K. K. Wagh Institute of Engineering Education and Research, Nashik. Department of Computer Engineering
Academic Year 2022-23

Course: Laboratory Practice III	**Course Code: 410246**
	Name:Atharva Kalidas
Pandharikar	**Class: BE**
Roll No.: 07	**Div: B**

ML Mini Project

Problem Statement:

Build a machine Learning model of Placement prediction. Placement Prediction data (sl_no, gender, ssc_p, scc_b, hsc_p, hsc_b, hsc_s, degree_p, degree_t, workex,etest_p, specialision,mba_p,status,salary, etc).
Dataset Link : (https://www.kaggle.com/datasets/barkhaverma/placement-data-full-class)

Process of a ML Project Development

The development of ML projects involve systematic procedure and steps. Basic steps that should be followed is as follows:

- Understand the given problem. Whatever the problem is it need to understand first.
- Frame the problem statement and look at the big picture.
- Collect the data.
- Prepare the data to better expose the underlying data patterns and increase the performance of machine learning algorithms.
- Explore many different models and select the best ones.
- Train the model.
- Test and evaluate the model.
- Deploy the model.

DATASET PREPARATION

1. **Dataset Description :**
   The data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients. The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Sr.No	Attributes
1	**sl_no**
2	**gender**
3	**ssc_p**
4	**ssc_b**
5	**hsc_p**
6	**hsc_b**
7	**hsc_s**
8	**degree_t**
9	**degree_p**
10	**etest_p**
11	**specialisation**
12	**mba_p**
13	**status**
14	**salary**

0. Data Preprocessing

Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.

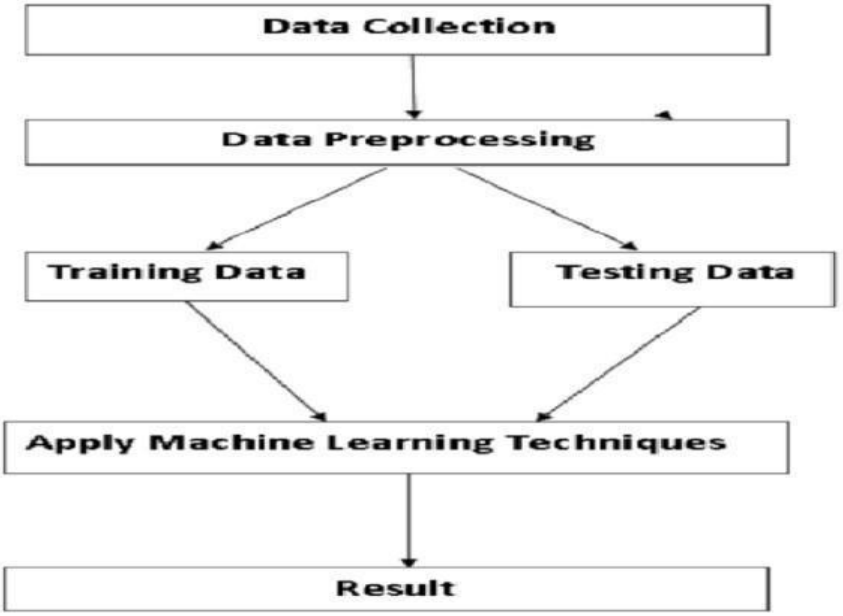
1). Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.
2). Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training

process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

0. Apply Machine Learning

When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction.
The Techniques are follows-
1)Logistic Regression: Logistic regression is a classification technique and it is very good for binary classification. It's decision boundary which is generally linear derived based on probability interpretation. The results are in a nonlinear optimization problem for parameter estimation. Parameters can be estimated by maximising the expression using any nonlinear optimization solver. The goal of this technique is given a new data point, and predict the class from which the data point is likely to have originated. Input features can be quantitative or qualitative. Instead of a hyperplane or straight line, the logistic regression uses the logistic function to obtain the output of a linear equation between 0 and 1. The function is defined as logistic(x)=1/(1+exp(-x)

2) **Random Forest**: We have a plethora of classification algorithms at our disposal, including, but not limited to, SVM, Logistic regression, decision trees and Naive Bayes classifier, just to name a few. But, in the hierarchy of classifiers, the Random Forest Classifier sits near the top. The random forest classifier is a group of individual decision trees and so, we shall look into how decision trees work. It is basically a flowchart-like structure in which each node excluding the leaf node is a test on a feature (i.e, what will be the outcome if some activity, such as flipping a coin, is done), leaf nodes are used to represent the class label (the decision taken after all features are computed) and branches represent the conjunctions of features that lead to those class labels. The classification rules of a decision tree are the paths from the root node to the leaf node. So then, now let us look into random forest classifiers. As mentioned earlier, it is a collection of decision trees. The basic idea behind random forest is “the wisdom of the crowds”. It is a powerful concept wherein a large number of uncorrelated models, or in this case trees, operating as a group, would provide a much more solid output than any of the constituent models. So, in a random forest, each individual tree with different properties and classification rules would try to find an appropriate class label for the problem. Each tree would give out its own answer. A voting is done within the random forest to see which class label received the most votes. The class label with the most votes would be considered the final class label for the problem. This provides a more accurate model for class label prediction.



