

Campus Placements Prediction & Analysis using Machine Learning

Priyanka Shahane
Department of Artificial Intelligence & Data Science,
AISSMS Institute of Information Technology,
Pune, Maharashtra, India
priyankashahane04@gmail.com

Abstract — Campus placement is an activity of participating, identifying and hiring young talent for internships and entry level positions.

Reputation and yearly admissions of the institute invariably depend upon the placements provided by the institute to the students. Therefore, most of the institutions, assiduously, try to boost their placement department in order to improve their organization on a full scale. Any assistance during this specific space can have a good impact on the institute's capability to position it's students.

In this study, the target is to analyze student's placement data of last year and use it to determine the probability of campus placement of the present students. For this we have experimented with four different machine learning algorithms i.e. Logistic Regression, Decision Tree, K Nearest Neighbours and Random Forest.

Index Terms — Machine Learning, Campus placements prediction, Logistic Regression, Decision Tree, KNN, Random Forest

I. INTRODUCTION

NOWADAYS the number of educational institutes is growing day by day. The aim of each higher educational institute is to help their students to get a well-paid job through their placement cell. One of the biggest challenges that higher learning institutes face these days is to uplift the placement performance of scholars.

The goal of this system is to predict whether the student will get a campus placement or not based on various parameters such as gender, SSC percentage, HSC percentage, HSC stream, degree percentage, degree type, work experience & e-test percentage.

This research focuses on various algorithms of machine learning such as Logistic Regression, Decision Tree, K-Nearest Neighbours and Random Forest in order to produce economical and correct results for campus placement prediction. This system follows a supervised machine learning approach as it uses class labelled data for training the classification algorithm.

II. LITERATURE SURVEY

Sharma et. al. developed the placement predictor system i.e. PPS by using a model of logistic regression. For this he has

considered the features such as matriculation score, senior secondary score, scores of the subjects in various semesters & demographics. Here, dataset used is of GuruNanak Dev Engineering College (GNDEC), Ludhiana. This model gave an accuracy of around 83.33%.

Elayidom et. al. constructed multi way decision trees using various parameters such as branch, sector, sex & rank. Here, the dataset used is received from the National Technical Manpower Information System (NTMIS) via the Nodal center. This model gave an accuracy of 80%.

Nagaria et. al. used the Random Forest model where he has considered various parameters such as degree type, work experience, e test percentage, specialization, MBA percentage. The dataset used is taken from Kaggle. This model gave the highest accuracy of 85 %.

S.Venkatachalam et. al. designed the fuzzy inference system using Naive Bayes algorithm for campus placement prediction. The dataset is prepared with the help of primary & secondary data collection sources. This model gave the highest accuracy of 86.15%.

Manvitha et. al. designed used the Random Forest model where she has considered various parameters such as credit , backlogs , whether placed or not, b.tech %. The dataset is collected from the placement department of Sreenidhi Institute of Science and Technology. This model gave the highest accuracy of 86%.

III. METHODOLOGY

The steps involved in this system are as follows,

A. Data Acquisition:

The campus placement dataset is collected from Kaggle website. Here is the link for the dataset: https://www.kaggle.com/benroshan/factors-affecting-campus-placement?select=Placement_Data_Full_Class.csv

The dataset consists of various attributes such as Serial Number, Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E -test Percentage, Degree Specialization, Degree Percentage, Placement Status & Salary. The size of dataset is 19.71 KB & it has total 215 records.

1) Handling missing values:

In our dataset missing values are present only in the salary column as these values correspond to the students who didn't get placed in any placement drive. So it is assumed that the missing values in Salary Column are Zero & replaced them by zero using `fillna(0,inplace=True)` function in Python.

2) Handling categorical data:

Since we cannot deal with categorical values directly, mapping is done for attributes having categorical values.

Gender attribute has values M (Male) & Female (F). Here, M is replaced by 0 & F is replaced by 1. SSC & HSC Board attributes has values 'Central' & 'Other.' Here, Central is replaced by 1 & Other is replaced by 0. Work Experience attribute has values 'Yes' & 'No'. Here, 'Yes' is replaced by 1 and 'No' is replaced by 0. Degree specialization attribute has values 'Marketing & Finance' & 'Marketing & HR'. Here, 'Marketing & Finance' is replaced by 1 and 'Marketing & HR' is replaced by 0. Status attribute has values 'Placed' and 'Not Placed'. Here, 'Placed' is replaced by 1 and 'Not Placed' is replaced by 0. This is achieved through map function in Python.

