Proyecto IA

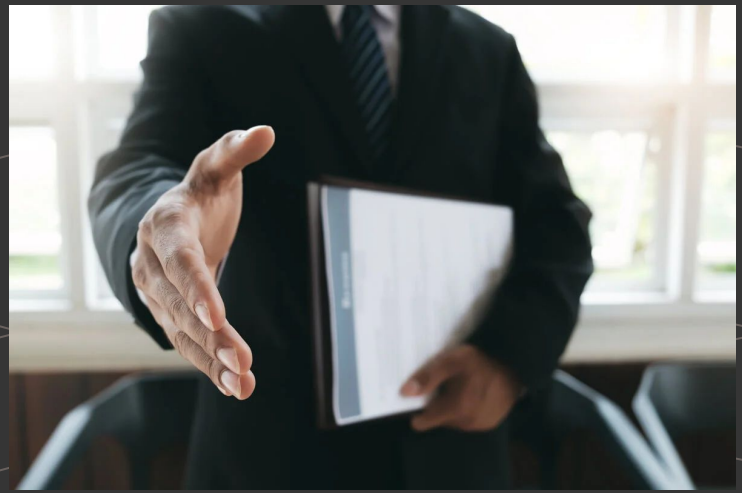
Predicción del éxito laboral en estudiantes de diferentes áreas del conocimiento

Daniel Alejandro Ayala Vallejo 2220084
Nelson Felipe Moreno 2220064
David Fernando Naranjo 2220046

Introducción

Este proyecto aborda un problema de clasificación utilizando el dataset Education & Career Success, que contiene información de 5.000 estudiantes, como edad, género, universidad, promedio, pasantías, habilidades y ofertas laborales.

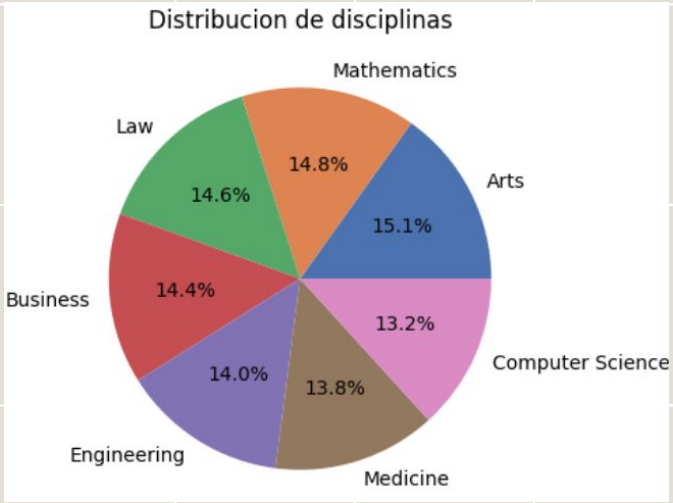
El objetivo es predecir una variable categórica, es decir, conocer el cargo laboral que tendrá un estudiante en su primer trabajo, aplicando modelos clásicos de clasificación: Árbol de Decisión (DT), Random Forest (RF) y SVM, y evaluando su desempeño con métricas y curvas de aprendizaje.

