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)**
- **Random Forest**

- Note: We also tried implementing our own version of a Logistic Regression algorithm, coded entirely from scratch.

Questions Chosen:

- To get placed with the highest salary package, which Degree should one opt for?
- People with which degree and specialization are more likely to be placed?

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 and after rechecking for
- null values and not finding any, we moved on.
- Lastly we checked for any columns/features that could be dropped, and decided that the school board (for both secondary and higher secondary) and serial number were not relevant and thus dropped them.

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 plotted all relevant columns in various types of graphs and plots to get a nice visual representation of our data set.

Assigning Variables and Setting Train/Test Ratio:

- Keeping in mind the questions we chose to answer with this dataset, we assigned 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 the Train/Test to a 0.75/0.25 split and 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.
- The team then generated a confusion matrix and a classification report using more functions from the sklearn package (metrics module). An accuracy percentage was also calculated using similar methods.

KNN(K Nearest Neighbors):

- For this model, we imported the KNN and GridSearchCV models from SKlearn library. Then we created a dictionary called ’parameters' containing a set of hyper parameters for the KNN model. The KNN model was then fitted with the training data set and then the GridSearchCV model was fitted with our trained KNN model and hyper parameters were set. GridSearchCV model chose the best hyper parameters from the set and used them to carry out KNN classification.

- The team again generated a confusion matrix and a classification report using the same functions as before, along with an accuracy percentage.

Random Forest:

- 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.

- The team again generated a confusion matrix and a classification report using the same functions as before, along with an accuracy percentage.

Comparing Accuracies:

- Logistic Regression: 83.33%
- KNN(K nearest neighbors): 78%
- Random Forest: 74%

Conclusions:

- We can clearly see that Logical Regression wins in terms of accuracy, followed closely by KNN.
- To answer our first question, we isolated the required data (placement status and degree chosen) and got value counts for both features. We then used these value counts to make an lmplot which we then used to finally make a violinptot, which gave us our answer. We can conclude from the plot that people who opt for Commerce and Management get placed with the highest salary packages.

Conclusions:

- To answer our second question, we again isolated data features relevant to our question (degree, specialisation and placement status) and got value counts for all three of them. We used these value counts to calculate probabilities (of being placed) for each possible combination of unique values from all 3 columns. Then we simply passed these calculated probabilities through a max function, which gave us our answer - people with a degree in Sci&Tech and specialisation in Mkt&Fin have the highest probability of being placed.