MODEL BUILDING

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. KNearest Neighbor**,** Support Vector Machine, Decision Tree, Logistic regression, Random Forest and Gradient boosting algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analyzing based on various measures conclude the best performing algorithm.

Prediction of Campus Placement

Visualizing the data

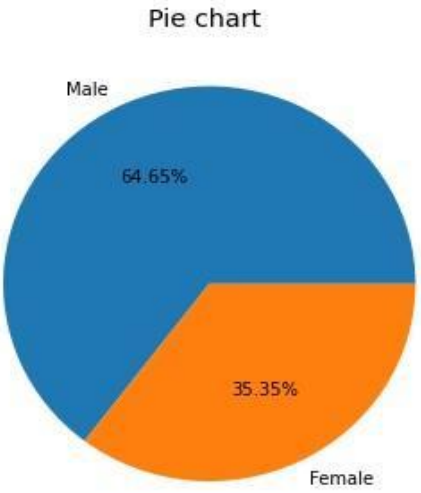


Fig. No of Male and female

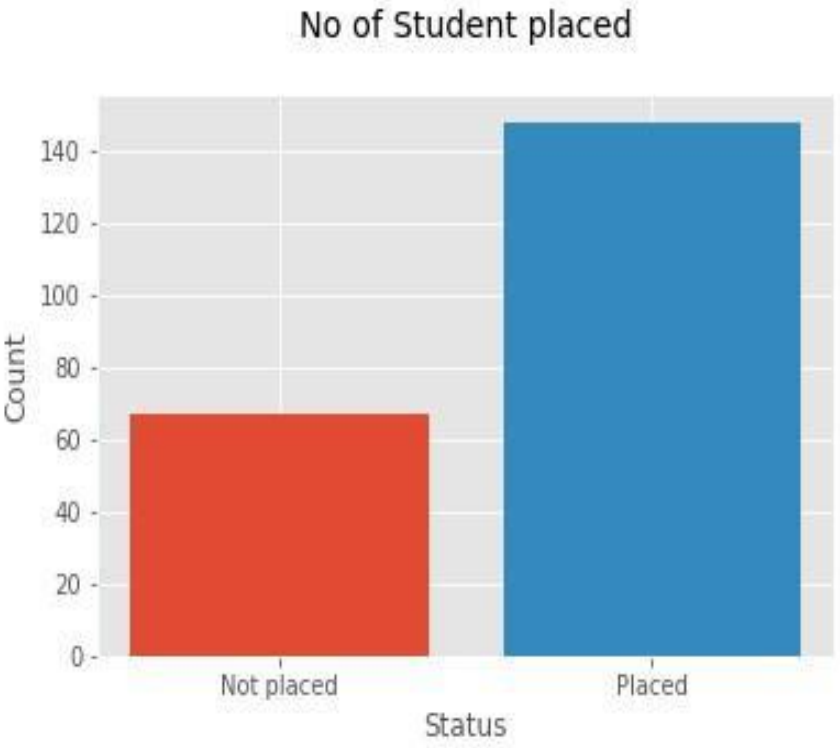


Fig. No of Placed Student

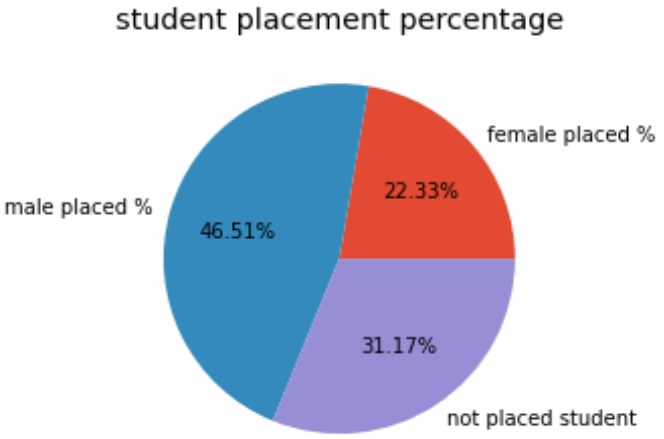


Fig. Student Placement Percentage

Conclusion:

Placement prediction system is a system which predicts the placement status of final year B-Tech students. For data analysis and prediction different machine learning algorithms are used in the python environment. We analyse the accuracy of different algorithms and it is shown in the above table. Logistic Regression is also good which gives an accuracy of 97.59 based on the given dataset. The accuracy of Machine learning algorithms may differ according to the dataset. Random Forest is good for binary classification problems since they all give accuracy of above 95.