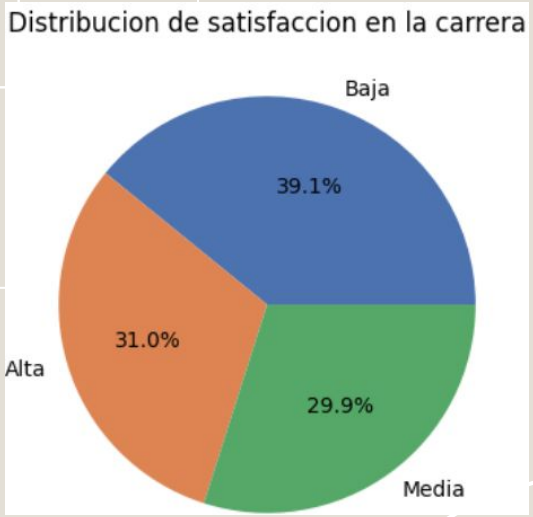


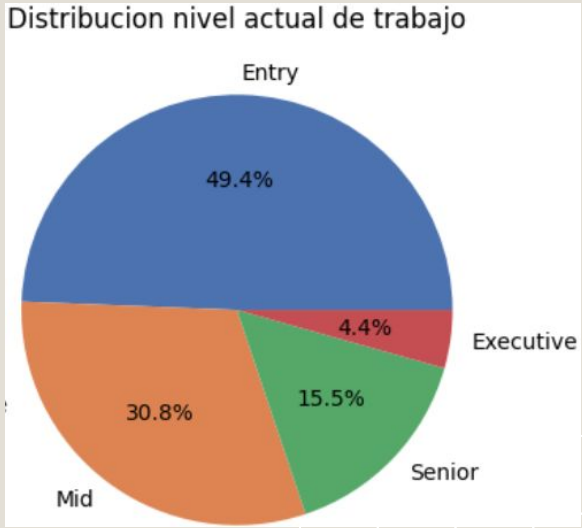
Limpieza de los datos

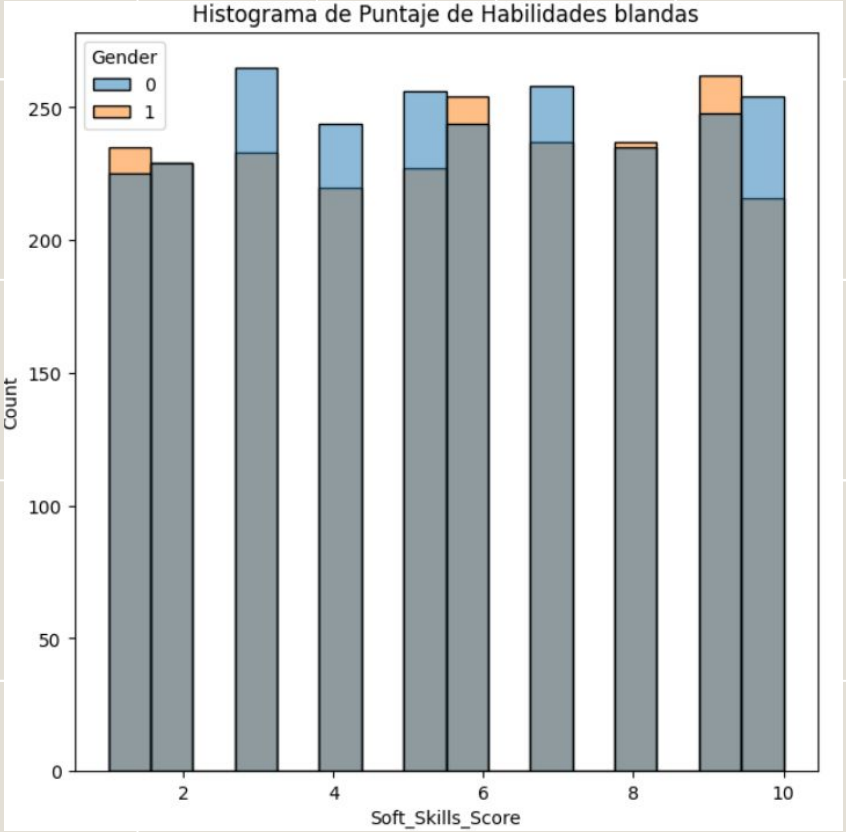
- Se elimina cualquier dato que sea nulo.
- Se elimina el dato “Other” para la columna Género.
- Se reemplazan “Male” y “Female” por 0 y 1 respectivamente.

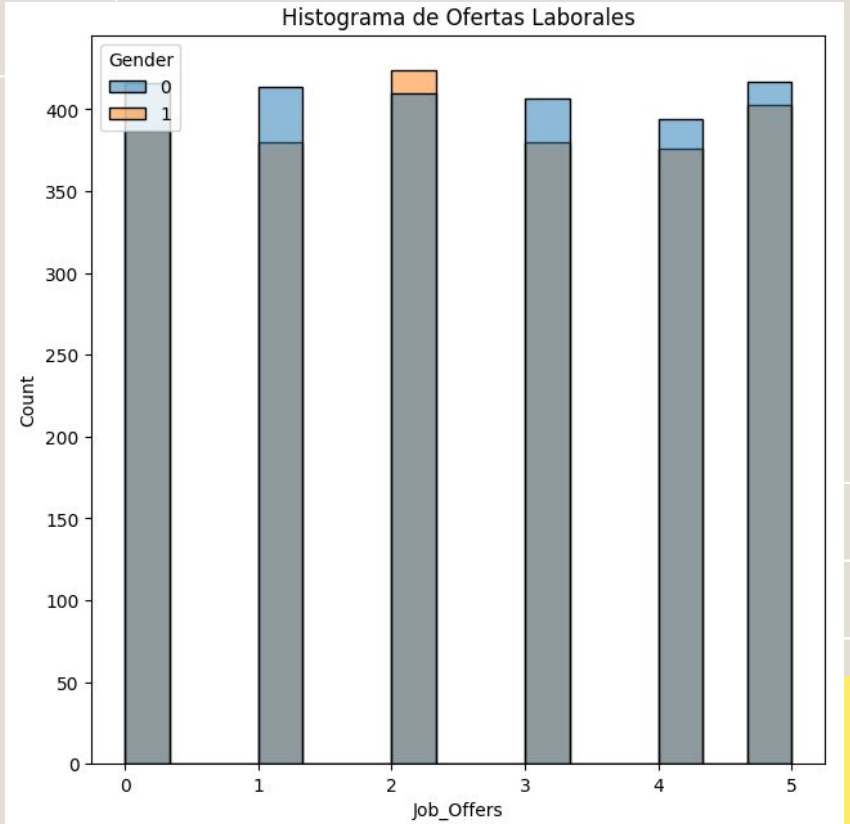
Distribución de algunas columnas








Histograma de Puntaje de Habilidades blandas
Gender
0
1
Count
Soft_Skills_Score
0
50
100
150
200
250
2
4
6
8
10

