Project report

Student Placement Predicator

Submitted By:

Sumit Kumar

58

Data Cleaning and Preprocessing, E.D.A

Ritik Kumar

46

Model training and evaluation.

Under the guidance of:

Dr. Bharti Rana

Department of Computer Science

November 2024

Content

Abstract

Campus placement plays a vital role in every educational institution in helping students to achieve their goals. Data mining classification can be used as a useful tool for extracting the associated information from the large scale student dataset. Data Science methods have been used broadly in the area of the education system which involves various methods and approach for discovering knowledge. In this paper, a predictive model is designed which can predict the category of placements (dream companies, super dream companies and mass recruiter companies) in which students are eligible by considering their past performance in academics and other curricular activities. The model will also suggest further skills required for future recruitment's which may help the students for placement preparation. The paper also provides real-time experimental results and findings along with performance measures used for model validation which helps in achieving the milestone of outcome-based education in educational institutes as it is given utmost importance in present scenario to ensure better placement prospects in students, which would in turn help the students for carrier building

1. Introduction

The primary aim of students who join professional courses in higher learning institutions is to secure a well- paid job in a reputed organization. Professional education can be either completely technical or it can be managerial as well. This degree is aimed at making students experts in state of the art conjectural as well as practical knowledge in various technical branches. The prediction of placement status that MCA students are most likely to achieve will help students to put in more hard work to make appropriate progress in stepping into a career in various technical fields. It will also help the teachers as well as placement cell in an institution to provide proper care towards the improvement of students in the duration of course. A high placement rate is a key entity in building the reputation of an educational institution. Hence such a system has a significant place in the educational system of any higher learning institution. We used decision tree classifier within Scikit-learn-a machine learning module in python having simple and efficient data mining and data analytics capability- for the implementation of the system.

2. Data

2.1 Data Preprocessing

Data preprocessing is a crucial step, encompassing the cleaning and transformation of raw data to render it suitable for analysis. The goal of data preprocessing is to improve the quality of the data and to make it more suitable for the specific data mining task.

Dropped “StudentId” column because it not relevant for our data analysis.

StudentId does affect our model or is not significant for analysis

All the data in the dataset is in int or float except for 3 columns/variables that are of object data type:

- ExtracurricularActivities
- PlacementTraining
- Placement Status

In sci-kit-learn, the LabelEncoder is a utility class used to encode categorical labels into numerical labels. It essentially converts categorical data (text labels) into numerical abels so that machine learning algorithms can handle them more effectively.

For instance, if you have a categorical feature like "Yes" and "No". The LabelEncoder would assign them numerical labels, such as 0 and 1, respectively.

3. Related Work

Machine Learning techniques has a significant role in deriving innovative knowledge in the educational field so as to help students for their better performance in placement. Many scientists across the world has done considerable amount of work in determining the methodologies for performance analysis and placement. Few of the relevant works in this field are listed out so as to obtain an idea on what has been done so far and what further growth is expected in this area of work.

Hijazi and Naqvi [11] conducted a study to find the factors affecting the academic performance of students.

They made use of questionnaires to elicit information from students highlighting factors such as income factor, parents' educational background, size of the family, regularity of teachers, subject interest created by the teachers and student's interest in co-curricular activities.

They used Pearson Correlation Coefficient to highlight the important factors and they found that mother's education and family income played an important role in students' academic performance.

Pal and Pal [6] conducted a study on student data that have information on their academic records and proposed a classification model to find an efficient method to predict student placements. They concluded that Naïve Bayes classifier is the best classification method for use in placements in comparison with Multilayer Perceptron and J48 algorithms.

Ramanathan, Swarnalatha and Gopal [7] conducted a study using sum of difference method for students' placement prediction. They used the attributes such as age, academic records, achievements etc. for the prediction. They concluded that based on their results higher learning institutions can offerits students a superior education.

Arora and Badal [8] conducted a study to predict student placements using data mining. They made predictions on MCA students in Ghaziabad in UP, considering parameters such as MCA result, Communication skills, programming skills, co-curricular activity participation, gender, 12th result and graduation result. They concluded that their model based on decision tree algorithm can assist the placement cell and faculties in identifying set of students that are likely to face problem during final placements.

Elayidom, Idikkula and Alexander [9] designed a generalized data mining framework for placement chance

prediction problems. They considered the students' Entrance Rank, Gender, Sector and Reservation Category to predict the branch of study that is Excellent, Good, Average or Poor for him/her using decision trees and neural networks.

Naik and Purohit [10] made a study to use prediction technique using data mining for producing knowledge about students of MCA course before admitting them.

4. Methodology

Training:

Logistic regression:

Logistic regression is a statistical method used for binary classification tasks. It predicts the probability of a categorical outcome by fitting data to a logistic function, transforming values into probabilities between 0 and 1.

Decision Tree:

In machine learning, a decision tree is a model that resembles a tree-like structure used for classification and regression tasks. It starts with a root node that represents the entire dataset and splits the data into branches based on specific features and their values. This process continues down the tree until reaching leaf nodes, which provide final decisions or predictions.

K-Nearest Neighbors (KNN):

The K-Nearest Neighbors (KNN) Classifier is a straightforward yet effective supervise machine learning algorithm for classification tasks. It operates on the principle of similarity classifying new data points based on their proximity to labeled data points in the training set. The algorithm determines the class of a new data point by identifying its K nearest neighbors in the feature space using a distance metric (like Euclidean distance).

5. Results

- Primary objective achieved: Successfully predicted students' placement status post-final year using four classification algorithms (LR, DTC, KNN, RFC).
- In Future work we can apply this predictor Model on real time application instead of synthetic data.
- System's efficacy: Elevates institution's placement rates. Enhances institution's reputation.
- Signifies a substantial advancement in classification techniques for placement prediction.
- Stands as a pivotal tool to improve placement prediction methodologies significantly.

6. Discussion and Conclusion

Out of the 289 students, the system predicted 33.91% students to get placed in mass recruiter or common company class. The number of students who would get into Core Company and Dream Company are found to be in and around 18% each. System predicts 150% more boys than girls are most likely not to get placed in any tier of companies. These predictions when compared across the real life data of 60 students, gave 71.66% accuracy, which is a significant accuracy measure to consider a prediction system as reliable. The accuracy of the system will be improved once, the outliers in the test data are removed.

The system has a considerable matching when compared to results obtained from weka tool. Since weka is a highly used data analysis tool, 84% matching with weka results further underscores the reliability of the system. The analysis of various algorithms in weka proves that decision tree classifier stands out with 0.01 seconds of running time and 84.42% accuracy. This proves the efficiency of the methodology employed in the system.

7. References

- Senthil Kumar Thangavel, Divya Bharathi P and Abijith Sankar: Student Placement Analyzer: A Recommendation System Using Machine Learning 2017 International Conference on Advanced Computing and Communication Systems.
- Student Placement Prediction Model: A Data Mining Perspective for Outcome-Based Education System Article in International Journal of Recent Technology And Engineering(IRJTE) September 2019:-By Abhishek Rao, NMAM Institute of Technology
- 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) A Review on Student Placement Chance Prediction
- 2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN) Campus Placement Predictive Analysis using Machine Learning
- https://www.kaggle.com/datasets/chandhurubaskar/campus-placement-data-for-engineering-colleges/code?datasetId=3678158&sortBy=voteCount