WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predictive Modeling of Student Placements with Decision Tree

Nagaraj M Lutimath[1], Shruti M Jolad[2], Sohan G[3], Sumant Kulkarni[4], Udith Hegadera M S[5], Sidda Yashaswini[6], Shwetha N[7], Sangamesh Gangadhar Madagi[8], Sneha[9], Sneha D L[10], Shivanandan V M[11], Sanjana B[12]

[1]nagarajlutimath@gmail.com, [2]shrutijolad2005@gmail.com, [3]sohan040927@gmail.com, [4]sumantskulkarni777@gmail.com, [5]udithmsudith@gmail.com, [6]siddayashaswini8@gmail.com, [7]shwethanaganna93@gmail.com, [8]sangumadagi@gmail.com, [9]snehha2205@gmail.com, [10]lokeshlokesh.1234m@gmail.com, [11]shivanandannandu548@gmail.com, [12]sanjanabetageri60@gmail.com

[1,2,,3,4,5,6,7,8,9,10,11,12]Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology and Management, Udayapura, Kanakapura Main Rd, Bengaluru-560082

Abstract:

Machine Learning (ML) has become an essential tool for analyzing large and complex datasets, enabling the discovery of hidden patterns and supporting data-driven decision-making. In the field of higher education, the placement outcomes of graduates serve as a key indicator of institutional performance and effectiveness. As such, the analysis of student placement data is instrumental in shaping future academic strategies and enhancing institutional growth.

This study applies ML techniques to analyze placement records of students in a higher educational institution. Through the integration of visual analytics and statistical modeling, the research uncovers meaningful insights into placement trends. The analysis reveals that students pursuing an MBA with a specialization in Marketing and Finance tend to have higher placement rates. Additionally, a significant portion of the placed students hold academic backgrounds in Commerce and Management. Interestingly, the study finds that employability test scores do not have a substantial impact on placement outcomes.

Overall, the findings demonstrate the effectiveness of Machine Learning in generating actionable insights that can inform academic planning, curriculum development, and alignment with industry needs, ultimately enhancing the employability of graduates. Python programming language is used to analyze the data set.

Keywords: Decision Tree, Machine Learning, Support Vector Machines

1. Introduction

In today's data-driven world, high-volume and rich datasets act as the new fuel powering intelligent systems and strategic decision-making. Much like crude oil in the energy sector, raw data in its unprocessed form holds immense potential but must be refined to unlock its true value. Systematic processing, much like the principle of fractional distillation, enables the extraction of meaningful insights from complex data at multiple levels. These insights serve as a foundation for informed decision-making, enhancing strategic outcomes across both business and institutional environments. Machine Learning (ML) serves as a transformative approach for extracting such insights. By analyzing datasets and often presenting results through visualizations, ML facilitates a clearer understanding of complex information and supports better decision-making processes for organizations and educational institutions alike.

In the realm of higher education, campus placement is a key performance indicator reflecting the effectiveness of an institution's academic model. It provides tangible evidence of the institution's ability to prepare students for the workforce and serves as a benchmark for prospective students evaluating future educational paths. Thus, placement data becomes an essential asset in assessing both institutional success and student outcomes.

By applying ML techniques and statistical models to this data, institutions can uncover valuable patterns and trends influencing placement outcomes. Visualizing this data not only enhances comprehension but also enables stakeholders to derive meaningful interpretations. ML-based analysis of placement records allows for deeper exploration and interpretation of educational impact, supporting future planning, curriculum improvements, and strategic development within the institution.

The rest of the paper is organized as follows, Section 2 gives the related works, Section 3 explains the decision tree constructed, Section 4 explain the support vector machine studied, Section 5 ends with the conclusion.

2. Related Works

ML is the initial step of deciphering data by first showing the visual representation using different tools available in a data processing tool. It helps in summarizing your findings and display the data with graphics and can help in interpretation and finding the underlying patterns or trends [1]. Humans, by nature, are attracted to visuals, and it is one of the fastest methods to recognize a result rather than reading the text. Visualizations can simply explain any complex idea.