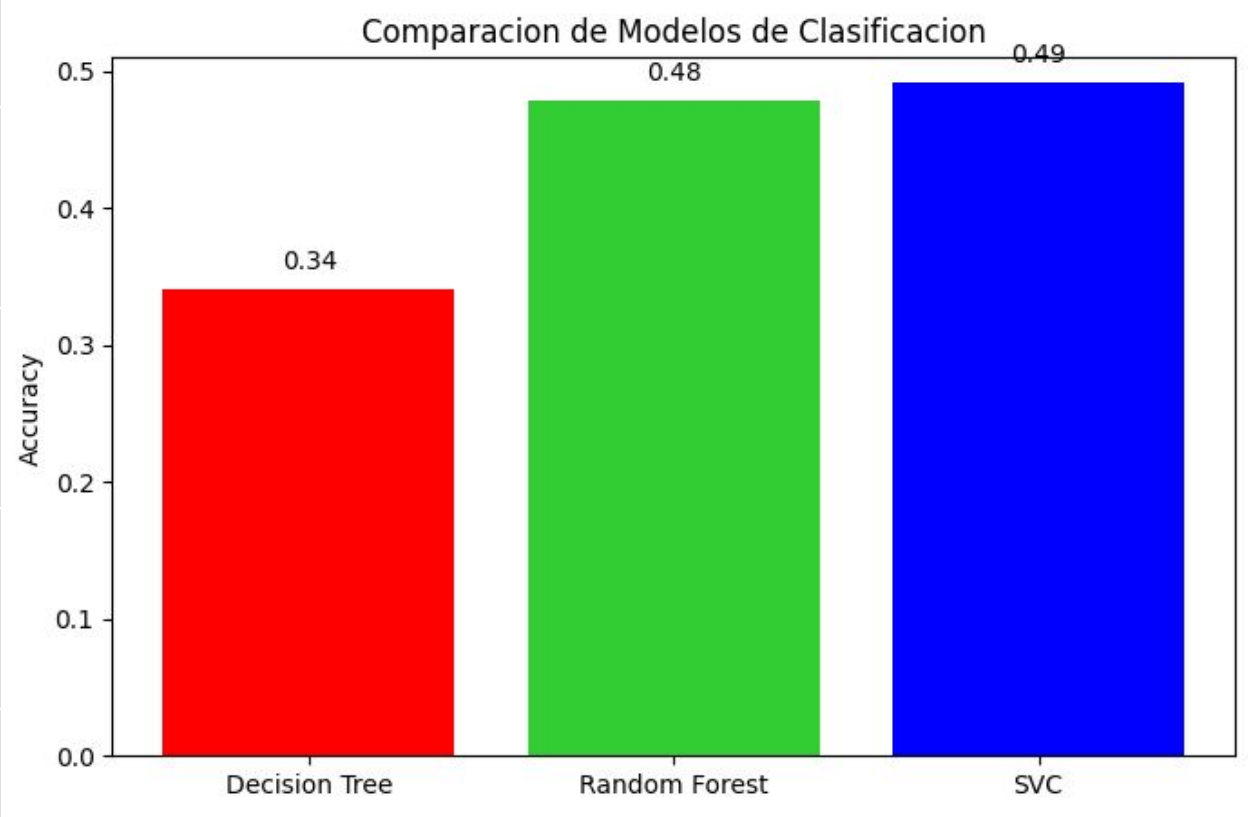

Histograma de Ofertas Laborales
Gender
0
1
Count
Job_Offers
0
50
100
150
200
250
300
350
400
0
1
2
3
4
5

CLASIFICACIÓN

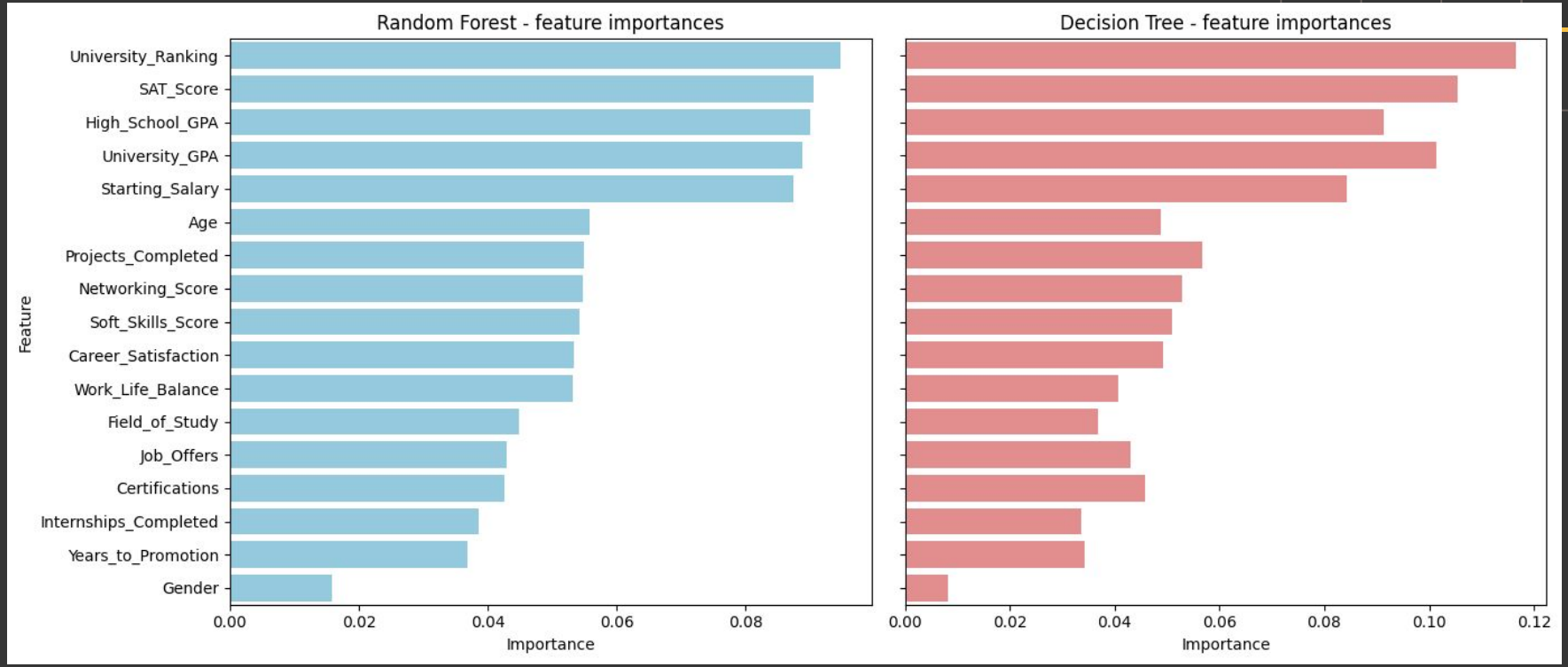
Para este conjunto de datos se realizó un proceso de clasificación a través de Machine Learning, con el fin de predecir la clase o ground truth “Current Job Level” para los estudiantes universitarios. Para ello se implementaron los métodos:

