How to Cite:
Manimaran, P., Priyadharshini, G., & Yamuna Devi, N. (2022). Predicting the eligibility of placement for students using data mining. *International Journal of Health Sciences*, *6*(S8), 460–467. https://doi.org/10.53730/ijhs.v6nS8.9723

Predicting the eligibility of placement for students using data mining

Dr. P. Manimaran
Professor, Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode-637 215, Namakkal District, Tamil Nadu, India
Corresponding author email: manimaran@ksrct.ac.in

Priyadharshini G
Students, Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode-637215, Namakkal District, Tamil Nadu, India

Yamuna Devi N
Students, Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode-637215, Namakkal District, Tamil Nadu, India

***Abstract*---**Every educational institution uses campus placement to help students achieve their goals. Data mining techniques are frequently used in the educational system, which employs a variety of learning methods and approaches. A predictive model is built based on academic and extracurricular achievements to determine which category of placements (dream businesses, super dream companies, and mass recruiter companies) students are suitable for. Admission and the name of the institution are heavily influenced by placements. The main goal of this paper is to analyze previous year's applicant data in order to predict current students' placement chances and help institutions increase their placement percentage . Our project's main goal is to determine location using student data and to implement KNN models, logistic regression models, random forest models, and SVM models. Our project represents the candidate placement classification provided in the available data set. These algorithms predict the results on their own, and then we compare the algorithms' efficiency based on thedataset. This model assists a company's position cell in screening potential candidates and focusing and improving their technical and soft skills at regular intervals. In our project, the parametersare their grades/grades in quants, verbal, logical reasoning, programming, cgpa, networks, cloud computing, web services, data analysis, QA, AI, and location.

Manuscript submitted: 27 March 2022, Manuscript revised: 9 May 2022, Accepted for publication: 18 June 2022

***Keywords*---**placement prediction, educational data mining, logistic regression, random forest, support vector machine, accuracy comparison.

Introduction

Businesses use data mining to extract specific data from massive databases in order to solve business problems. Its primary purpose is to transform raw data into useful information. Data mining is the process of identifying anomalies, patterns, and correlations in large datasets in order to predict future outcomes. This is achieved by combining three intertwined disciplines: statistics, artificial intelligence, and machine learning. The techniques, tools, and research used to automatically extract meaning from large repositories of data generated by or related to people's learning activities in educational settings are referred to as educational data mining. This information is frequently comprehensive, fine-grained, and precise. Student performance, dropouts, and teacher performance can all beclassified

and predicted using educational data mining. It can help educators track academic progress to improve the teaching process, students choose courses, and educational management to be more efficient and effective. Educational data mining techniques have produced a variety of phenomena relating to student learning on an online platform, as well as consistently achieving higher accuracy. There are important aspects that must be investigated in order to justify the extraordinary progress for educational data, which is developing recognition that not all critical data is stored in a single data stream.

Literature Review

This paper proposes that all college students wish to complete a task that will provide for their arms before they leave college. A placement risk predictor assists students in understanding where they stand and what needs to be done to achieve an excellent placement. A placement predictor is a machine that anticipates the opportunity or type of organization in which a pre-very-last-year student may be placed. Many predictor fashions have been introduced with the emergence of information mining and system learning by reading the previous 12 months pupil's dataset [1].This paper suggests Data mining is a technique for discovering patterns in large amounts of data. Knowledge discovery in data reveals within the software of cutting-edge gadget learning techniques such as regression, classification, clustering, and so on. A complex technique of information preprocessing, together with information cleansing and information transformation, was applied to the considered data set so that it could be used in numerous class responsibilities. For class responsibilities, K-nearest friends and selection tree algorithms were used, and the accuracy of the class was determined using the holdout method. The outcome of using the more advanced state-of-the-art approach for information set partitioning, which employs the K-method clustering approach, is also presented [2]. This work provides a version that could be anticipating the student's possibility of dropping the use of information from the primary three semesters attended with the aid of Computer Science

Undergraduate college students (N=1516) from Federal University of Pelotas. The CRISP-DM technique information from the Cobalto Management System is used in this work. The outcomes are demonstrated for three algorithms, and for the Random Forest set of rules, a precision of 95.12 percent and a recall of 91.41 percent are provided, indicating that it is far more feasible to apply a prediction model utilizing only information from the first three semesters of the course [3]. Machine learning is used in this paper to forecast placement using Naive Bayes and the K-nearest neighbor (KNN) algorithm. The methodology considers USN, tenth, and diploma results, as well as CGPA, technical, and aptitude abilities. They use two machine learning classification algorithms: the Naive Bayes Classifier and the K-Nearest Neighbors [KNN] algorithm. We evaluate the algorithms' effectiveness using the dataset after they each predict the results independently. Several previous research articles focused on a small number of characteristics for predicting placement status, such as CGPA and arrears, yielding less accurate results; however, the proposed study includes many educational parameters to predict placement status, yielding more accurate results [4]. We will use data mining principles and techniques under Classification in this paper to accomplish this. We are also attempting to create a data collection containinginformation about students' gender, marks and rank in entrance examinations, and third-year results from the previous batch of students. These data sets were analyzed to produce the final answer. Data is mapped into specified groups or classes in the classification data mining approach. It is a supervised learning method that uses training data to create rules for categorizing test data into specific groups or classes. The procedure is divided into two sections. The first phase, which is the learning phase, examines the training data and generates classification rules. The second phase is the Classification phase, which involves categorizing test data based on the training data set. On this data, the ID3, C4.5, Improved weighted modified ID3 classification algorithms are used to forecast general and individual performance of third-year students in future examinations [5].