For e.g.,

- `df['gender']=df['gender'].map({'M':0,'F':1})`
- `df['ssc_b']=df['ssc_b'].map({'Central':1,'Others':0})`
- `df['workex']=df['workex'].map({'Yes':1,'No':0})`

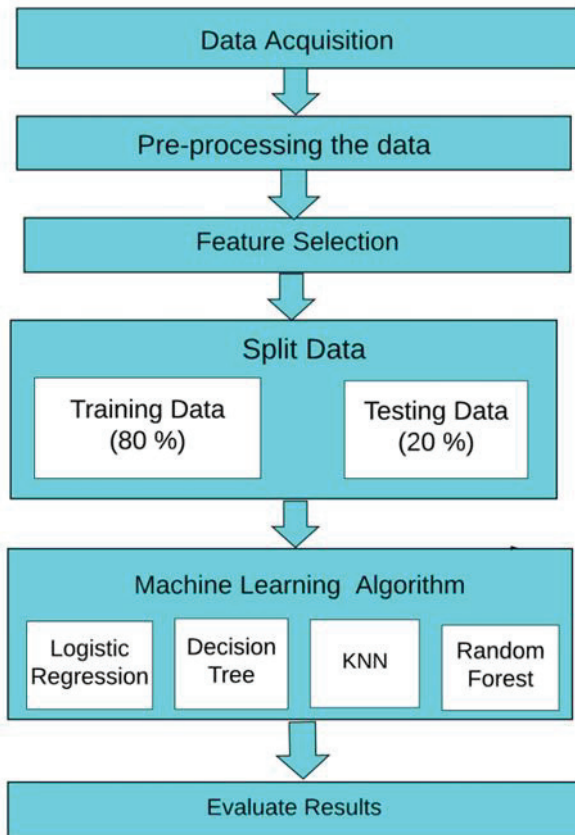


Fig. 1. Architecture Diagram

3) Feature Selection:

Here, various features are visualized to understand their correlation with the target feature.

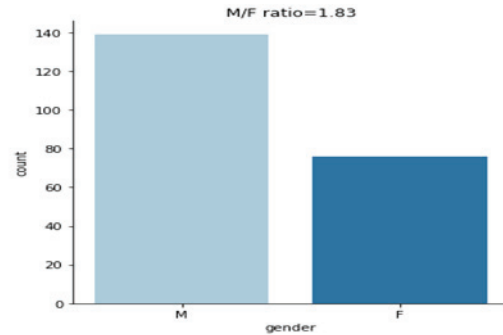


Fig. 2. M/F ratio

Here, male : female ratio for one batch of students is approximately equal to 2. It means that there are 2 male candidates appearing for placement drives for every 1 female candidate.

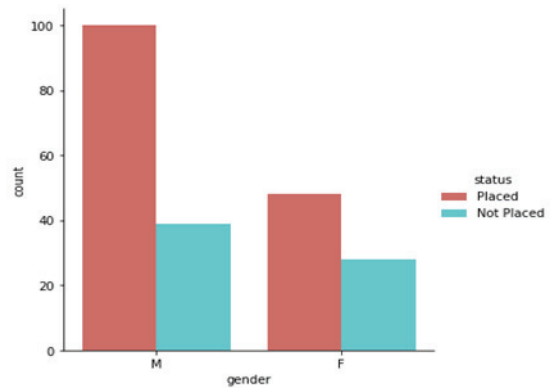


Fig. 3. Placement count vs. gender

From the above graph it can be concluded that the count of placed male candidates in a batch is higher as compared to female candidates & the placement count is dependent on gender.

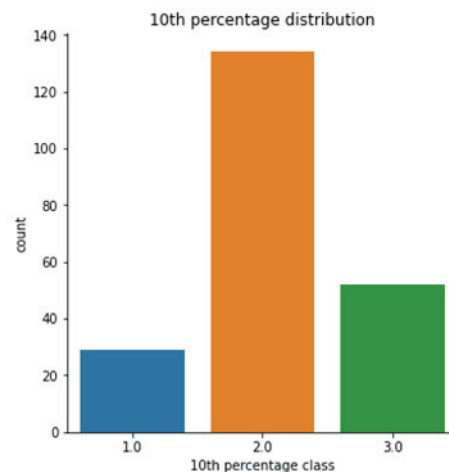


Fig. 4. 10th standard percentage distribution

In the above graph, class 1 represents students having scores between 80-100%, class 2 represents students having scores between 60-80% and class 3 represents students having less than 60 % score in 10th standard.