- Decision tree
- Random forest
- SVC (Support Vector Machine)

Adicionalmente como pre procesamiento se empleó el método “train_test_split” para seguir el pipeline recomendado por el docente y tener mejor desempeño en el proceso de clasificación.


Comparacion de Modelos de Clasificacion
Accuracy
0.0
0.1
0.2
0.3
0.4
0.5
0.34
0.48
0.49
Decision Tree
Random Forest
SVC

Feature importances



Curvas de aprendizaje

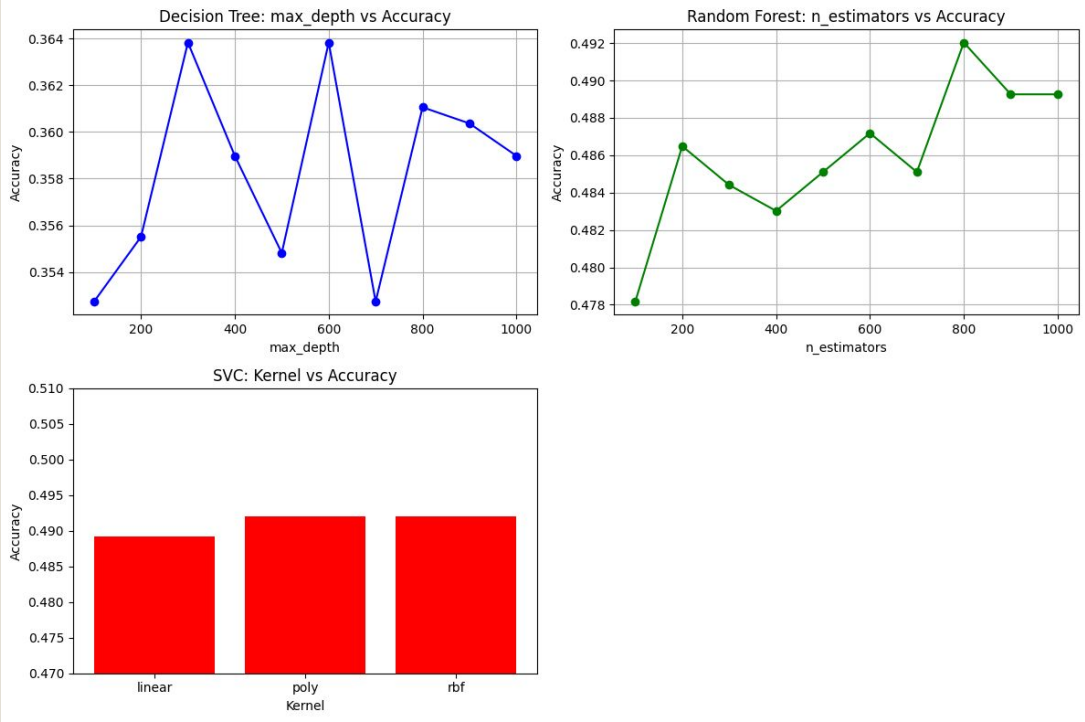
A continuación se realizó un proceso de curvas de aprendizaje para optimizar los procesos de clasificación de los 3 métodos mostrados anteriormente encontrando el MEJOR valor para el hiper parámetro clave a la hora de entrenar con cada uno

Árbol de Decisión (DT):
Se varió el valor de “max_depth” que indica que tan profundo ira el Decision tree para entrenar con el conjunto de datos.

