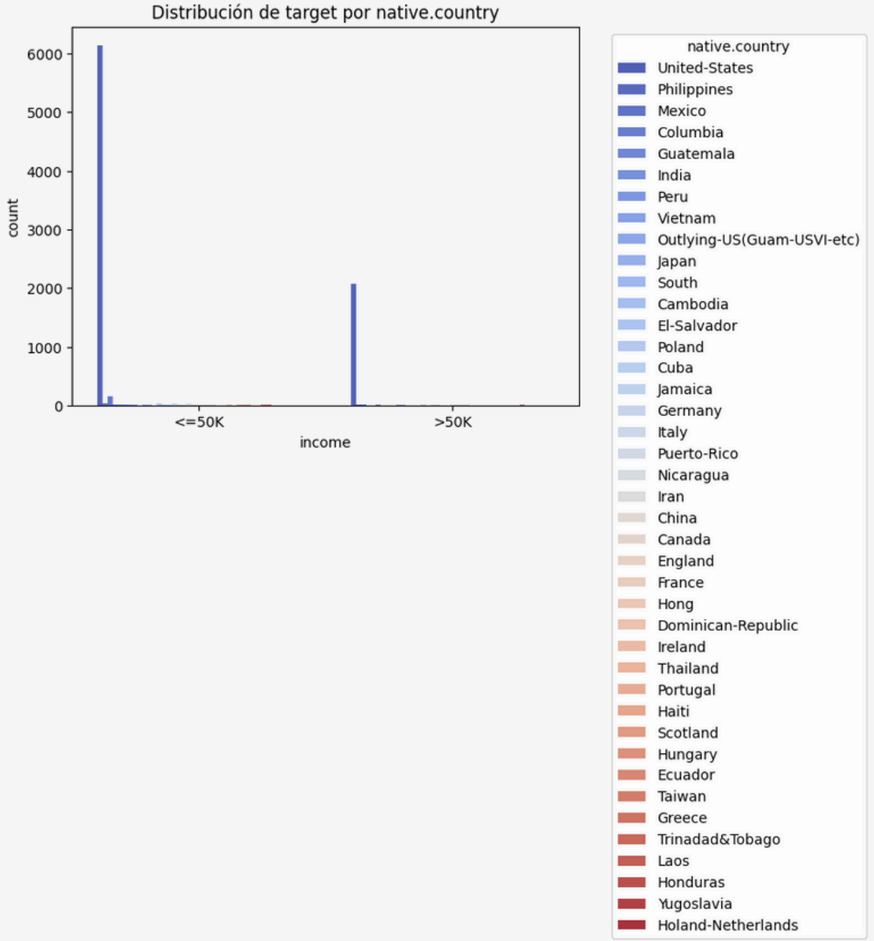
Conjunto de datos





Conjunto de datos



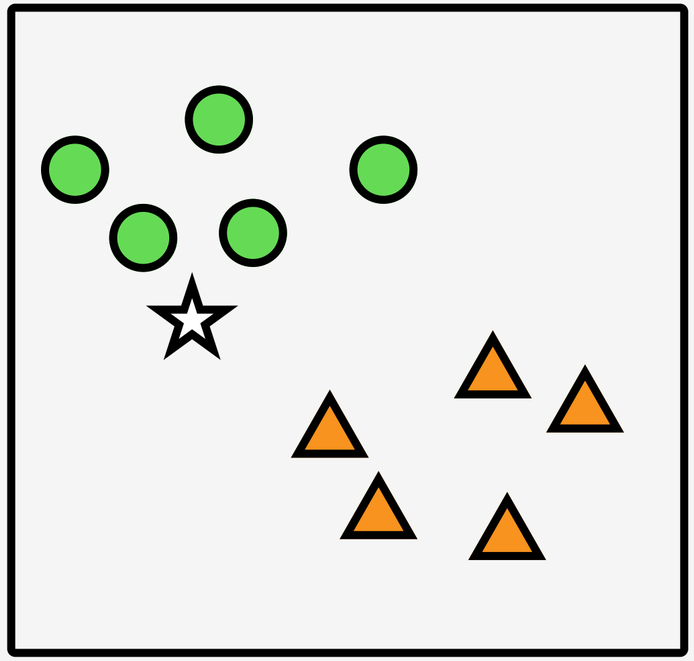


Algoritmos de clasificación

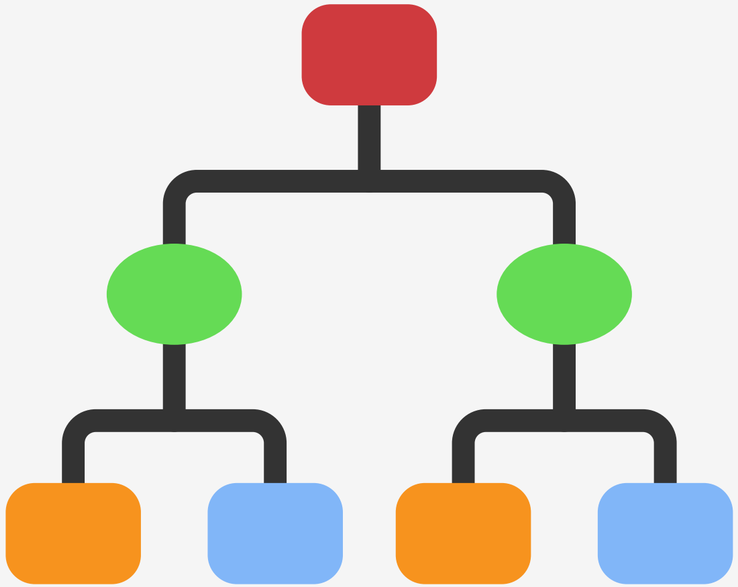
Los arboles de decisión son un algoritmo de clasificación, en la que el proceso de clasificación se realiza por medio de un conjunto de decisiones con jerarquía que se construyen a partir de los atributos de un conjunto de datos

Algoritmos de clasificación

