# LOW LEVEL DOCUMENT

**Campus Placement Prediction**

| | |
|---|---|
| Written by | **Sourav Dey** |
| Document Version | **1.0** |
| Last Revised Data | **20th December 2022** |

**1 Document Version Control:**

| Version | Date | Author | Comments |
|---|---|---|---|
| **1.0** | **20 December 2022** | **Sourav Dey** | |
| | | | |

# Contents

# 1.Introduction

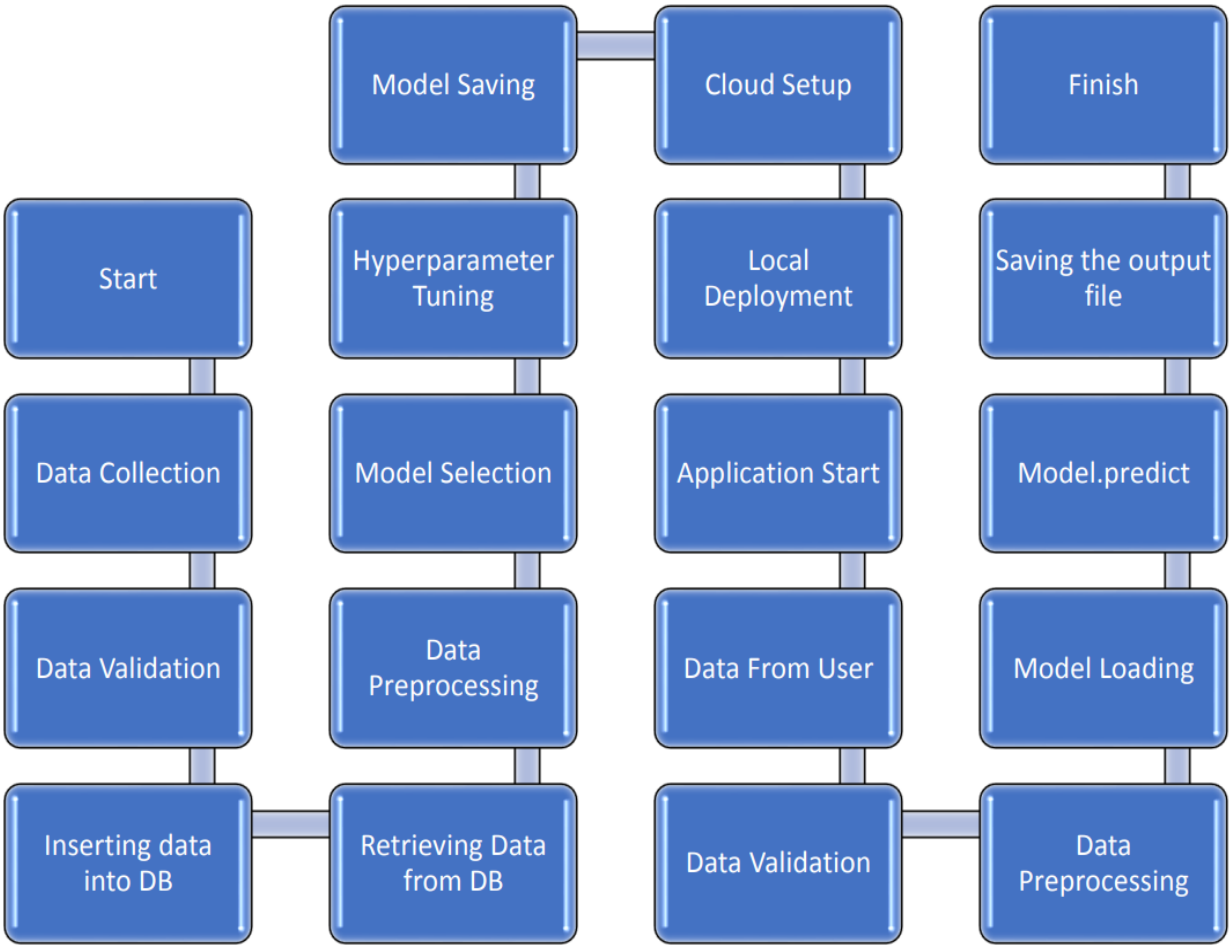
## 1.1 What is Low-Level design document?

The goal of LLD or a low-level design document (LLDD) is to give the internal logical design of the actual program code for flight fare estimation System. LLD describes the class diagrams with the methods and relations between classes and program specs. It describes the modules so that the programmer can directly code the program from the document.

## 1.2. Scope

Low-level design (LLD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work.

## 2. Architecture



# 3. Architecture Description

## 3.1 Data Description

**File descriptions**

- train.csv - the training set
- test.csv - the test set
- SampleSubmission.csv - a sample submission file in the correct format

**Data fields**

- gender - sex of the student
- secondary education percentage-marks obtained in secondary education
- higher secondary percentage-marks obtained in higher secondary education
- degree percentage-marks obtained in degree
- Under-graduation(Degree-type)-Field of degree education
- Work-experience
- Employability-test-package
- specialisation-field of study

## 3.2 Data Transformation

In our dataset a lot of categorical values are present, we transform those attributes into numerical values using one hot encoding. A one hot encoding is a representation of categorical variables as binary vectors.