Random Forest (RF):
Se varió el parámetro “n_estimators” que indica el número de árboles de decisión que poseerá el método de RF.

SVM (SVC):
Se varió el “kernel” utilizado por el metodo para realizar el proceso de clasificación con el dataset.

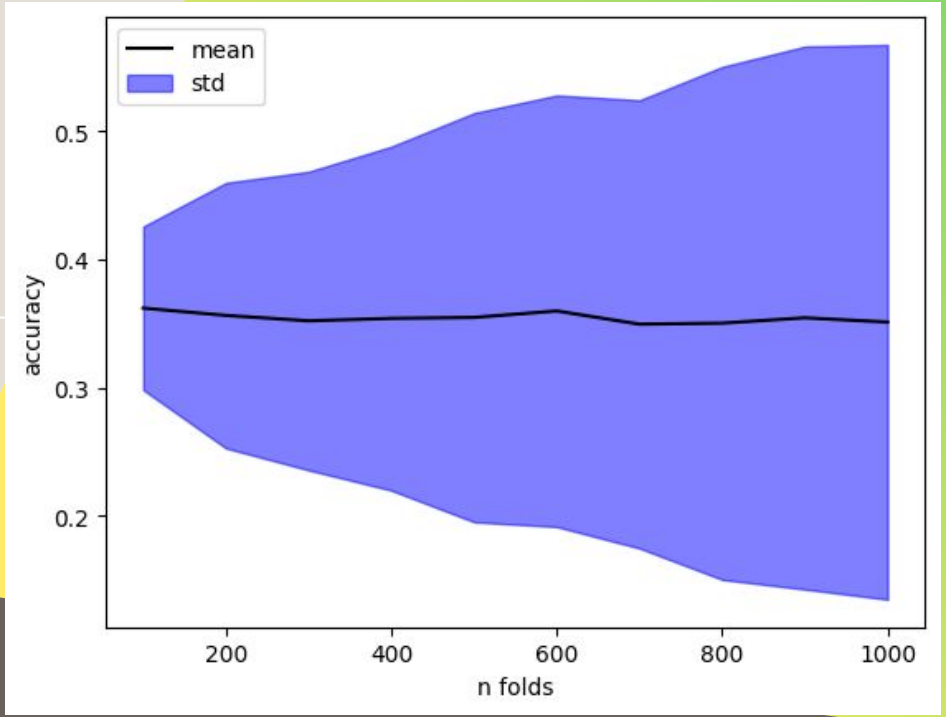
Resultado


Decision Tree: max_depth vs Accuracy
Accuracy
max_depth
0.364
0.362
0.360
0.358
0.356
0.354
200
400
600
800
1000
Random Forest: n_estimators vs Accuracy
Accuracy
n_estimators
0.492
0.490
0.488
0.486
0.484
0.482
0.480
0.478
200
400
600
800
1000
SVC: Kernel vs Accuracy
Accuracy
Kernel
0.510
0.505
0.500
0.495
0.490
0.485
0.480
0.475
0.470
linear
poly
rbf

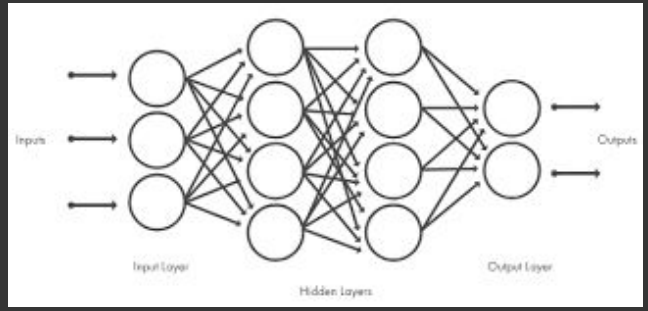
Cross Validation

Para este apartado, se volvera a realizar un procedimiento de curvas de aprendizaje pero unicamente para el metodo Decision tree, utilizando adicionalmente la funcion de validacion cruzada cross_val_score() para asi poder obtener el numero de ventana mas apropiado para que al igual se optimice el proceso de clasificacion.