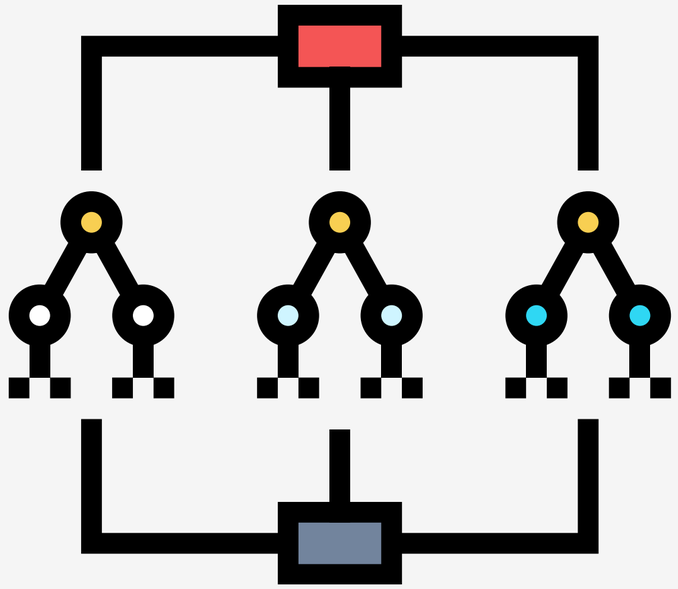
K-nearest-neighbor



Árbol de decisión



Random Forest



Experimentación

Herramienta	Versión	Sitio Web	Licencia
Python	3.10	https://www.python.org/	MIT
Jupyter Lab	3.5.3	https://jupyter.org/	CCO
Pandas	1.5.2	https://pandas.pydata.org/	BSD
Sckit-Learn	1.2.2	https://scikit-learn.org/stable/	BSD 3
Matplotlib	3.6.2	https://matplotlib.org/	BSD
Seasborn	0.12.2	https://seaborn.pydata.org/	BSD 3

