Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Placement Analysis for Students using Machine Learning

C. M. Nalayani[1], Thanga Akilan .V[2], Hariharan .S[3], SaranArulnathan[4], Venkatanathan .S[5]

[1]Assistant Professor, Information Technology, Velammal Engineering College, Tamil Nadu, India

[2,3,4,5]Department of Information Technology, Velammal Engineering College, University, Chennai, India

E-mail: [1]nalayini@velammal.edu.in, [2]thangaakilan112@gmail.com, [3]hariselva3478@gmail.com, [4]saranarulnathan5@gmail.com, [5]venkatanathansrinivasan1@gmail.com

Abstract

Every University/College want their students to get placed in a good company with a better package. They create the syllabus so, that students will gain the most knowledge out of the study period. But they are not sure whether the students are getting trained with proper guidance and instruction to be placed. There is a need for metric to find the progress of the student's placement. So, the student can speed up his/her preparation to meet the demands of the minimum criteria of placement. This metric is called the placement analysis system, it takes attributes like internships completed, papers published, aptitude scores, etc to predict where the individual will get placed i.e., Dream company, Core company, Normal company or not get placed. For this the machine learning algorithms are used to predict the results using three basic algorithms Random Forest, Decision tree and K-means clustering the accuracy of the algorithms are determined to find the optimal algorithm. The past data on the placement results were fed as the training dataset and a part of it is used for testing the accuracy of the model. Then if the accuracy is good, this can be used to predict the possibility of a student getting placed. If the student is unhappy with the result, then the model can be used to find the area where the student needs to improve to get to his desired goal. If properly implemented and the students work consistently, this aids in providing solutions to meet every student's goal.

DOI: https://doi.org/10.36548/jitdw.2023.3.001
Received: 09-05-2023. received in revised form: 09-06-2023, accepted: 24-06-2023, published: 05-07-2023

Keywords: Machine Learning, Classification, Decision Tree, Random Forest, Sci-Kit learn, K-Means clustering, Unsupervised learning.

1. Introduction

In the history, Engineers have always revolutionized the world with their innovative ideas, and their works could be never forgotten. It is a myth that that those who hold a degree on Engineering or Technology is an Engineer rather Engineer is a way of thinking and sorting out the solution for the problem. This idea paved the way for the greatest engineers in history who never had a degree.

To find the best engineers for their organisation, HR representatives explore the best institutions and schools. They not only look at the skill of the student but also look for a lot of other qualities, this includes social behaviour, problem solving, hardworking and so on. But educational institutions focus on teaching only the so-called technical skills for the students, this may lead to a lot of problem for the students at the beginning of their job.

All the students are not the same but they are thought to be so. Hence, all the students are taught the same syllabus in the same method. For e.g. student1 may find topic1 to be difficult and may require some extra teaching, whereas student2 may find topic2 to be difficult. Student1 may find the topic 2 to be easy and may not require any revision for it and vice versa.

To find one's proficiency in a particular subject exam are used. Exams were conducted to checks one's strength and weakness. So, the exam scores are used to find the particular student's weakness and work on strengthening it.

Today almost all colleges have a separate department called the placement cell, that shows how important placement is. More than 80 percent of the students wants to get placed and the rest go for either higher studies or entrepreneurship.

Skilled students are having a high possibility of getting placed and this is beneficial first of all to the student as the student gets an employment , second the college as it has secured a job for one of its students and finally the company as it has got a skilled professional for itself. But it does not happen in all cases most of the students are pressurised at the end to learn new skills that does not end in a positive way always. So, it is necessary to find the students who require placement training before a year of their placements I this would be very

helpful in satisfying the demands for getting placed. The students are trained by guiding them in the path they need to take to get placed.

A machine learning model that can predict whether a student can get placed or not is constructed to determine where the student needs to focus. Three training methods are used in the proposed work to develop the model and find which of those is accurate.

2. Related Works

Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar [1] conducted a study to analyse the student placement using Decision tree machine learning algorithm. They used parameters such as school board, 12^{th} mark percent, department, number of arrears, arrear history and CGPA to find the student's placement in dream company or core company or common company or not placed. The Accuracy of the system is found to be 71.66% with tested real-life data.

Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew [2] conducted a study to predict the placement status of students using machine learning algorithms such as Naïve Bayes and K-Nearest Neighbour algorithms (KNN). It considers the parameters such as USN, Tenth and PUC/Diploma results, CGPA, Technical and Aptitude Skills. They conclude by stating that a greater number of factors will give more accuracy to the system.

Syed Tahir Hijazi and S.M.M. Raza Naqvi [3] conducted a study on the factors affecting the student's performance particularly on those of 3^{rd} and 4^{th} year of college. Usually, we expect factors like study hours to influence the results of students but from the study it is proved that study hours do not influence the results of students. Hence, it proves true dependence factors of results.

Surendra Raj Nepal, Bijay Lal Pradhan [4] conducted a study on the statistical analysis of college students' academic performance based on the factors like gender, group, level and system to get the result whether fail or pass. They used binary logistic regression model for their analysis. They concluded that group, level and system have significant impact on the student's academic performance.

Joshita Goyal, Shilpa Sharma [5] conducted a study on Placement Prediction Decision Support System using Data Mining in which they have used naïve bayes and improved naïve bayes. When the algorithms are applied on dataset having 560 instances, naïve bayes gave an accuracy of 80.96% and improved naïve bayes gave an accuracy of 84.7%.

Siddu P. Algur, Prashant Bhat, Nitin Kulkarni [6] conducted a study on Educational Data mining on the classification techniques for recruitment analysis. They have used Random tree and J48 models and it was found that the accuracy of random tree is 85% and that of J48 is 74%.

3. Machine Learning Method

Machine learning [7] is a technique in which the input and its output is given to the computer and allow it to frame the algorithm to get to the result. In other words, machine uses the input and the output to learn to build the code by itself but for that we need to specify the ML algorithm to use. The three main types of machine learning are

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Supervised Learning

It is a type of machine learning [3] which the labelled input as well as the expected outputs were given to the algorithm during the training and the algorithm generates the mapping function that can predict the output for the given input. The major focus of supervised learning is regression and classification problems. Some famous algorithms include linear regression, logistic regression, decision trees, random forest, etc

Unsupervised Learning

It is a machine learning algorithm [3] which takes the unlabelled and unclassified inputs to identify the hidden information by looking at the patterns of the data. The major focus of this kind is clustering and association problems. Some famous unsupervised learning are k-means clustering and appropriate algorithms.

3.1 Algorithms

There are various algorithms available to be used for machine learning for the particular problem but, we are going to analyse our model with three of the machine learning algorithms

- Decision Tree
- Random Forest
- K-Means clustering

3.1.1 Decision Tree

This is also a classification algorithm that uses the features of the input to split it once per feature. This produces a binary tree which can be traversed from top to bottom to reach an output. The model builds the tree such that it has the optimal Gini index. With a lot of training data, the tree becomes more complex and lengthier but the length of the tree can be controlled.

3.1.2 Random Forest

This classification algorithm consists of multiple different decision trees which produce different results, then the majority within the results is taken as the overall result of this algorithm.

3.1.3 K-Means Clustering

This is a classification algorithm in which the data is plotted in a graph to obtain clusters which have a similar property. K stands for the number of clusters. When the testing input is given the model checks to which cluster the point belongs to.

3.2 Framework

Various Python frameworks are used in the project to implement the various processes. The algorithms are implemented using the sci-kit learn framework in python. It is a specialized framework for machine learning that has all the built-in algorithms in it. We also use matplotlib to view the various plots for the model. We use the pandas framework for importing the dataset present in various other formats like csv or excel.