The most crucial step is to study the data thoroughly- identify if it is structured data or unstructured data, observe the number of columns and rows of structured data, read the structure of the data, find out the important columns that are useful in the dataset, find out the number of null values or missing data and decide either to replace it or remove it from the dataset. The data is first cleaned and trimmed using inbuilt functions of the tool such as *filter*, *select*, *group by* and others.

Fig 1 displays the generic Exploratory Data Analysis (EDA) process to a different group of datasets. The significant subdivisions are the categorical data and numerical- continuous or discrete data. Followed along with its type of plots such as histogram can be plotted through continuous, whereas a box plot can be plotted between the continuous and categorical data.

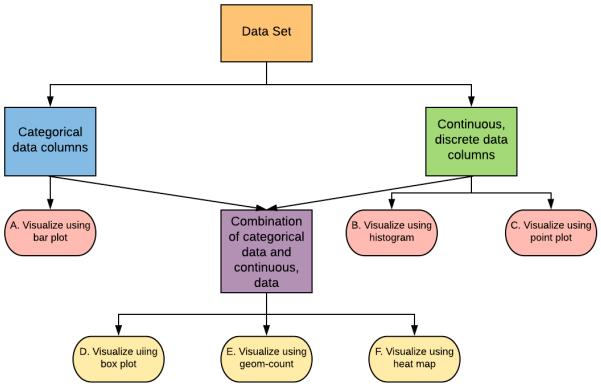

Fig 1: A generic EDA process

The dataset is about campus recruitment shows the influencing factors of academic and employability that helps in the placement of a student. The dataset was downloaded from Kaggle [2]. The Campus Recruitment data includes information regarding gender, Board of education, secondary schools, higher secondary schools percentage, and specialization. Along with the degree type, work experience, degree percentage, the salary of the placed students. The dimension of the dataset is 215 records and 15 columns.

The second most essential step is the data preprocessing part where the *NA* or *Null values* are handled [3]. Data preprocessing allows us to deal with inconsistent values such as the address in place of phone number and removable of duplicate values and also the extra spaces.

So we have removed the salary component as it coincides with the students who are placed and not placed in the status attribute. ML is comparable to the storytelling of the analysed data. The data was analysed using R. R is open-source and can import many formats of data [4][5]

ML is applied to the dataset. The plots and helps us to gain insight into the dataset [6]. The package 'ggplot2' was used to create graphics by providing the data and mapping variables to aesthetics [7][8]. A histogram plot breaks the data into bins and shows its frequency distribution of these bins.

A bar plot provides a comparison between cumulative tools across several groups. The Point plot is also known as Scatter Plot. In R programming language Geom_point() function helps in understanding how one variable changes with respect to other variables. It is used for two continuous values. Geom_box() can be used to visualize the spread of data and derive inference accordingly [9]. A box plot shows five most important numbers- the minimum, the 25th percentile, the median, the 75th percentile, and the maximum [10]. Heart disease prediction using support vector machine is studied [11]. Heart disease prediction using hybrid technique is studied [12]. Regression analysis for liver disease is studied for better understanding of machine learning techniques [13]. We study and compare decision tree and support vector machine for placement data in this research paper.

3. Decision Tree

The Decision Tree algorithm is among the most commonly used machine learning techniques. As a supervised learning approach, it is suitable for both categorical and continuous variables [14]. The model's accuracy depends significantly on how the data is split at each node. An optimal split is one that clearly separates instances with different class labels into distinct subsets [15].
In contrast, the Random Forest algorithm is a powerful and flexible ensemble method used for both classification and regression tasks [16]. It constructs multiple decision trees from different subsets of the training data and aggregates their results, thereby producing a more generalized and accurate model. This ensemble approach reduces overfitting and enhances predictive performance.

In this study, the campus recruitment dataset was divided into a training set and a test set using a 70:30 ratio, comprising 151 training instances and 64 test instances. A Random Forest model was trained with 'status' as the target variable and all other features as input variables.