Experimentación

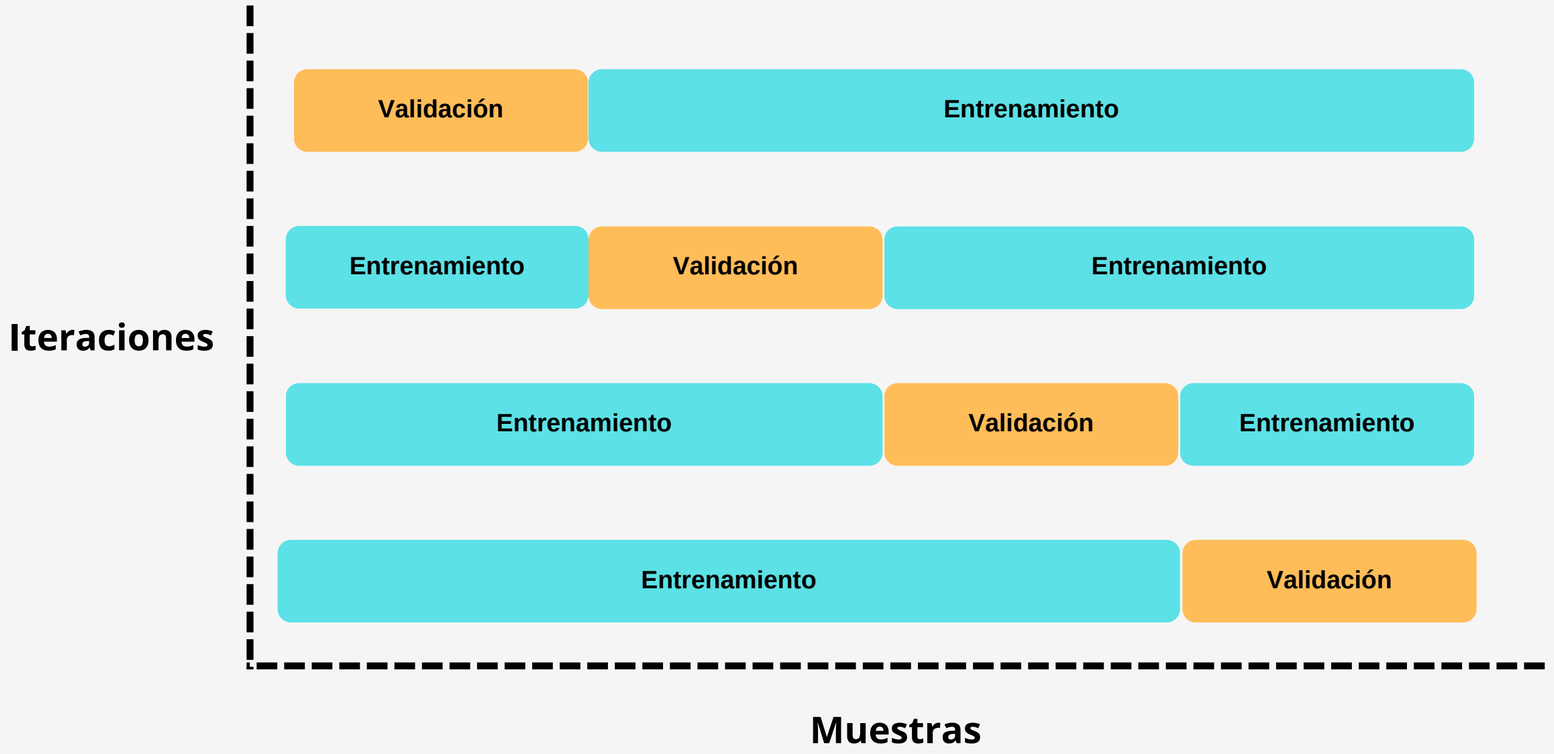
En la experimentación, se emplearon distintas técnicas para la construcción de los distintos modelos de clasificación. Obtener una representación de los datos que puede ayudar a obtener mejores resultados de precisión y exactitud para el modelo de clasificación.