This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1.

## 3.3 Data Pre-processing

Data preprocessing is a technique that is used to convert raw data into a clean dataset. The data is gathered from different sources is in raw format which is not feasible for the analysis. Pre-processing for this approach takes 4 simple yet effective steps.

Attribute selection some of the attributes in the initial dataset that was not pertinent (relevant) to the experiment goal were ignored. The attributes name, roll no, credits, backlogs, whether placed or not, b.tech %, gender are not used. The main attributes used for this study are credit, back-logs, whether placed or not, b.tech %.

Cleaning missing values in some cases the dataset contain missing values. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you're inadvertently removing crucial information? After all we might not need to try to do that. One in every of the foremost common plan to handle the matter is to require a mean of all the values of the same column and have it to replace the missing data. The library used for the task is called Scikit Learn preprocessing. It contains a class called Imputer which will help us take care of the missing data. to split our dataset into two. Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to examine how accurately it will predict. A general rule of the thumb is to assign 80% of the dataset to Training and Test data splitting the Dataset into Training set and Test Set Now the next step is training set and therefore the remaining 20% to test set.

Feature scaling the final step of data preprocessing is feature scaling. But what is it? It is a method used to standardize the range of independent variables or features of data. But why is it necessary? A lot of machine learning models are based on Euclidean distance. If, for example, the values in one column (x) is much higher than the value in another column (y), (x2-x1) squared will give a far greater value than (y2-y1) squared. So clearly, one square distinction dominates over the other square distinction.

## 3.4 Data Clustering

The campus placement activity is incredibly vital from institution point of view as well as student point of view. In this regard to improve the student's performance, a dataset has been analyzed and predicted using the classification algorithms like KNN algorithm, Random Forest, Naïve Bayes, Logistic Regression, Decision Tree and the SVM algorithm to validate the approaches.

## Models used:

**KNN ALGORITHM**- K Nearest Neighbor (KNN) is intuitive to understand and an easy to implement the algorithm. It is a versatile algorithm also used for imputing missing values and resampling datasets.

**Random Forest**- Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

**Naïve Bayes**- Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes Theorem, used in a wide variety of classification tasks. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

**Logistic Regression**- Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

**Decision Trees**- Decision trees use multiple algorithms to decide to split a node into two or more subnodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that the purity of the node increases with respect to the target variable. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

**SVM Algorithm**- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. o The goal of the SVM algorithm is to create the best line or decision boundary that can segregate ndimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

## 3.5 Model Building

After clusters are created, we will find the best model for each cluster. For each cluster, algorithms will be passed with the best parameters. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for KNN algorithm is 74%, Random Forest is 82%, Naïve Bayes is 84%, Logistic Regression 87%, Decision tree is 84% and for the SVM is 89.7%. Hence, from the above said analysis and prediction it's better if the SVM algorithm is used to predict the placement results.

## 3.6 Data from User

- gender - sex of the student
- secondary education percentage-marks obtained in secondary education
- higher secondary percentage-marks obtained in higher secondary education
- degree percentage-marks obtained in degree
- Under-graduation(Degree-type)-Field of degree education
- Work-experience
- Employability-test-package
- Specialisation-field of study is collected.

### 3.7 Data Validation

Here Data Validation will be done, given by the user

### 3.8 Data Clustering

The model created during training will be loaded, and clusters for the user data will be predicted.

### 3.9 Model Call for Specific Cluster

Based on the cluster number, the respective model will be loaded and will be used to predict/Recommend the data for that cluster.

### 3.10 Deployment

We will be deploying the model to HEROKU.

# 10. Test Cases

Test cases are given below

| Test Case Description | Pre-Requisite | Expected Result |
|---|---|---|
| Verify whether the Application URL is accessible to the user | 1. Application URL should be defined | Application URL should be accessible to the user |
| Verify whether the Application loads completely for the user when the URL is accessed | 1. Application URL is accessible<br>2. Application is deployed | The Application should load completely for the user when the URL is accessed |
| Verify Response time of url from backend model. | 1. Application is accessible | The latency and accessibility of application is very faster we got in Heroku service. |
| Verify whether user is giving standard input. | 1. Handeled test cases at backends. | User should be able to see successfully valid results. |
| Verify whether user is able to edit all input fields | 1. Application is accessible | User should be able to edit all input fields |
| Verify whether user gets Custom File Predict, Default File Predict button to submit the inputs | 1. Application is accessible | User should get both buttons to submit the inputs |
| Verify whether user is presented with recommended results on clicking submit | 1. Application is accessible | User should be presented with recommended results on clicking submit |
| Verify whether the recommended results are in accordance to the selections user made | 1. Application is accessible | The recommended results should be in accordance to the selections user made |