**Project Report**

*Data Science Live Project*

Sakshi Chahande

## Abstract

A binary classification model has been created for predicting the eligibility/ineligibility of interns using the dataset provided. The model generated a 100% accuracy level as well as 100% F1 score. Along with this, different visualization methods have been used to compare and analyse the data.

## Keywords

Areas of interest, Eligibility/Ineligibility of Students, Binary Classification, F1 score.

## Introduction

This project aims at giving interns a hands-on experience for a real-world data science project. The interns learn data pre-processing, generating visualization models, as well as applying classification algorithms in Machine learning.

Cloud Counselage Pvt. Ltd. Provided us with a [dataset](DS_DATESET.csv) which contained the details of the applicants who had applied for this internship training program. The dataset contained the details of 1000 applicants. We were provided with 32 columns of different attributes of data, including their degree, branch, area of interest, college, city, CGPA, languages known and eligibility for internship.

## Project Objective:

Students from different cities from the state of Maharashtra had applied for the Cloud Counselage Internship Program. We have the dataset of consisting information of all the students. Using this data, we want to get more insights and draw out more meaningful conclusions. Interns are expected to build a data visualization model and find the best data segmentation model using the student’s dataset. Following are the tasks interns need to perform-

1. Interns need to pre-process the data for missing values, unknown values, encoding

categorical values.

1. Create a data visualization model to build graphs from the dataset answering the

following questions:

* 1. The number of students applied to different technologies.
  2. The number of students applied for Data Science who knew ‘’Python” and who did not.
  3. The different ways students learned about this program.
  4. Students who are in the fourth year and have a CGPA greater than 8.0.
  5. Students who applied for Digital Marketing with verbal and written communication score greater than 8.
  6. Year-wise and area of study wise classification of students.
  7. City and college wise classification of students.
  8. Plot the relationship between the CGPA and the target variable.
  9. Plot the relationship between the Area of Interest and the target variable.
  10. Plot the relationship between the year of study, major, and the target variable.

1. Identify the best binary classifier to classify data into “eligible/1” and “not eligible/0”.

*Input*

Path of the input dataset file

*Output*

For the data classification, In the output, print accuracy of the best model identified.

For the data visualization model, in output code need to create a pdf file of all the graphs

generated by the model.

*This project needs to be performed using Python programming language.*

## Exploratory Data Analysis

The dataset was implemented using Python programming. Different Python libraries like pandas, numpy, matplotlib, scikitlearn, seaborn and pdfpages were used. Following are some of the different exploratory analysis we have implemented:

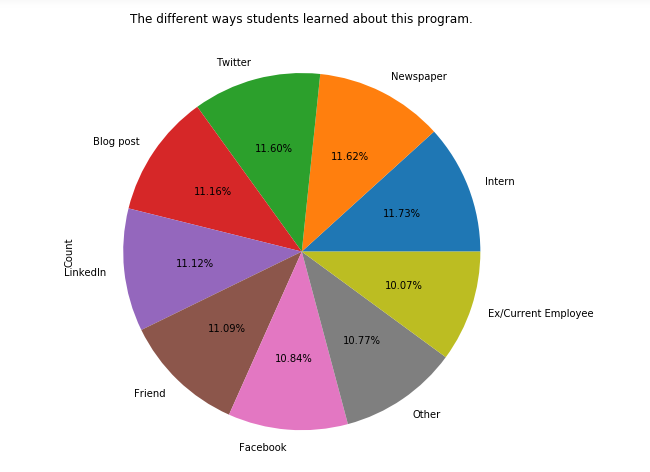


Fig above shows the percentage by which different applicants learnt about this internship program.

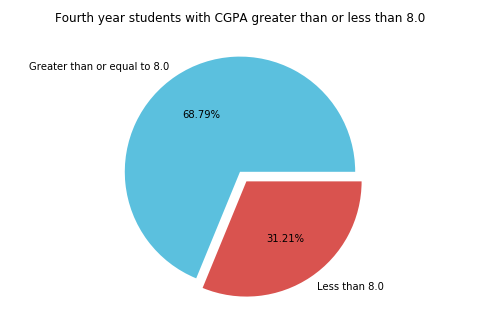
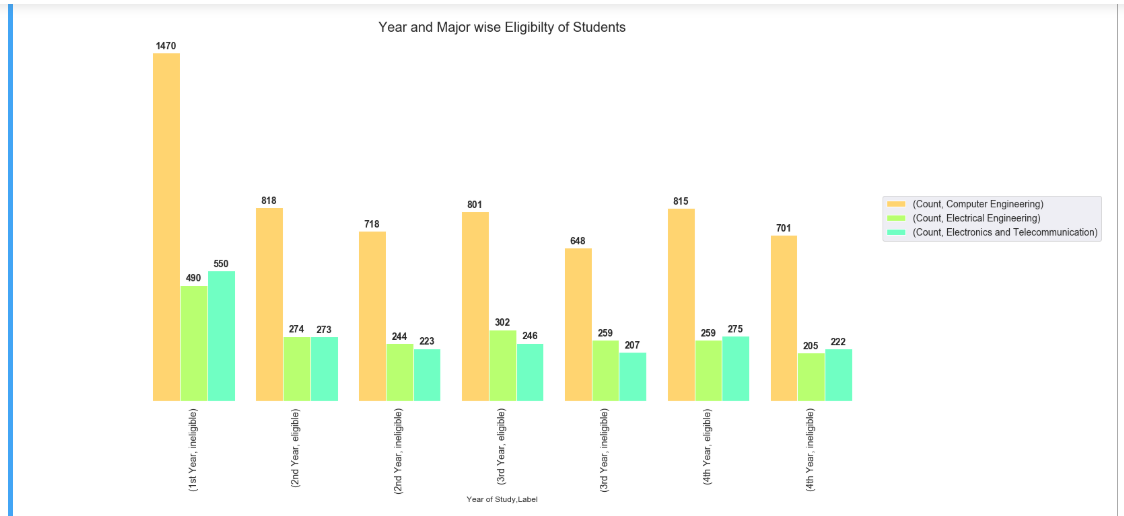


Figure above shows the percentages of 4th year students who have a CGPA of greater than 8.0



The above figure shows the number of students who are eligible and ineligible for the internship, based on their stream.

## Model Selection

Various classification models have been implemented in order to identify the best binary classifier for this project. These classification models include:

1. Support Vector Classifier Model
2. Decision Tree Classification
3. Logistic Regression
4. Random Forest Classifier

After applying these models, we have compared the classification report of these models, specifically the F1 score and accuracy and determined the best option for the same. According to these comparisons, we have concluded that the Random Forest Classifier is the best model to train the dataset for Binary Classification of data.

The F1 score and accuracy for the Random Forest Classifier is 1.0.

## Conclusion

This project analysis different parameters relating to the eligibility and ineligibility of students applying for this internship project. After completion of this project, we have been able to handle the data for missing values and null values, analyse the given data for different exploratory analysis and visualization methods, and applied different models for classifying the same. The best model, according to me, is the Random Forest Classifier.

This project report is submitted along with the following files:

* Project Synopsis
* Software Requirement Specification Document
* Project Plan
* Development Log
* Design Model
* RAID Log
* Visualization models
* Lessons Learned
* Visualization modelling .py file
* Classification .py file