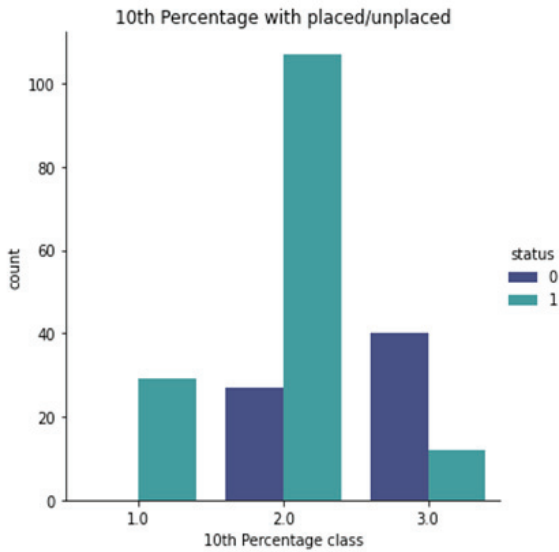


Fig. 5. Placement count vs. 10th percentage

From the above graph, it's observed that all the students having scores between 80-100% in 10th standard got placed. Very few students having scores between 60-80% in 10th standard couldn't get placed. Whereas, most of the students having below 60% score in 10th standard couldn't get placed.

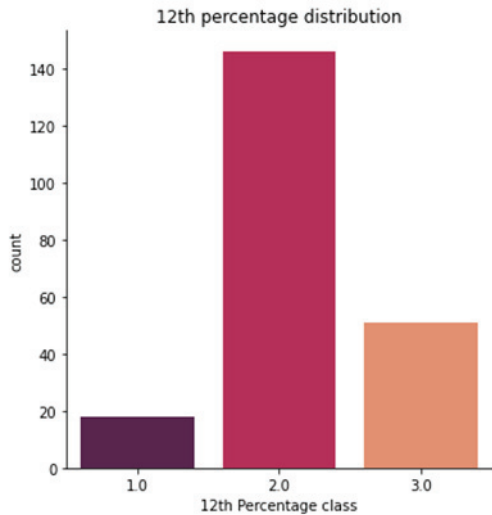


Fig. 6. 12th standard percentage distribution

In the above graph, class 1 represents students having scores between 80-100% , class 2 represents students having scores between 60-80% and class 3 represents students having less than 60 % score in 12th standard.

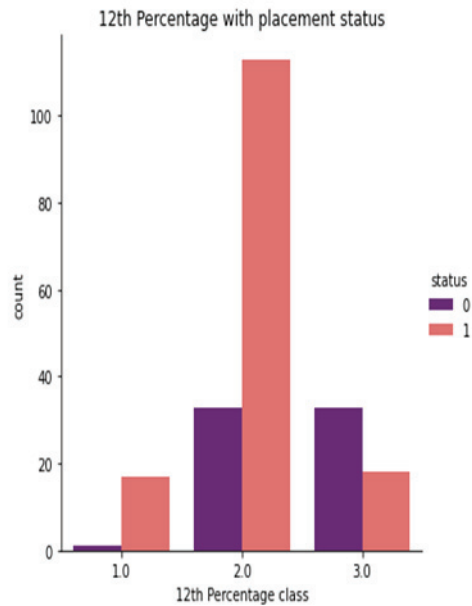


Fig. 7. Placement count vs. 12th percentage

From the above graph, it's observed that all the students having scores between 80-100% in 12th standard got placed. Very few students having scores between 60-80% in 12th standard couldn't get placed. Whereas, most of the students having below 60% score in 12th standard couldn't get placed.

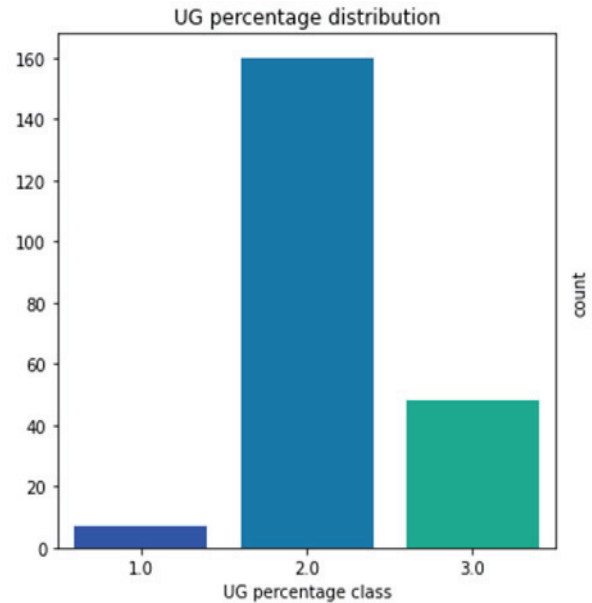


Fig. 8. UG percentage distribution

In the above graph, class 1 represents students having scores between 80-100%, class 2 represents students having scores between 60-80% and class 3 represents students having less than 60 % score in UG degree.

