**DATA SCIENCE LIVE PROJECT REPORT**

**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:**

STUDENTS DATASET

**OUTPUT:**

1. ML MODEL
2. DATA VISUALIZATION PDF

**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.

**METHODOLOGY:**

1. Data Understanding
2. Data Quality validation
3. Data Cleaning: Drop all the unwanted columns from the dataset which have Nan values.
4. Pre-processing: We will normalize the data in such a way that is usable for modelling.
5. Data Visualization: Using matplotlib, seaborn, squarify (for Treemap) we will generate and plot the figures which will be saved as pdf file by using PdfPages.
6. Model for Classifier:Try different models on the dataset and perform the evaluation by F1 score as a metric to compare the scores.
7. Find best model: Check the highest F1 score achieved to choose the best Model.

**MODELLING:**

1. Gaussian Naive Bayes Classifier
2. Logistic Regression
3. SVM
4. Neural Networks MLPClassifier
5. K neighbours classifier
6. SVM with poly kernel and degree 8
7. Decision Tree
8. Random Forest

**RESULTS:**

|  |  |
| --- | --- |
| **MODEL** | **F1score** |
| Gaussian Naive Bayes Classifier | **0.88** |
| Logistic Regression | **0.75** |
| SVM | **0.9** |
| Neural Networks MLPClassifier | **0.75** |
| K neighbours classifier | **0.8** |
| SVM with Poly kernel and degree 8 | **0.75** |
| Decision Tree | **1.0** |
| Random Forest | **1.0** |

**CONCLUSION:**

Hence it can be concluded as Random Forest Classifier and Decision Tree are the best classifiers to perform binary classification on Students Dataset.