# DATA SCIENCE LIVE PROJECT REPORT

**Shreya Pawaskar**

**Data Science Intern**

**PROBLEM STATEMENT:**

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 preprocess the data for missing values, unknown values, encoding categorical values.

2. 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 didn’t.
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.

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

***INPUT:***

DATASET PROVIDED

***OUTPUT:***

1. MACHINE LEARNING MODEL WITH F1 SCORE
2. VISUALIZATION PDF OF THE VARIOUS ATTRIBUTES

**INTRODUCTION:**

This Live Project is a part of Internship Program of Cloud Counselage Pvt Ltd in the field of Data Science. We have to make a binary classifier for the dataset to classify data. Also we need to perform data visualization for knowing the data better.Data analysis will help us in the process of cleaning, transforming, and modelling data to discover useful information for business decision-making.

**METHODOLOGY:**

1. Finalize design
2. ***Data Cleaning***: It is the process of detecting and correcting (or removing) corrupt or inaccurate records .We will drop all the unwanted columns from the dataset which have Nan values, redundant columns like DOB as age is already given and so on.
3. ***Pre-processing***: It is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency. We will normalize the data in such a way that is usable for modelling.
4. ***Data Visualization:*** Data visualization is the graphical representation of information and **data**. By using visual elements like charts, graphs, and maps. Using Matplotlib we will generate graphs which will be saved in PDF by using Pdf Pages.
5. **Model for Classifier:** A **machine learning model** is the output of the training process and is defined as the mathematical representation of the real-world process. Try different models on the dataset to find the best model.
6. **Accuracy Testing:**  **F1 score** conveys the balance between the precision and the recall. Check the F1 score of the models to choose the best Model.

***MODELLING [CLASSIFIERS USED]*:**

1. Bayes Classifier:
2. K neighbours classifier
3. Random Forest

***RESULTS:***

|  |  |  |
| --- | --- | --- |
| *MODEL* | *Accuracy* | *F1score* |
| Bayes Classifier | **0.8635** | **0.8953** |
| K neighbours classifier | **0.993** | **0.9946** |
| Random Forest | **1.0** | **1.0** |

***CONCLUSION:***

1. The most important input variables are CGPA, Written communication skills and Verbal communication skills.
2. Most of the eligible students are from the second, third and last year.
3. Most of the students applied for the Internship are from the Computer Engineering background.
4. Most of the students enrolled knew Python.
5. Around 70 per cent fourth year students enrolled had CGPA greater than 8.
6. Random forest is one of the most accurate and versatile learning algorithms available. It runs efficiently on large databases. It handles thousands of input variables without variable deletion. Hence Random Forest is the best classifier model.