To evaluate feature relevance, Information Gain was computed for each independent variable using the model's feature importance function. This metric identifies how effectively each feature contributes to data splitting. A higher Information Gain value indicates a stronger predictive capability. Fig 2 displays the structure of the decision tree built from the dataset. Table 1 presents the Information Gain values for each feature. Fig 3 visualizes the relative importance of the features using a bar chart.

From the results, mba_1 had the highest Information Gain, signifying its strong influence on the placement outcome, while hsc_b had the lowest. Furthermore, Table 2 shows the classification report for the decision tree model, which achieved an accuracy of 74%, indicating a reasonable level of predictive performance.

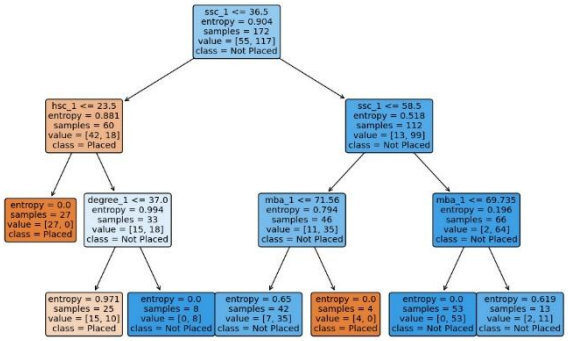

Fig 2: Decision Tree for the given data set

S.No	Feature	Information Gain
1	gender	0.0058
2	ssc_1	0.6227
3	ssc_b	0.0010
4	hsc_1	0.5411
5	hsc_b	0.0002
6	hsc_s	0.0035
7	degree_1	0.5066
8	degree_t	0.0091
9	workex	0.0600
10	etest_1	0.3899
11	specialization	0.0454
12	mba_1	0.8485

Table 1: Information Gain of Various features in the data set

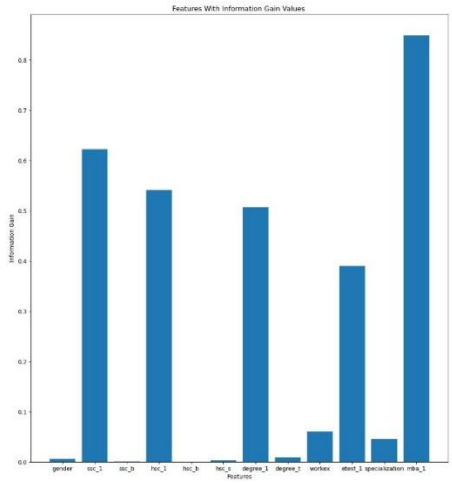

Fig 3: The bar chart diagram of the information gain values of various features of data set

Attribute	Precision	Recall	f1-score	Support
0	0.56	0.42	0.48	12
1	0.79	0.87	0.83	31
accuracy			0.74	43
macro avg	0.67	0.64	0.65	43
weighted	0.73	0.74	0.73	43

Table 2: Classification Report for the decision tree for the given data set

4. Support Vector Machine

Support Vector Machine (SVM) is a widely used supervised learning algorithm for classification tasks in machine learning. It functions by determining the most effective hyperplane that separates data points into distinct target classes. In this analysis, an SVM model was implemented using the sigmoid kernel to evaluate its performance on the campus recruitment dataset. The results of this model are presented in Table 3, which contains the classification report. The SVM model achieved an accuracy of 72%.

When compared to the Decision Tree model, which achieved an accuracy of 74%, the SVM demonstrated slightly lower predictive performance for this particular dataset.