3.3 Architecture

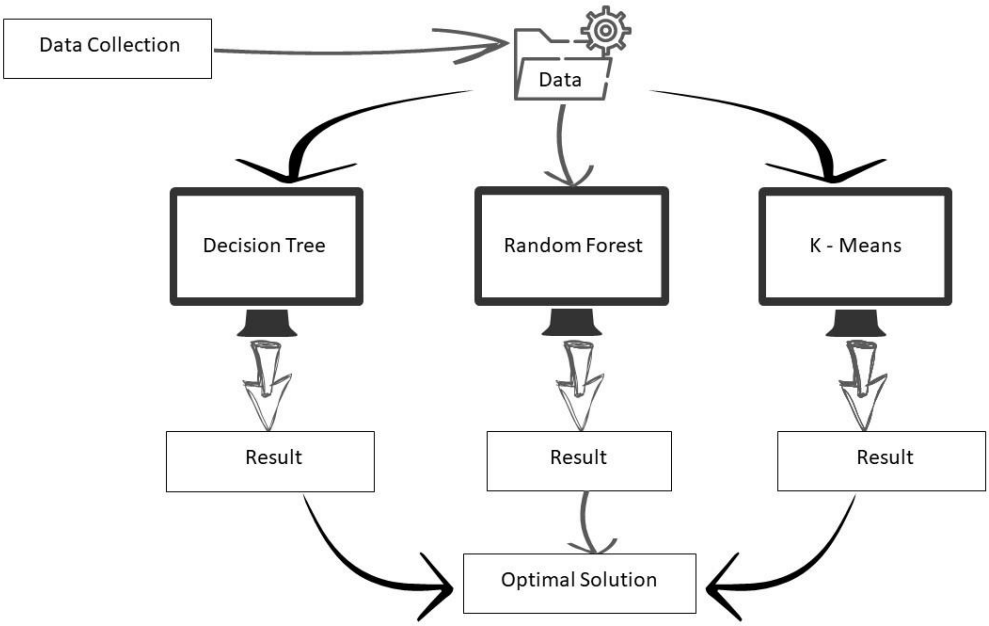

Figure 1. Overall Architecture

3.4 Problems in ML

3.4.1 Underfitting

A Machine learning model is said to have Underfitting if, it performs well in the training data but performs poorly in the testing data. It greatly affects the accuracy of the model. It occurs if the size of the training data is not enough, the model is simple and training data is not cleaned.

3.4.2 Overfitting

A Machine learning model is said to be Overfitted [9] if, it is trained with so much data that it starts learning from the noise and inaccurate data entries. It occurs mainly if the model is too complex and the training data is too large. It is said to occur, if the accuracy is 1 or nearly 1.

3.4.3 Metrics

The correctness of the algorithm in predicting the solution is measured using various metrics. Each type i.e., classification, regression and clustering have their own metrics to check

the accuracy of the algorithm in predicting the outcome. The metrics used for analysing the models are Accuracy Score [11] and F1 Score. The higher these values are the more accurate the model in predicting the results.

3.4.4 Accuracy Score

In multilabel classification, this function computes subset accuracy, the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true. It is a basic accuracy metric for classification.

3.4.5 F1 Score

It is the harmonic mean of precision and recall. A F1 score of 1 is considered best and that of 0 is considered worst. The formula to calculate it is given as

$$F1 = 2 * (precision * recall) / (precision + recall) \quad (1)$$

4. Proposed Work

4.1 Dataset

The data for the model is collected from college students who are pursuing final year and it has the presumed placement result. The dataset contains 50 samples, out of which 75% is used for training the model and 25% is used for testing the accuracy of the model. This splitting is done using the train_test_split() function in python. This data is sent into three models with each of different algorithms. Then the models are coded then the data is passed to the models for training, after which the model can be used to predict the results for the provided input.

4.2 Data Preparation

The dataset for the process is collected from the 3rd and final year students of college using google forms. The data is mostly taken as integer values for reducing the further work on formatting the data for processing. The required information is formatted for excluding the null values.

4.3 Data Transformation

The dataset is then stored in a suitable format like excel (.xls or .xlsx) for easy and efficient retrieval of the same in the future. The dataset is checked for any null value or any other mistakes. These errors were removed manually by checking the whole dataset since the data needs to clean to reduce the possibility of underfitting. The dataset is now ready to be processed into the model.

The dataset is divided 8:2 in which the larger part is used to train the model whereas the smaller part is used to test the accuracy of the model used. The ratio of the division can be changed with the desired ratio.

