UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

Maestría en Ciencia de Datos

Aprendizaje Automático

Alvaro Pequeño Mondragón
1726520

Reporte #3

21 de julio de 2024

Para la practica 3 se requiere realizar un modelo de clasificación por lo que se decidió realiza un árbol de decisiones para poder determinar en base al numero de empleados y a la cantidad de pago promedio si la dependencia es una preparatoria o una facultad.

Primero se realiza el modelo de un árbol sin la validación cruzada:

```
#Se mantienen los datos a utilizar en el modelo de clasificación, en este caso se guardan los datos donde se incluye el tipo de facultad y preparatoria:
model3 = sueldo_dependencia.loc[(sueldo_dependencia['tipo'] == 'PREPARATORIA') | (sueldo_dependencia['tipo'] == 'FACULTAD')]
model3_x = model3.drop(['dependencia', 'tipo', 'sueldo_neto'], axis = 1, inplace = False)
model3_y = model3.drop(['dependencia', 'sueldo_neto', 'num_empleados', 'pago_diario_promedio'], axis = 1, inplace = False)
```

[4]

```
model3_xtrain, model3_xtest, model3_ytrain, model3_ytest = train_test_split(model3_x, model3_y, test_size=0.2, random_state=73)
model3_decision_tree = DecisionTreeClassifier(max_depth=30, random_state=40)
model3_decision_tree = model3_decision_tree.fit(model3_xtrain, model3_ytrain)
model3_ypred = model3_decision_tree.predict(model3_xtest)
model3_accuracy = accuracy_score(model3_ytest, model3_ypred)
model3_class_report = classification_report(model3_ytest, model3_ypred)
print("Accuracy", model3_accuracy)
print(model3_class_report)
```

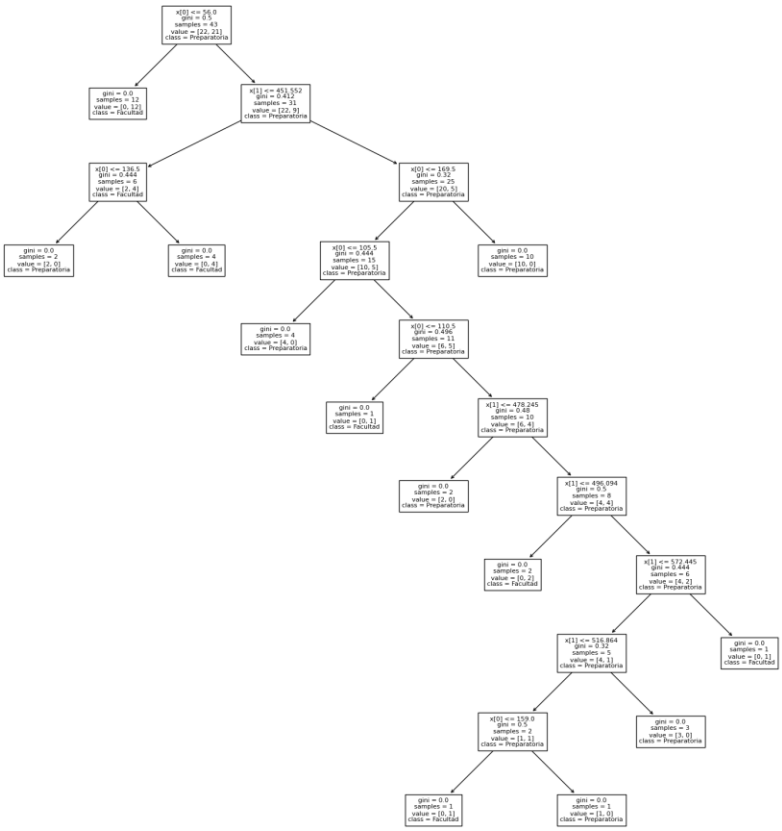
[5]

```
Accuracy 0.45454545454545453
              precision    recall  f1-score   support

     FACULTAD       0.38      0.75      0.50         4
PREPARATORIA        0.67      0.29      0.40         7

     accuracy                           0.45        11
    macro avg       0.52      0.52      0.45        11
 weighted avg       0.56      0.45      0.44        11
```

Como se puede ver el modelo tiene un desempeño muy bajo y a continuación se muestra el árbol obtenido:



Por último, se realiza la validación cruzada dónde se obtienen los mejores parámetros para resolver este problema de clasificación, la métrica elegida es la de accuracy ya que me interesa saber qué porcentaje de todas las clasificaciones que hizo el modelo fueron correctas :

Con lo anterior podemos observar que los mejores parámetros para el árbol de decisión son los siguientes:

- Max_depth = 5
- Random_state = 0

Lo anterior nos da un accuracy de 0.8142 lo cual es una gran mejora respecto al primer modelo realizado sin la validación cruzada.