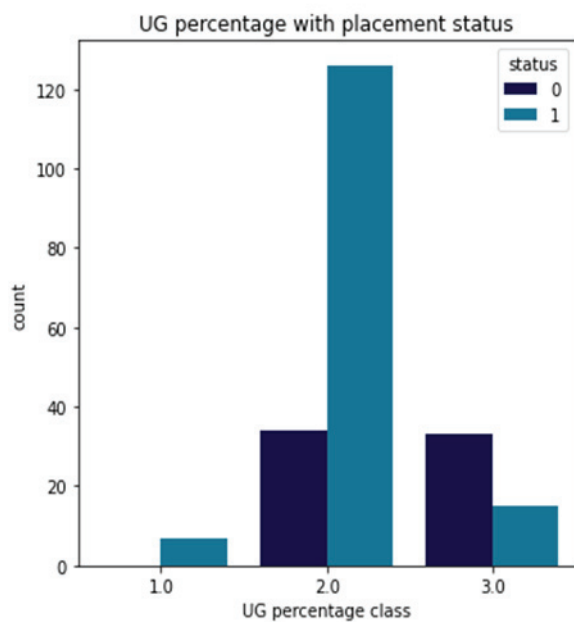


Fig. 9. Placement count vs. UG percentage

From the above graph, it's observed that most of the students having scores between 80-100% in UG got placed. Very few students having scores between 60-80% in UG couldn't get placed. Whereas, most of the students having below 60% score in UG couldn't get placed.

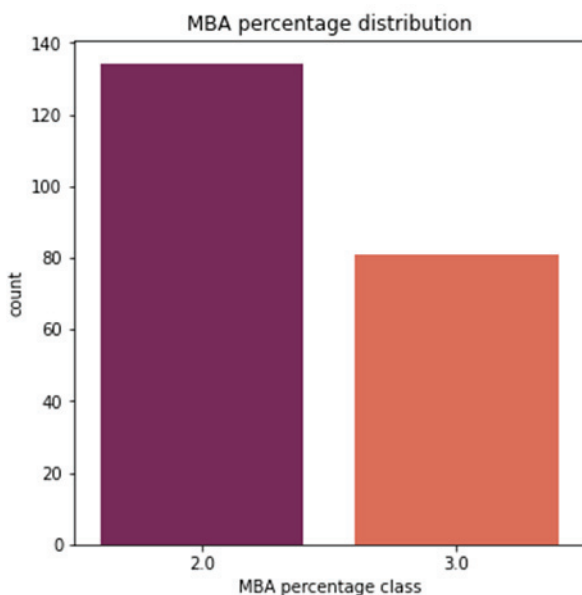


Fig. 10. MBA percentage distribution

After studying MBA percentage data it is observed that no student has secured more than 80% marks. So the class 1 data isn't available for percentage of MBA.

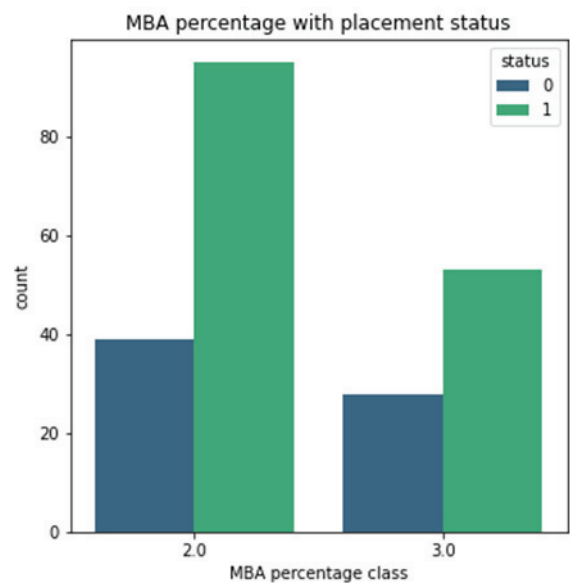


Fig. 11. Placement count vs. MBA percentage

In the above graph we can see that more students from class 2 got placed as compared to class 3.

Hence, it is clear that placement count of the students is dependent on various features such as Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E -test Percentage, Degree Specialization, Degree Percentage.

4) Split data:

Here, data is divided into two parts i.e. training data & testing data. Where 80 % data is taken for training our machine learning algorithm and remaining 20 % data is used for testing whether our trained machine learning model is working correctly or not.