There are a lot of features that influence the probability of placement of the students including personal reasons like educated parents [3] but we choose some of the very professional and educational factors that influence the placement result. We take the following features specified in the below table.

Table 1. Attributes used for Analysis

Variables	Description	Possible values
Class 12th marks	Mark percent in class 12th	Percentage
CGPA	CGPA till the last semester	Float value (0-10)
Certification	No. of certifications done	Integer
Internships	No. of internships attended	Integer
Prog. Lang.	Known Programming lang.	Integer
Mini-Projects	No. of mini-projects done	Integer
Projects	No. of projects completed	Integer
Competitions	No. of competitions won	Integer
Aptitude	Marks in Aptitude exam	Percentage

These features are provided as the input to get output on where the particular individual will be placed i.e., Dream company, Core company, Normal company, Not-placed or Entrepreneurship.

4.3.1. Decision Tree

Beginning with the decision tree algorithm, as specified we use sci-kit learn to implement this algorithm. Since we have multiple features associated it affects the output of the model to a greater extend. The model is first trained with the larger part of the dataset which optimises the tree with various attributes at the nodes. Then the testing data is sent into the model to predict the output.

The tree created by the model can be viewed using matplotlib.pyplot . This helps to understand how the algorithm have classified the data samples.

4.3.2. Random Forest

The same sci-kit learn framework from python is used to implement the algorithm. The dataset prepared has been passed onto the model, that creates the n number of decision trees as specified and takes the majority result from the trees as the output of the whole random forest.

Since it is an extended version of the decision tree, it's accuracy mostly is higher than decision tree.

4.3.3. K-Means Clustering

The dataset cannot be sent as such like the above cases for this algorithm. It only takes in two-dimensional data. So, we convert the multi-dimensional data into two dimensional using PCA (Principal Component Analysis).

Then the resulting dataset is passed onto the k-means algorithm to plot the points in a clustery manner. Then these clusters will be well defined with more training data. When the testing data is then sent to this model the model predicts the cluster to which it belongs.

5. Result

The tree generated by the training data in the decision tree algorithm can be saved as Png file, which is shown below.

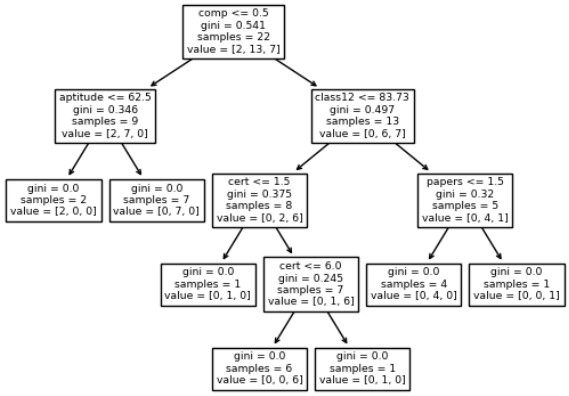

Figure 2. Tree generated by Decision Tree

The mentioned study was conducted and the testing dataset is used to check the accuracy [7] obtained in every algorithm. There are many metrics available to check the accuracy of the model in predicting the output. We use two basic accuracy metrics.

Table 2. Accuracy of Algorithms

	Accuracy Score	**F1_Score**
Decision Tree	0.58	0.53
Random Forest	0.75	0.72
K-Means clustering	0.57	0.27

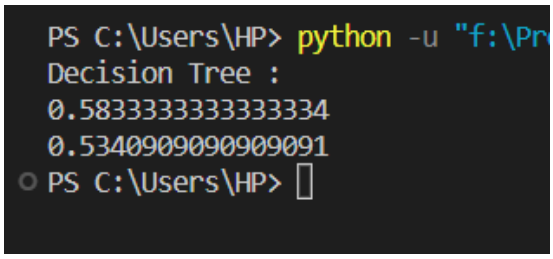

Figure 3. Result for Decision Tree

Accuracy Score and F1 score for decision tree algorithm is low because the algorithm generates a single decision tree randomly using the training data provided and this tree has no

threshold height. Hence, these values reach as low as 0.3 in least efficient tree and as high as 0.75 in some rare high efficient trees. But, the modal value is taken into consideration.