ID	Atributo
1	A
2	B
3	C

ID	Atributo	A	B	C
1	A	**1**	0	0
2	B	0	**1**	0
3	C	0	0	**1**

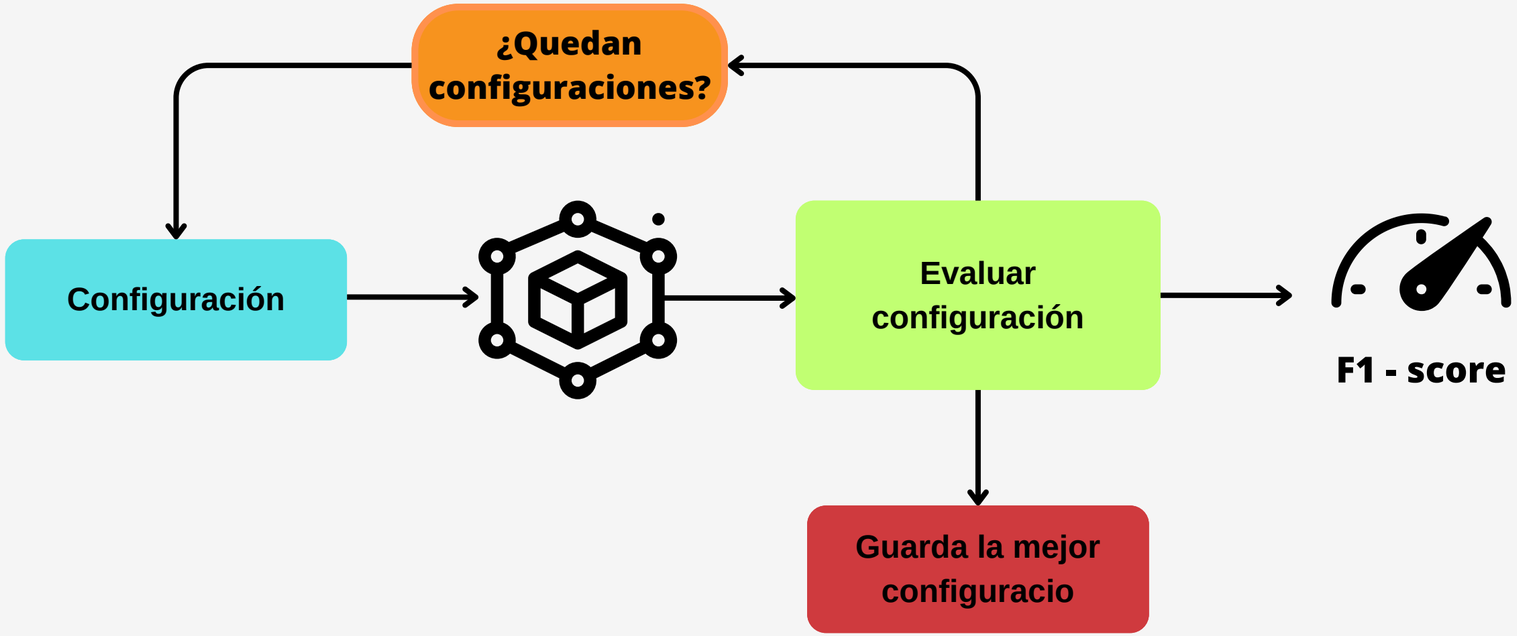
Experimentación

Con el conjunto de datos preprocesados, se realizo una partición estratificada del conjunto de datos en un 70% para entrenamiento y un 30% para evaluación



Experimentación

Los algoritmos de aprendizaje automático supervisado dependen en gran medida de la correcta configuración de sus hiperparametros para alcanzar un rendimiento optimo.



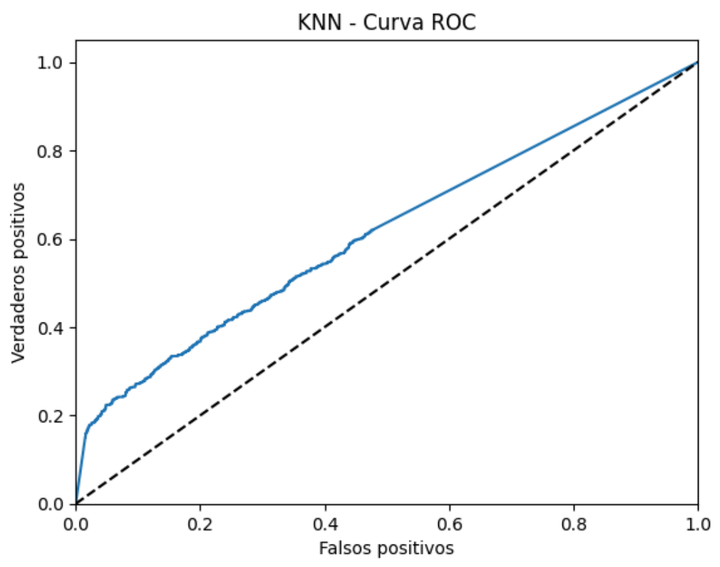
Resultados

Modelo	Clase	Precisión	Recall	F1-score
K-nearest-neighbor	0 1	0.80 0.43	0.85 0.35	0.82 0.38
Árbol de decisión	0 1	0.87 0.73	0.93 0.58	0.90 0.65
Random Forest	0 1	0.88 0.79	0.96 0.53	0.92 0.64

