Machine Learning ***Assignment # 2***

MSDSF22M001 ***Maida Nadeem Butt***

1. Summary of training at least three linear regression models or any three models of your choice, all use the same training and test splits, or the same cross-validation method.

As the dataset used here is categorical, firstly, we need to transform it into numeric variables using "***LabelEncoder***" of "***sklearn***". Then we get to know about the class imbalance by plotting the target variable in "***boxplot***" i.e. We have many values of a class and few values of others; we balance them out by Oversampling using "RandomOverSampler" from "imblearn". Then applied the following machine learning models:

i. **Naïve Bayes:** Gaussian model to test our normal data.
ii. **Logistic Regression: S**tatistical method
iii. **XGBoost:** Parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

2. A paragraph explaining which of your models you recommend as a final model that best fits your needs in terms of accuracy and explain ability.

When running the models, we had some good models like *Naïve Bayes* and *Logistic Regression* with accuracies of ***75.34%*** and ***83.56%*** some models exceptional, the best model was the "***XGBoost***" with ***93.15%*** accuracy, the model was able to predict both results with the following ***F1-Scores*** for True and False classes respectively ***94***% and ***93***% and gave the lowest ***False Positive*** of ***1*** only and ***False Negative*** of ***4*** values .

3. Summary Key Findings and Insights, which walks your reader through the main drivers of your model and insights from your data derived from your model.

- Looking at the correlation we can see that there is no strong correlation between our data, when looking at our Exploratory Analysis we can see that we have no outliers, when we look at our categorical variables we can say most of our data are from men, usually from Central, and more focused on Commerce and Science, most of our data has no experience and when we look at our target variable, most of the data is from Placed, When we look at our continuous variables, most tend towards the mean and others are well distributed across all values.
- When we compare our categorical variables with our Target variable, we can see that the Not Placed result is usually for women, and that generally those with previous experience

manage to be relocated, looking at our continuous variables, we can see that those employees who have a higher grade are more likely to be placed.

- When we take the "***ssc_percentage***" variable to analyze, we can see that employees who are not from the central and who are linked to science are more likely to have a higher grade, when compared to our continuous variables, practically in most cases, those who have a high grade are always ***Placed***, but when compared to our "***Work_Experience***" variable, this employee does not necessarily need to have had previous experience, just having good grades increases your chance a lot.
- Now talking about the most important variable for the machine learning models to reach the result, it was the variable "***ssc_percentage***", followed by "***hsc_percentage***," something that we could verify in our data analysis, which confirms our suspicions made earlier that the higher the score, more likely to have a positive result.

4. Suggestions for next steps in analyzing this data, which may include revisiting this model adding specific data features to achieve a better explanation or a better prediction.

- ❖ The first thing is to make data balanced manually for better training rather than doing it programmatically.
- ❖ Then we can use any Bagging Algorithms like Random Forest or use Decision Tree model with high power to improve results.
- ❖ Thirdly, more features like institute of higher education are added as it plays an important role in the aptitude improvement and results in a positive result for Placement.