Attribute	Precision	Recall	f1-score	Support
0	0.00	0.00	0.00	12
1	0.72	1.00	0.84	3
accuracy			0.72	43
macro avg	0.36	0.50	0.42	43
weighted	0.52	0.72	0.60	43

Table 3: Classification Report for the support vector machine for the given data set

5. Conclusion

This study applied various machine learning techniques, including Decision Tree and Support Vector Machine (SVM), to analyze campus recruitment data and predict student placement outcomes. The performance of each model was evaluated using classification metrics. Among the models tested, the Decision Tree classifier achieved the highest accuracy at 74%, followed closely by the SVM with a sigmoid kernel, which attained 72% accuracy. The Decision tree model further reinforced the importance of individual features through Information Gain analysis, with mba_1 emerging as the most influential variable. The results emphasize the value of machine learning approaches, particularly decision tree-based model in uncovering actionable insights from educational data. These insights can play a crucial role in helping institutions identify placement patterns, evaluate the effectiveness of academic programs, and implement strategic improvements. Ultimately, such data-driven decisions can lead to more industry-relevant curricula and increased employability for graduates.

REFERENCES:

1. S. M. Thaung et al., “Exploratory Data Analysis Based on Remote Health Care Monitoring System by Using IoT,” Communications, vol. 8, no. 1, pp. 1–8, 2020.
2. “Campus Recruitment | Kaggle.” [Online]. Available: https://www.kaggle.com/benroshan/factors-affecting-campus placement/kernels. [Accessed: 30-Apr-2020].
3. “RStatTutorial_Basic.” [Online]. Available: http://cis.csuohio.edu/~sschung/CIS660/RStatTutorial_BasicLab1. [Accessed: 18-Jun-2020]. [
4. J. Tuimala and A. Kallio, “R, Programming Language,” in Encyclopedia of Systems Biology, New York, NY: Springer New York, 2013, pp. 1809–1811.
5. Cox, Victoria. "Exploratory data analysis." Translating Statistics to Make Decisions. Apress, Berkeley, CA, pp. 47-74, 2017.
6. X. Qin, Y. Luo, N. Tang, and G. Li, “Making data visualization more efficient and effective: a survey,” VLDB J., vol. 29, no. 1, pp. 93–117, 2020.
7. H. Wickham, “ggplot2 by Hadley Wickham,” Media, vol. 35, no. July, p. 211, 2009.
8. K. Ito and D. Murphy, “Tutorial: Application of ggplot2 to pharmacometric graphics,” CPT Pharmacometric Syst. Pharmacol., vol. 2, no. 10, p. e79, Oct. 2013.
9. J. S. Kendrick, D. F. Williamson, P. D, R. A. Parker, and J. S. Kendrick, “The box plot : a simple visual method to interpret data . Ann Intern Med 110 : 916 The Box Plot : A Simple Visual Method to Interpret Data,” Acad. Clin., vol. 10, no. July 1989, pp. 916–921, 1989.
10. C. Thirumalai, M. Vignesh, and R. Balaji, “Data analysis using box and whisker plot for lung cancer,” 2017 Innov. Power Adv. Comput. Technol. i-PACT 2017, vol. 2017-January, no. March, pp. 1–6, 2017.
11. Nagaraj M. Lutimath, Arathi B N, Shona M, “Prediction of Heart Disease using SVM”, International Journal of Recent Technology and Engineering

(IJRTE) ISSN: 2277-3878, Volume-8 Issue-2S6, August 2019.

12. Nagaraj M. Lutimath, Chandra Mouli, B. K. Byre Gowda, K. Sunitha "Prediction of Heart Disease Using Hybrid Machine Learning Technique", Transactions on Computer Systems and Networks, Paradigms of Smart and Intelligent Communication, 5G and Beyond, Springer Series.2023.
13. Nagaraj M. Lutimath, D. R. Arun Kumar, C. Chetan , "Regression Analysis for Liver Disease Using R: A Case Study", International Conference on Innovative Computing and Communications, Lecture Notes in Networks and Systems, Proceedings of ICICC 2018, Volume 2, 2018.
14. J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random Forests and Decision Trees," Int. J. Comput. Sci. Issues, vol. 9, no. 5, pp. 272–278, 2012.
15. D. Berrar and W. Dubitzky, "Decision Tree," in Encyclopedia of Systems Biology, Springer New York, 2013, pp. 551–555.
16. L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.