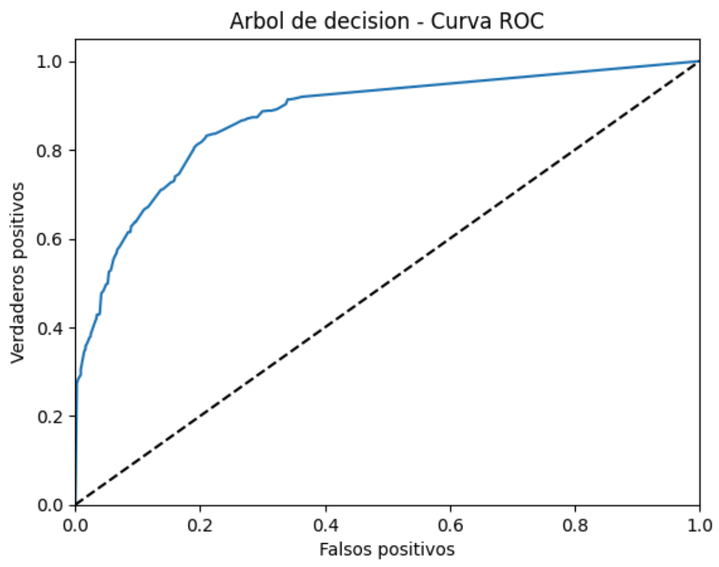
Resultados

Curvas ROC de los modelos de clasificación

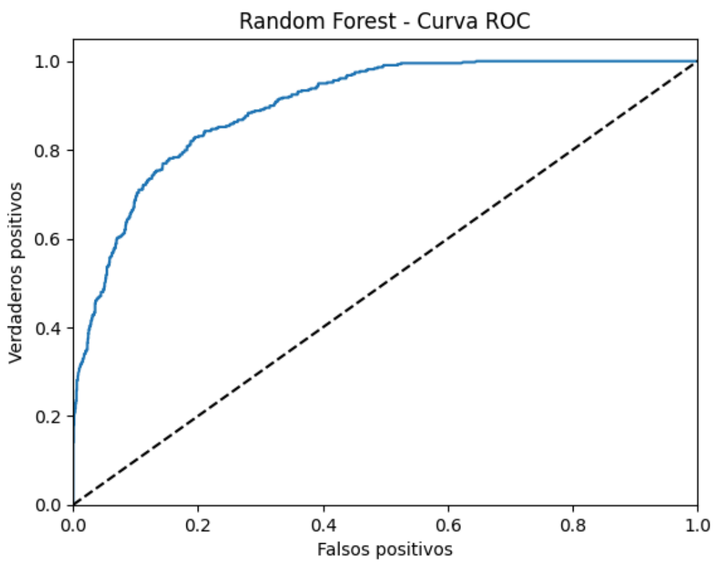
K-nearest-neighbor



Árbol de decisión



Random Forest



Conclusiones

Con los resultados presentados se puede concluir que el modelo de Random Forest es el mas adecuado para este problema de clasificación binaria. Su capacidad de generalizar y su buen desempeño en ambas clases lo convierten en una excelente opción.

- la elección del mejor modelo siempre depende del conjunto de datos especifico.
- Realizar diferentes preprocesamientos para la construcción de los modelos de clasificacion
- Ajustar los hiperparametros de los modelos de clasificación ayudan a obtener mejores resultados.

Bibliografia

[1] Charu C Aggarwal. Data mining. en. 2015th ed. Cham, Switzerland: Springer International Publishing, Apr. 2015.

[2] Jiawei Han, Micheline Kamber, and Jian Pei. Data mining concepts and techniques, third edition. 2012. url: http://www.amazon.de/Data- Mining- Concepts- Techniques- Management/dp/0123814790/ref=tmm_hrd_title_0?ie=UTF8&qid=1366039033&sr=1-1.