Resultado



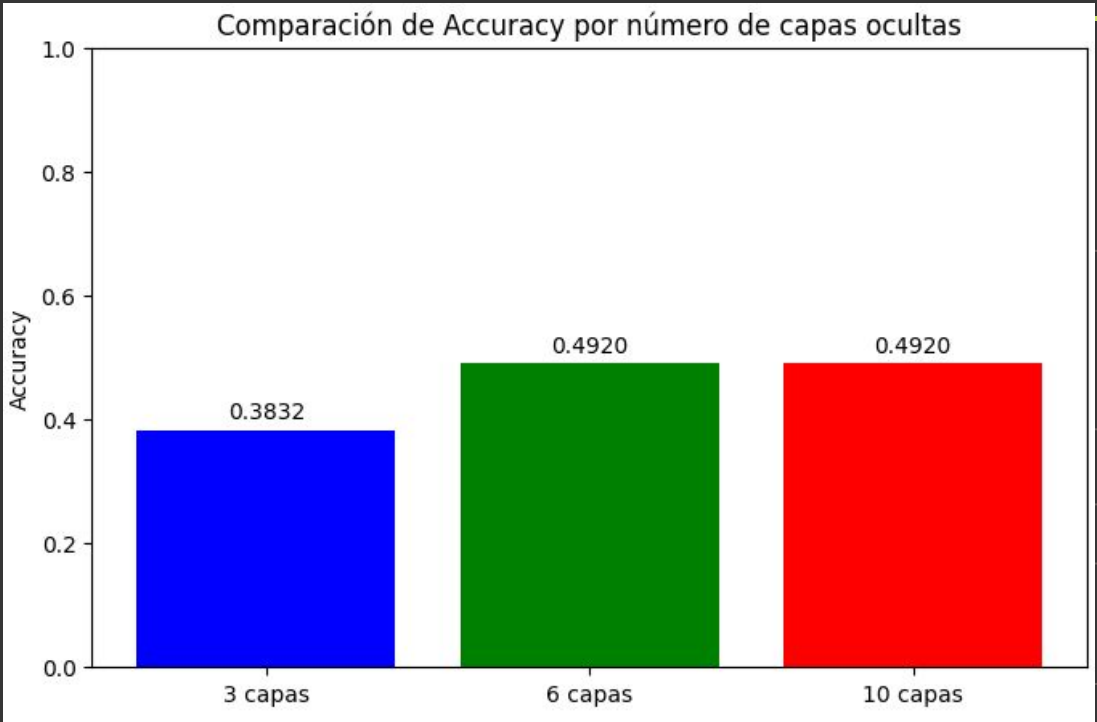
Deep Learning



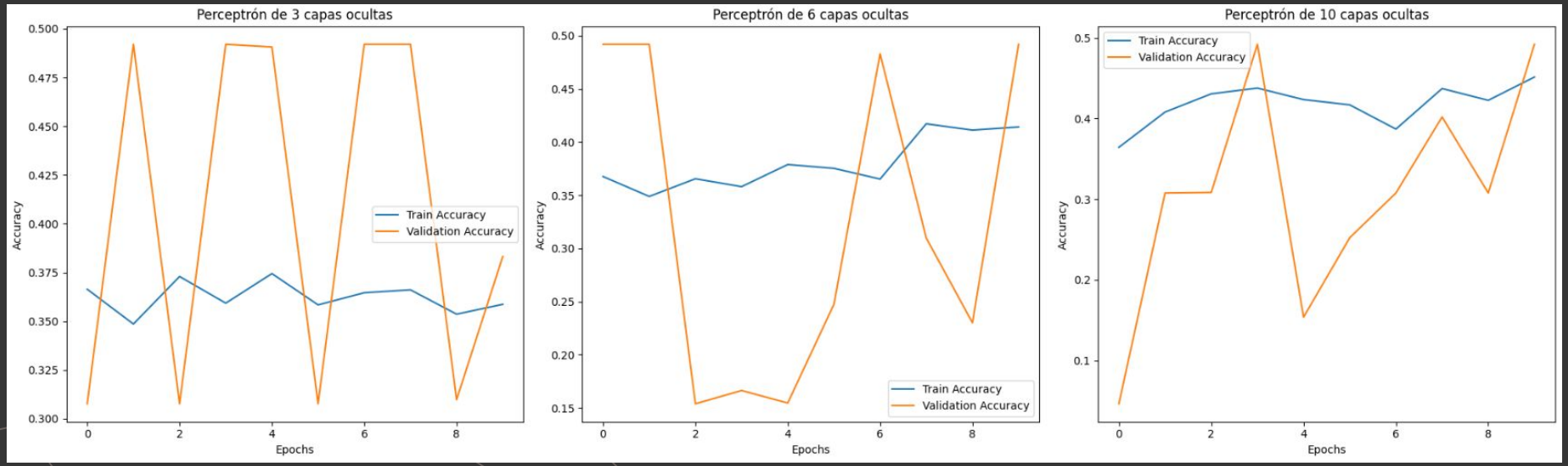
La implementación cuenta con 3 modelos, cada uno con 3, 6 y 10 capas de neuronas ocultas respectivamente. La métrica usada para evaluar el proceso de clasificación fue “accuracy”.

Para realizar Aprendizaje Profundo, se implementó un perceptrón multicapa con ayuda de la librería TensorFlow, esto para realizar una segunda vez un proceso de clasificación teniendo en cuenta el mismo ground truth “Current Job Level”.

