**DATA SCIENCE 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:**

DATASET

**OUTPUT:**

1. ML MODEL
2. 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. Finalize design
2. Data Cleaning: 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: We will normalize the data in such a way that is usable for modelling.
4. Data Visualization: Using matplotlib we will generate graphs which will be saved in pdf by using PdfPages.
5. Model for Classifier:Try different models on the dataset to find the best model.
6. Accuracy Testing: Check the fscore of the models to choose the best Model.

**MODELLING:**

1. Logistic Regression
2. Naïve Bayes Classifier:
3. K neighbours classifier
4. SVM
5. Random Forest

**RESULTS:**

|  |  |  |
| --- | --- | --- |
| **MODEL** | **Accuracy** | **F1score** |
| Logistic Regression | **0.77** | **0.71** |
| Stochastic Gradient Descent | **0.78** | **0.72** |
| Naïve Bayes Classifier | **0.8695** | **0.82** |
| K neighbours classifier | **0.84** | **0.81** |
| SVM | **0.971** | **0.964** |
| Random Forest | **1.0** | **1.0** |

**CONCLUSION:**

In this study, I analysed the candidates scores and their skills based on which their eligibility was depended for the Internship. I identified CGPA, Written communication skills and Verbal communication skills are the most important features which affect the selection of the candidate. I built classification models to predict which candidates are eligible/ineligible. These models can be very useful in helping other companies in number of ways. For example, it could help identify the candidates who are more deserving and eligible for the posts available in the company.

Some of the analysis I observed as follows:

1. Most of the students with CGPA above 7.5 are eligible for the Internship.
2. Most of the eligible students are from the second, third and last year.
3. Highest selection percentage is for Blockchain and the lowest is for the IoT.
4. Most of the students applied for the Internship are from the Computer Engineering background.
5. Using Random forest, every time we were able to get almost correct predictions.