MACHINE LEARNING MAJOR PROJECT SUMMARY

CLASS ID-ML063B17

DATASET

Name: PlacementData_Full_Class.csv

Source: https://www.kaggle.com/benroshan/factors-affecting-campus-placement

- Description: The above data set consists of placement data pertaining to students enrolled in Jain University, Bangalore. It includes secondary and higher secondary grades (in percentages) and the stream that students chose for their higher secondary education. The data set also includes degree specialization, degree type, work experience and salary offers to the students who got placed.

Classification Algorithms Chosen:

- **Logistic Regression(LR)**
- **KNN(K Nearest Neighbors)**
- **SVM**

Questions Chosen:

- To get placed with the highest salary, which degree should be opted?
- Which specialisation has the highest “mba_p”?

EDA and Working of Algorithms:

- First, we converted all the data set columns to a list to get an overview of the data we had on our hands.
- We first checked for null values. All the null values in the dataset were in the salary column and after a quick glance at the data set we found that the null values were only for those students who didn’t get placed. So, we filled all the non-placed students' salaries as 0.
- Lastly to check for columns/features that could be dropped, we label encoded the all the non-numeric columns and after looking at the heat map we decided that “ssc_b”, ”mba_p”, ”degree_t”, ”ssc_p”, ”hsc_p”, ”salary” were more correlated and as serial number was not relevant and thus dropped them all.

Data Visualisation:

- Let’s first see what Data visualization is -:
  Data visualization is the graphic representation of data.It involves producing images that communicate relationships among the represented data to viewers of the images.
- We answered EDA questions using Data Visualization.

Assigning Dependent & Independent Variables:

- Wassigned a set of variables as Independent variables and put them in a list and put the dependent variable in another list.
- Dependent Variable - Status of placement
- Independent Variables - Everything else that remained.
- After this, we set proceeded with our chosen models.

HERE WE GO WITH THE

MACHINE LEARNING MODELS

Logistic Regression:

- For this model, we simply imported the concerned module from the sklearn package and passed our data set to the concerned function from said module.
- Then an accuracy percentage was calculated using accuracy_score from sklearn.metrics .

KNN(K Nearest Neighbors):

- For this model, we imported the KNN model from sklearn library. The KNN model was then fitted with the training data set to carry out KNN classification.

- Then an accuracy percentage was calculated using accuracy_score from sklearn.metrics.

SVM:

- For this model, we again imported the concerned module from the sklearn package and passed our data set to the relevant function from said module.

- Then an accuracy percentage was calculated using accuracy_score from sklearn.metrics.

Comparing Accuracies:

- Logistic Regression: 77.77%
- KNN (K nearest neighbors): 75.92%
- SVM: 68.51%

Conclusions:

- We can clearly see that Logistic Regression wins in terms of accuracy, followed closely by KNN.
- To answer our questions, we got value counts for the required features ("degree_t" & "specialization"). We then plotted histograms which gave us our answers. We concluded from the plot that students who opt for Commerce and Management get placed with the highest salary and "MKT&FIN" has the highest "mba_p".