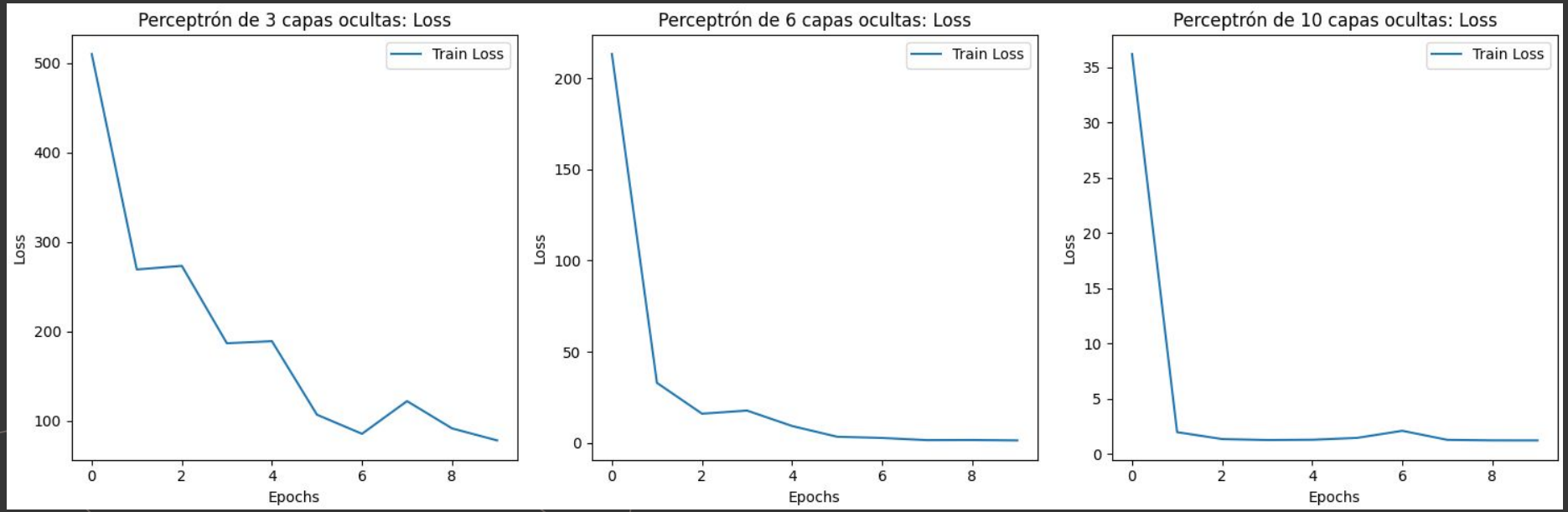
Resultado


Comparación de Accuracy por número de capas ocultas
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.3832
0.4920
0.4920
3 capas
6 capas
10 capas