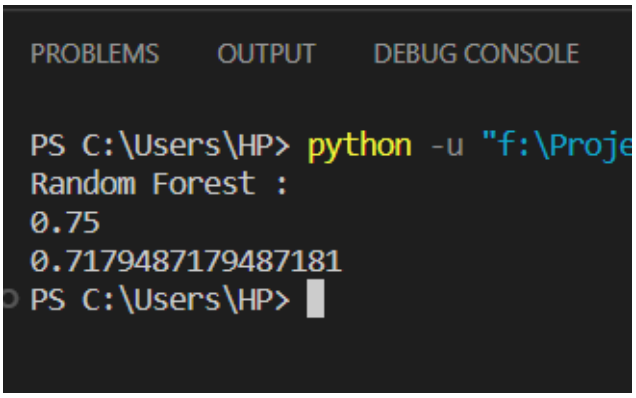

Figure 4. Result for Random Forest

Accuracy score and F1 score for random forest algorithm is high because it creates 100 decision trees with the training data provided and when testing data is sent to these trees the majority answer obtained from the forest is the solution. Hence, it uses multiple decision trees which increases it's accuracy.

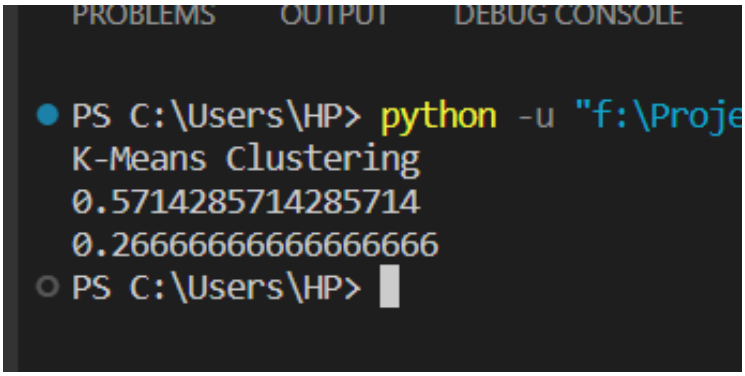

Figure 5. Result for K-Means Clustering

Accuracy score and F1 score for K-means clustering is less compared to others because this algorithm groups the samples using the training data and the output is just two groups.i.e., placed and not placed. Hence, this model is not good for finding the placement possibility but can be used to find the skills one lacks to get placements.

6. Discussion

Out of the 50 samples, 37 samples are used for training with an epoch of 20 to keep both the cost and the accuracy under control. Higher the epoch, higher the cost and the

efficiency of the model. The Loss function is the highest for K-means clustering while it is the least for Random Forest. The loss function is calculated using the correct and incorrect predictions made.

In the Previous Researches, Researchers used very few numbers of attributes that too non-related attributes to analyse the probability of placement. Hence, we used 9 attributes, which are closely related to college studies to analyse placement possibility.

7. Evaluation

The Accuracy scores of these models are above 0.5 which proves that the attributes chosen are relevant to the analysis under consideration. The Accuracy score in Random forest is above 0.75 which is the highest among the three, this assures that this model can be trained in future using more data.

8. Conclusion

In conclusion, the placement analysis system using machine learning algorithms provides a valuable tool for universities and colleges to assess students' placement progress and predict their chances of getting placed in desired companies. By considering attributes such as internships completed, papers published, aptitude scores, and more, the system classifies students into categories like dream company, core company, normal company, or not placed.

The system utilizes three main machine learning algorithms: decision tree, random forest, and k-means clustering. These algorithms analyze past placement data as the training dataset and use a portion of it for testing the accuracy of the model. The accuracy of the models can help predict the placement possibilities for individual students and identify areas of improvement to reach their desired goals.

Previous studies have shown the effectiveness of machine learning in analyzing student placement, and this work builds upon that research. The proposed system takes into account factors beyond just technical skills, considering social behavior, problem-solving abilities, and more.