5) Machine Learning Algorithm:

a) Logistic Regression:

Logistic regression is a statistical method used to determine the outcome of a dependent variable (y) based on the values of independent variable (x).

In our problem dependent variable is placement status and independent variables are the features selected by us in the previous step.

This algorithm is mostly used for the problems of binary classification.

b) Decision Tree:

A decision tree is a graph like a tree where nodes represent the position where we select the feature and ask a question, edges represent the answers of the question; and the leaves represent the final output or label of the class.

c) KNN:

K-NN stores all the training data into different classes based on the class labels and classifies new data by checking its similarity with data in the available classes.

d) Random Forest:

Random Forest classifier consists of a number of decision trees which apply on different subsets of our dataset and the average of outputs of all the decision trees is taken to improve the accuracy of output prediction.

6) Evaluate results:

Accuracy is calculated by following formula,

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

Where,

TP: True Positive (the number of cases correctly identified as placed)

TN: True Negative (the number of cases correctly identified as unplaced).

FP: False Positive (the number of cases incorrectly identified as placed)

FN: False Negative (the number of cases incorrectly identified as unplaced)

TABLE I. TP, FP, FN & TN VALUES OF DIFFERENT MODELS

Model	TP	FP	FN	TN
Logistic Regression	16	1	1	25
Decision Tree	13	3	4	23
KNN	14	1	3	25
Random Forest	13	2	4	24

TABLE II. CAMPUS PLACEMENT PREDICTION ACCURACY OF DIFFERENT MODELS.

Model	Accuracy
Logistic Regression	95.34 %
Decision Tree	83.72 %
KNN	90.69 %
Random Forest	88.67 %

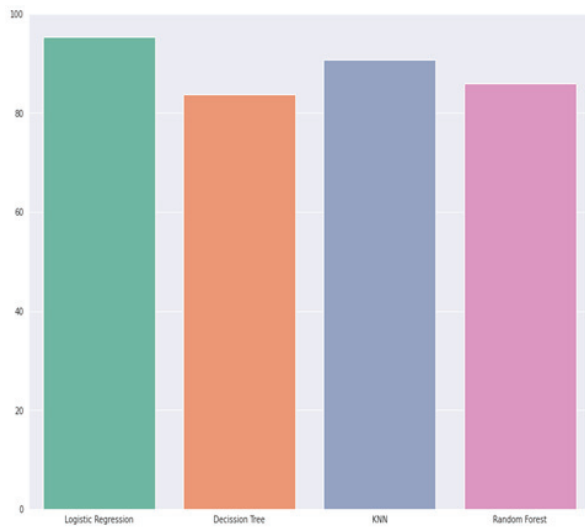


Fig. 12. Comparison of Campus placement prediction accuracy of different models.

III. CONCLUSION

The problem of campus placement prediction can be solved with the help of different machine learning algorithms such as Logistic regression, Decision Tree, KNN & Random Forest.

Here, the Logistic Regression algorithm gave the highest accuracy of 95.34 % for campus placements prediction.

The selected features i.e. Gender, SSC percentage, SSC Board - Central/ Others, HSC percentage, HSC Board, HSC Specialization, Degree Percentage, UG Degree Stream, Work Experience, E -test Percentage, Degree Specialization & Degree Percentage lead to higher classification accuracy.

IV. FUTURE SCOPE

Accuracy may further increase by application of more advanced techniques such as deep learning & experimenting with different activation functions of neural networks such as linear, sigmoid, tan h & ReLU.

We can also experiment with different cross validation techniques such as 3 Fold, 5 Fold, 10 Fold, 15 Fold cross validation in order to analyze the change in accuracy.

REFERENCES

- [1] A. S. Sharma, S. Prince, S. Kapoor and K. Kumar, "PPS — Placement prediction system using logistic regression," 2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE), 2014, pp. 337-341, doi: 10.1109/MITE.2014.7020299.
- [2] S. Elayidom, S. M. Idikkula, J. Alexander and A. Ojha, "Applying Data Mining Techniques for Placement Chance Prediction," 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009, pp. 669-671, doi: 10.1109/ACT.2009.169.
- [3] J. Nagaria and S. V. S, "Utilizing Exploratory Data Analysis for the Prediction of Campus Placement for Educational Institutions," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225441.
- [4] S.Venkatachalam, "Data Mining Classification and analytical model of prediction for Job Placements using Fuzzy Logic," 2021 IEEE International Conference on Trends in Electronics and Informatics (ICOEI), 2021.
- [5] Pothuganti Manvitha, Neelam Swaroopa "Campus Placement Prediction Using Supervised Machine Learning Techniques," 2019 International Journal of Applied Engineering Research, pp. 2188-2191.