**JOB RECOMMENDATION, STUDENT GRADE, GRADUATE ADMISSION PREDICTION.**

**AIDI 1003: CAPSTONE-1**

**MODULE ONE: SoW**

**FACULTY:**

**Dr. Uzair Ahmad**

**STUDENT:**

**SHRITA GAONKAR 100799307**

**EXECUTIVE SUMMARY**

Most students after graduation are open to innumerable job roles. Some of the students are good at analysis while some are good at development. Being open to a bunch of job roles might be confusing as it is difficult to analyze the best fit for yourself and match the interest. The goal of my model would be to find the best job role for students.

On the other hand, my model can also predict the chances of getting an admit from their desired university.

Three major project goal

1. Identify the best job role for students
2. Find out whether the student is eligible to apply at a particular university.

**RATIONALE STATEMENT**

Identifying the best job role based on analytical and spatial reasoning. Predicting whether a student gets an admit from their desired learning institution based on their expectation and university requirements.

**DATA REQUIREMENTS**

* The first data requirement will be to find out the key variables which define the quality of the data.
* For developing various types of models, the data should be large and diverse.
* The data should be in .csv or in excel format for better extraction.
* The data size must be feasible according to the available machine.
* The dataset is required to not have special characters in them so the data can be cleaned easily.

**ASSUMPTIONS**

* I am keeping in mind that there was no data manipulation before.
* I am assuming that we have enough data features available to build the machine learning models.
* I have enough data to conduct analysis.
* I am assuming that all columns are necessary, and all independent variables are correlated with the dependent variable and not changing any variable name.
* The dataset would be easily understandable to all the developers.

**CONSTRAINTS**

* One of the biggest constrain is that all the independent variables are not properly defined.
* The data is imbalanced because all the dependent variables have different total count.
* The variables that are most useful are yet to be defined.
* Few variables have outliers and we do not know that they are valuable for the model or not.
* All independent variables might not have correlated with the dependent variable.
* The dataset requires to have at least one dependent variable for conduct analysis.
* The data format should be .csv or in excel to build the model in Jupyter.

**DATA**

There are three datasets that I will be using which are downloaded from Kaggle

**Data\_job\_posts.csv**

* There are 24 columns of dataset.
* There are 6 columns which have maximum null values.
* Title is the dependent variable

**Admission\_predict.csv**

* The dataset contains GRE, LOR, SOP, IELTS, CGPA, etc.
* These are the independent variables
* The column ‘chances of admit’ is the dependent variables.

**MODELS**

* K-means
* Naive Bayes
* Logistic regression
* Decision Tree Classifier
* Gradient boosting

**PROJECT PLAN**

|  |  |  |
| --- | --- | --- |
| **TASK** | **HOURS** | **DELIVERY DATE** |
| Collect information | 15 | 30-SEP-2020 |
| Analyze dataset | 20 | 15-OCT-2020 |
| Clean dataset | 10 | 21-OCT-2020 |
| Develop data processing pipeline | 15 |  |
| Train all models | 30 | 15-NOV-2020 |
| Building all models | 25 | 20-NOV-2020 |
| Test models | 20 | 25-NOV-2020 |
| Evaluate models | 15 | 02-DEC-2020 |
| Feature selection | 15 | 02-DEC-2020 |
| Refine model | 10 | 20-DEC-2020 |
| Develop report | 8 | 23-DEC-2020 |
| Develop score card | 10 | 23-DEC-2020 |