The implementation of the system relies on Python frameworks such as sci-kit learn, matplotlib, and pandas for machine learning algorithms, data visualization, and data

manipulation. The dataset used for training and testing is collected from college students, and important attributes like 12th marks, CGPA, certifications, internships, programming languages known, and more are considered.

The results of the system can be measured using metrics such as accuracy score and F1 score, indicating the correctness of the algorithm's predictions. The accuracy of the models can be improved by increasing the number of factors considered.

Overall, the placement analysis system using machine learning algorithms offers a promising approach to assist students in their placement journey, helping them understand their placement possibilities and providing guidance on areas of improvement. By implementing this system effectively and consistently, it can help students, colleges, and companies in achieving their placement goals.

References

[1] Senthil Kumar Thangavel, Divya Bharathi P, Abijith Sankar (2017) “Student Placement Analyzer: A Recommendation System Using Machine Learning”

[2] Shreyas Harinath, Aksha Prasad, Suma H S, Suraksha A, Tojo Mathew (2019) “Student placement prediction using machine learning”

[3] Syed Tahir Hijazi and S.M.M. Raza Naqvi (2006) “Factors affecting students’ performance” Bangladesh e-Journal of Sociology. Volume 3. Number 1.

[4] Surendra Raj Nepal, Bijay Lal Pradhan (2022) “A Statistical Analysis of College Students Academic Performance: A Case Study of Amrit Campus” Vol. 3, Issue 1, 108-116

[5] Joshita Goyal, Shilpa Sharma (2018) “Placement Prediction Decision Support System using Data Mining” ISSN: 2395-1303

[6] Siddu P. Algur, Prashant Bhat, Nitin Kulkarni (2016) “Educational Data Mining: Classification Techniques for Recruitment Analysis” I.J. Modern Education and Computer Science, 2016, 2, 59-65 (DOI: 10.5815/ijmecs.2016.02.08)

[7] Kohavi, R. and F. Provost (1998) Glossary of terms. Machine Learning 30:271-274.

[8] Pinky Sodhi, Naman Awasthi, Vishal Sharma (2019) Introduction to Machine Learning and Its Basic Application in Python

[9] Xue Ying (2019) An Overview of Overfitting and its solutions (DOI:10.1088/1742-6596/1168/2/022022)

[10] J. Aggarwal, Essentials of Examination System: Evaluation Tests & Measurement. vikas publishing House, 1997.

Author's Biography

C. M. Nalayini is an Assistant Professor has nearly 15 years of experience in teaching. She has published over 10 research papers in international journals, 10 in International Conference and 11 in National Conference. She has also published book chapters in Springer Book Series. She has given Computer Training to Government School Higher Secondary Students in Chennai. She has taken classes for Women Welfare SHG Group. She has authored books such as C- Programming and Programming and Data Structures-II. Her area of specialization is Network Security. https://www.linkedin.com/in/c-m-nalayini-6a0875203/

Thanga Akilan.V is an aspiring Engineering student who is pursuing under graduation from Velammal Engineering College. He is basically from Information Technology department. He has done projects and attended symposiums and hackathons. He has done several projects in domains like Web development and Machine learning.

S.Hariharan is an aspiring Engineering student who is pursuing under graduation from Velammal Engineering College. He is basically from Information Technology department. He has done courses in NPTEL. He has participated in workshop in UI/UX design and Android development. He is very much interested in domain and loves to explore many new things.

Saran Arulnathan is a student from Velammal Engineering College currently doing me under graduation on Information Technology. He is currently studying Web development domain. He is very much interested in my domain and loves to explore many new things. He has also participated in symposiums and got certifications from online courses like NPTEL. He also like to learn new technologies and gain knowledge from it.

Venkatanathan.S is a student from Velammal Engineering College currently doing me under graduation on Information Technology. He has experience in Web development domain and have got online certifications. He has also participated in UI/UX design, Android app development Workshop and got certifications from online courses. He is very much interested in domain and loves to explore many new things. He also like to learn new technologies and gain knowledge from it