

# Data Science Live Project

15th June- 31st July 2020

---

## OVERVIEW

Cloud Counselage is a Pan India IT, Management, and Career services provider. To expand its reach across all the colleges in Mumbai and surrounding colleges that might be interested to collaborate with the company and develop a data visualisation and data segmentation model for the applications of Internship Program.

## PROBLEM STATEMENT

Students from different cities from the state of Maharashtra had applied for the Internship Program at Cloud Counselage. A dataset consisting of information of all these students is the dataset provided for this project. Interns are expected to build a data visualization model and find the best data segmentation model using the student's dataset.

## GOALS

The main objective of this project is to identify the best classification model that perfectly segments the data with respect to the target variable and develop a data visualisation model to display the results.

## SPECIFICATIONS

Hardware: Laptop/ Desktop with Linux/Windows/Mac OS

Software: Anaconda, Jupyter Notebooks,

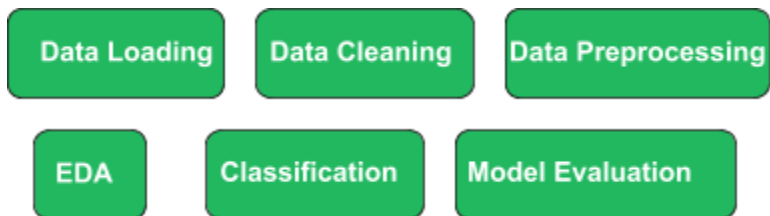
Python Libraries: Pandas, Numpy, Matplotlib, SciKit Learn

## Dataset

First Name	10000 non-null object
Last Name	10000 non-null object
City	10000 non-null object
State	10000 non-null object
Zip Code	10000 non-null int64
DOB [DD/MM/YYYY]	10000 non-null object
Age	10000 non-null int64
Gender	10000 non-null object
Email Address	10000 non-null object
Contact Number	10000 non-null int64
Emergency Contact Number	10000 non-null int64
College name	10000 non-null object
University Name	10000 non-null object
Degree	10000 non-null object
Major/Area of Study	10000 non-null object
Course Type	10000 non-null object
Which-year are you studying in?	10000 non-null object
CGPA/ percentage	10000 non-null float64
Expected Graduation-year	10000 non-null int64
Areas of interest	10000 non-null object
Current Employment Status	10000 non-null object
Have you worked core Java	10000 non-null object
Programming Language Known other than Java (one major)	10000 non-null object
Have you worked on MySQL or Oracle database	10000 non-null object
Have you studied OOP Concepts	10000 non-null object
Certifications/Achievement/ Research papers	0 non-null float64
Rate your written communication skills [1-10]	10000 non-null int64
Rate your verbal communication skills [1-10]	10000 non-null int64
Link to updated Resume (Google/ One Drive link preferred)	0 non-null float64
Link to LinkedIn profile	0 non-null float64
How Did You Hear About This Internship?	10000 non-null object
Label	10000 non-null object

---

## System Design



## Output

- F1 Score of data classification model
- Pdf of Data Visualisation Plots

## Comparison of various classifiers

Classification Model	F1 Score
Logistic Regression	0.71
Decision tree	0.86
SVM	0.82
Random Forest Classifier	1.0

## CONCLUSION

After analysing various classification models, the random forest classification algorithm was found to give the highest F1 score among all the others. Hence in this project the final model implemented was the random forest classifier and data visualisation model was built to plot the graphs. All the plots are combined into a single PDF file using PdfPages. This was the project assigned to me under my internship tenure at Cloud Counselage.