1. Introduction

Welcome to Predicting Job Offer Probability project!

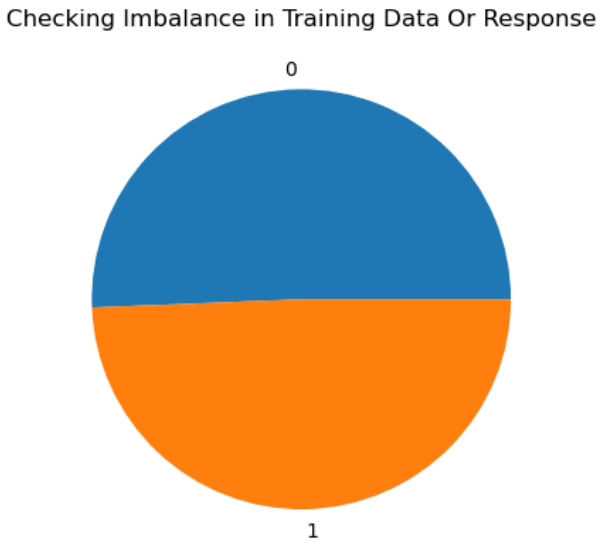
This is a tool which define the probability of a student receiving a job offer

2. Model Overview

I used Random Forest Classifier with (max_depth = 2, min_samples_split = 2, n_estimators = 100) to predict the likelihood for each student

3. Data Preparation

3.1. By checking the balance of this dataset shown that it's balanced according to the following chart:



3.2. By checking the dataset information and its features , determined that there are no missing values, but there is only **two duplicate** rows and one **categorical feature** (skills feature).

So to prepare this dataset for feeding it to the model I did the following:

- Remove duplicates
- Handle categorical feature using get_dummies function

- Applying feature scaling using StandardScaler package , so the all features will be in the same range which is good for training

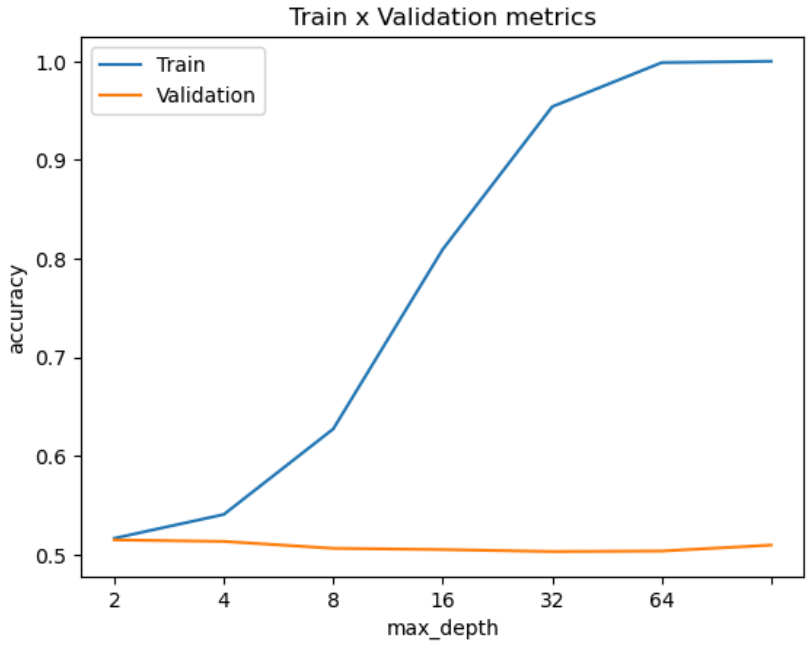
4. Training Process

As the random forest algorithm has different parameters , I tried different values for the following parameters to get highest accuracy:

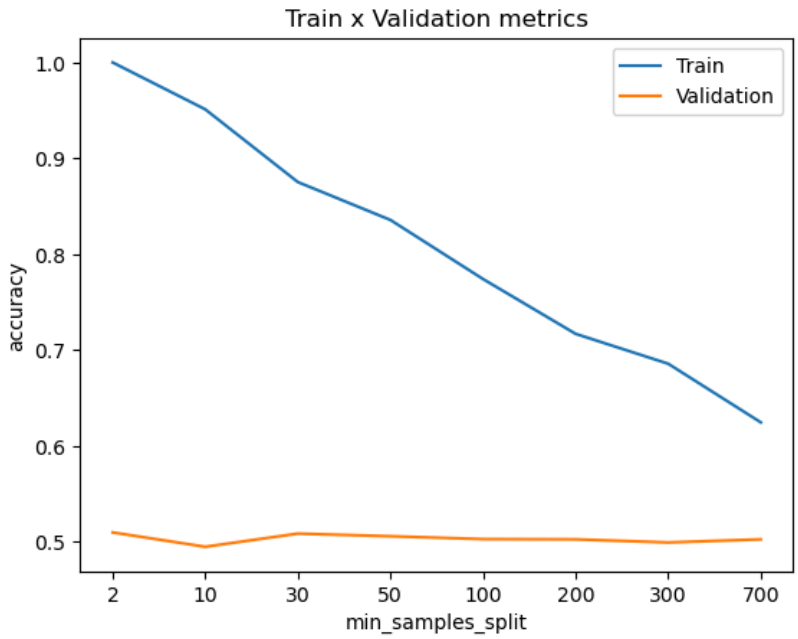
- Max Depth [2, 4, 8, 16, 32, 64]
- Min Samples split [2,10, 30, 50, 100, 200, 300, 700]
- N estimators [100,200,400,500]

After comparison between each values and plotting the results of these values as following :

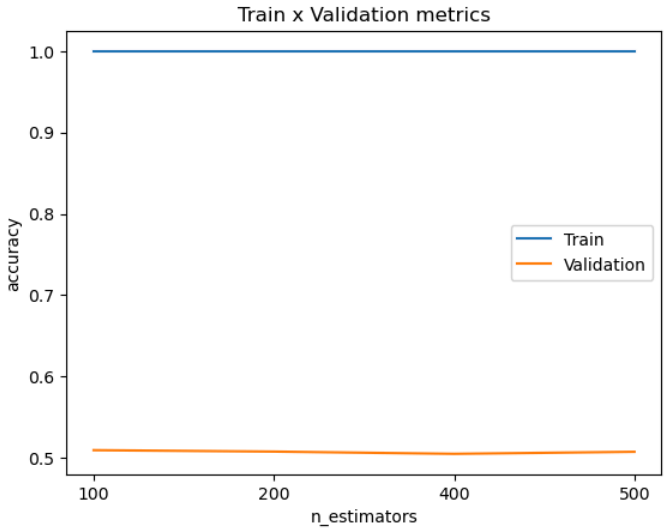
1. Max Depth values' results plotting shown that best value in validation is (2)



2. Min Sample Split values' result shown that best value in validation = (2).



3. N_estimators values' results shown that the best value in validation = 100



So I fit the model using the following parameters values:

- max_depth: 2
- min_samples_split: 2
- n_estimators: 100

5. Evaluation

- With this model and its parameters , I got accuracy = 51.35%
- According to this accuracy , I used **KFolds** method to evaluate the model and got higher accuracy but the accuracy increased for 0.05 % ONLY (Got accuracy = 51.4%)

6. Conclusion

According to K-fold method there is an **underfitting** , this is also shown in the graphs that applying different parameters in random-forest such as (min_samples_split, n_estimators, ..). Those graphs show that the accuracy of the training set is higher than the validation set which tells us that there is an **underfitting** . Future work we can use less number of features to avoid underfitting