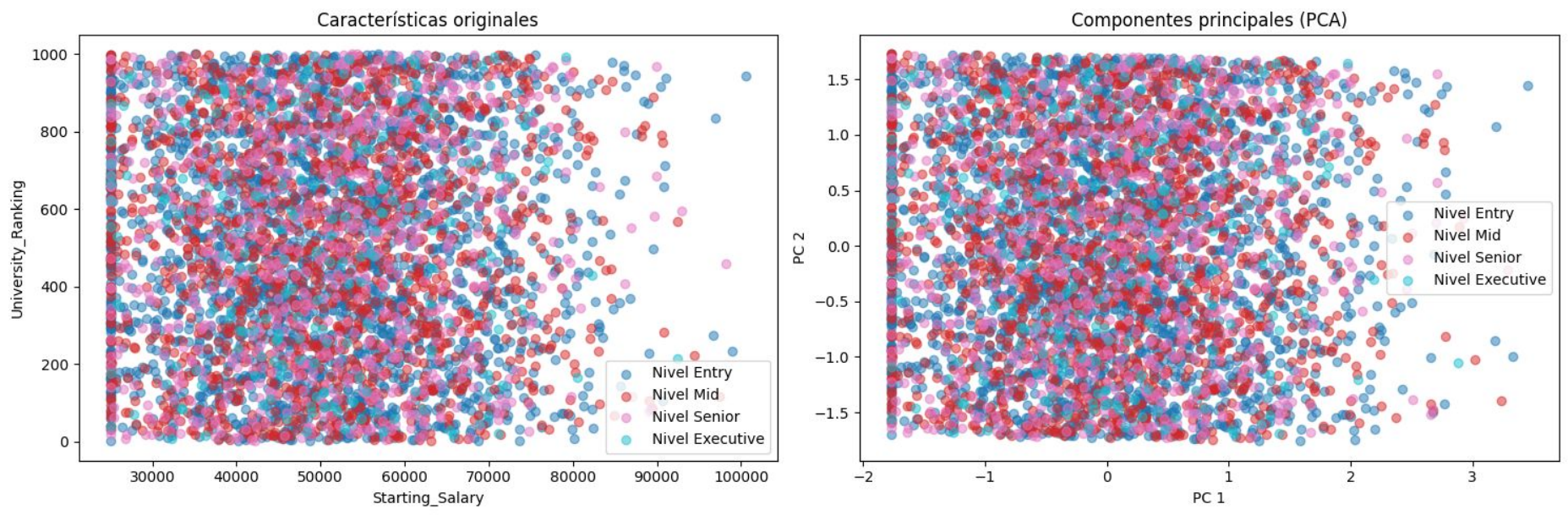
Accuracy vs Epochs



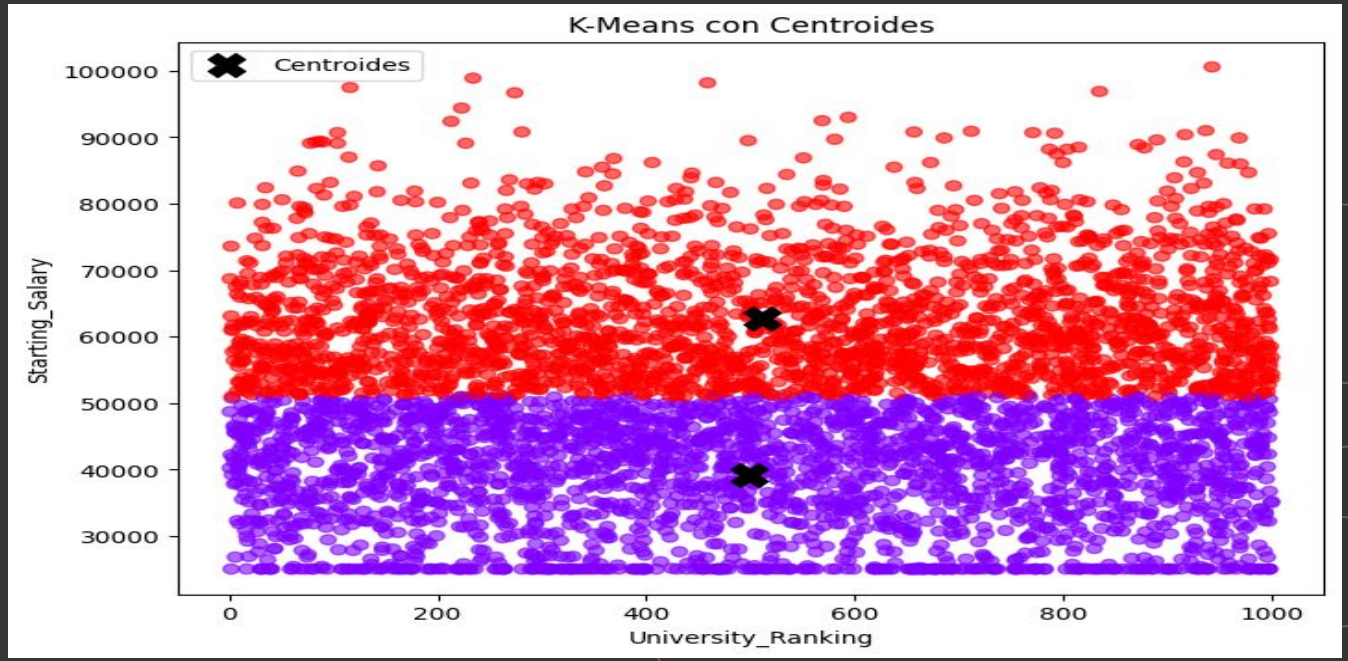
Loss vs Epochs



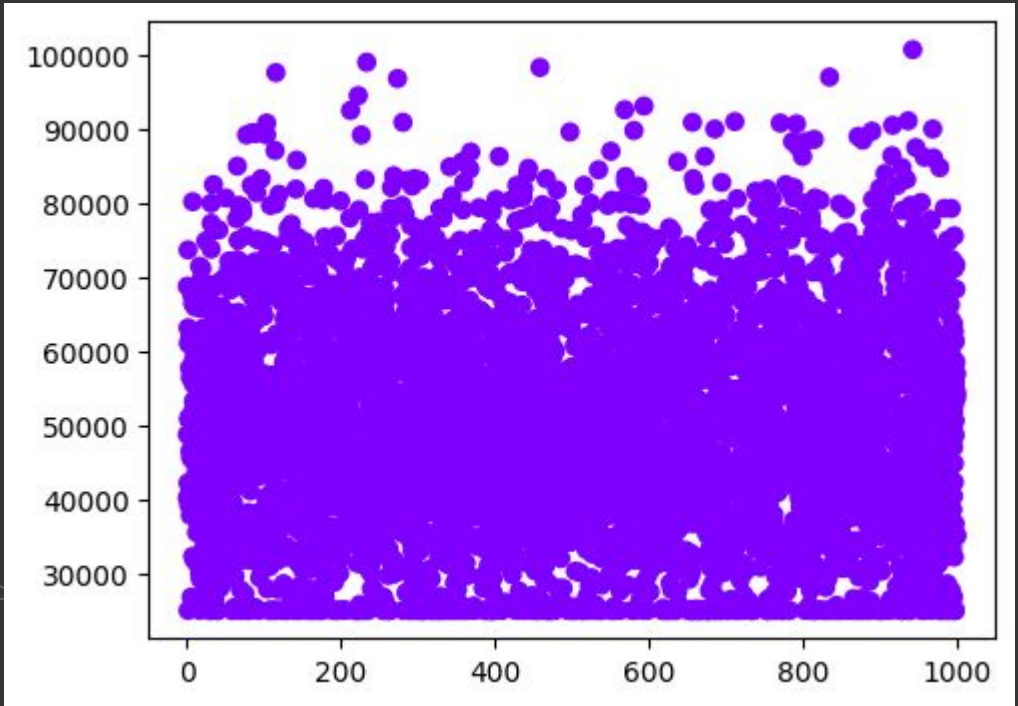
PCA


Características originales
University_Ranking
Starting_Salary
Nivel Entry
Nivel Mid
Nivel Senior
Nivel Executive
Componentes principales (PCA)
PC 2
PC 1
Nivel Entry
Nivel Mid
Nivel Senior
Nivel Executive

K-Means con Centroides



DBSCAN


100000
90000
80000
70000
60000
50000
40000
30000
0
200
400
600
800
1000