Methodology

Quants, logical reasoning, verbal, programming, cgpa, networking, cloud computing, web services, data analysis, quality assurance, AI, and placement are some of the categorical types in our dataset. Following that, preprocessing is used to remove non-essential characteristics such as Register no. that do not (will not) play a role in the analysis. It entails categorizing students or users into appropriate clusters using a data mining model based on the k-nearest neighbor method, with the goal of assisting them in developing their abilities and mentality.The results of various models, such as K-NN, logistic regression, random forest, and SVM, are compared to find the best solution. Applying all of those algorithms and obtaining their training and test sets for each one. Determine the accuracy of all algorithms based on the trainingand test sets, as well as which algorithm is effective.

Dataset

A data set (also known as a dataset) is a collection of information. In the case of tabular data, a data set corresponds to one or more database tables, where each

column of a table represents a specific variable and each row represents a specific record of the data set in question. The data set lists values for each variable, such as an object's height and weight, for each member of the data set. Every value is known as a datum. A data set can also include a collection of documents or files. We provide the placement dataset as input. Used to predict where candidates will be placed based on parameters. We'll go over how to forecast a student's placement status using techniques like the logistic regression method, KNN methods, random forest methods, SVM methods based on various student variables. In the future, parameters will be an important factor to consider when selecting and extracting features.

Pre-Processing

Data preprocessing is critical to the model-building process. If the data is not properly pre-processed, it can cause significant errors in the model's final output. We did the following as part of the data preprocessing:

1. Data Cleaning : The missing values in the dataset were addressed during data cleaning.
2. Normalization : Data normalization is the process of converting the values of numeric columns in a dataset to a common scale of 0 to 1. The range of the salary attribute and marks in our dataset required data normalization. As a result,the numeric values had to be brought into a common range.
3. Dropping columns :Dropping unneeded columns is critical and can have a significant impact on the model's performance.

K-NNAlgorithm

KNN is a non-parametric classification method. It is also one of the most well-known classification methods. The basic idea is that known data is ordered in a space defined by the features selected. When new data is fed into the algorithm, it compares the classes of the k closest data to determine the class of the new data. The KNN classification has several advantages, the most important of which is its ease of use. It is also a highly effectivemethod.

Logistic Regression

Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable in its most basic form, though many more complex extensions exist. Logistic regression (or logit regression) in regression analysis is used to estimate the parameters of a logistic model. Logistic regression is a statistical analysis method that uses prior observations of a data set to predict a binary outcome, such as yes or no. Logistic regression has grown in importance in the field of machine learning.

Random Forest

Random Forest is a classifier that combines and averages the results of several decision trees on different subsets of a dataset to improve the dataset's predicted accuracy. Random Forest is a Supervised Machine Learning Algorithm used to

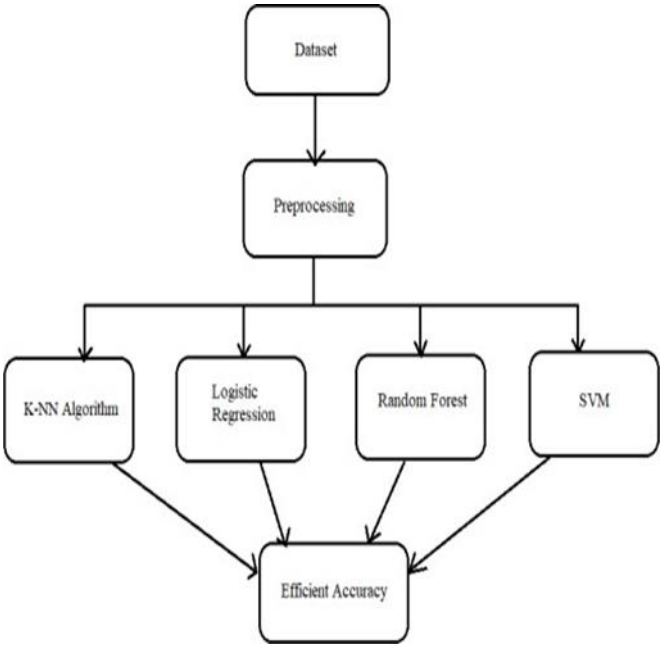
solve classification and regression problems. It builds decision trees from different samples by using themajority vote for classification and the average for regression.

Support Vector Machine

A Support Vector Machine (SVM) classifies data by locating the hyperplane with the smallest difference between two classes. The vectors (cases) that define the hyperplane are known as support vectors. The beauty of SVM is that if the data is linearly separable, there is a single global minimum value. A hyperplane should completely separate the vectors (cases) into two non-overlapping classes in an ideal SVM analysis. Perfect separation, on the other hand, may be impossible to achieve or may result in a model with too many examples to correctly classify. In this case, SVM identifies the hyperplane that maximizes the margin while minimizing misclassifications.

Efficiency Accuracy Comparison

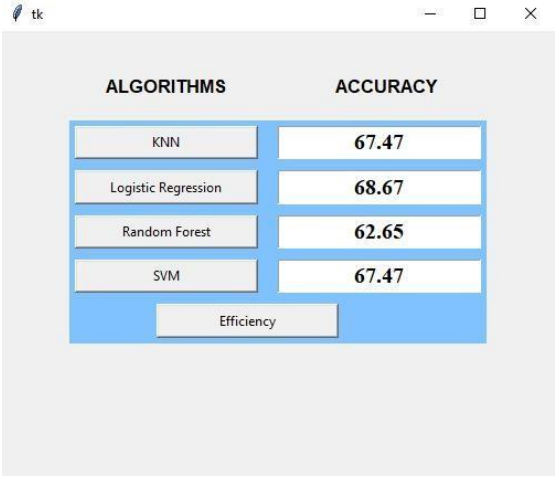
After calculating all of the efficiencies of these four algorithms, we will compare their efficiencies to determine which algorithm provides the highest accuracy.



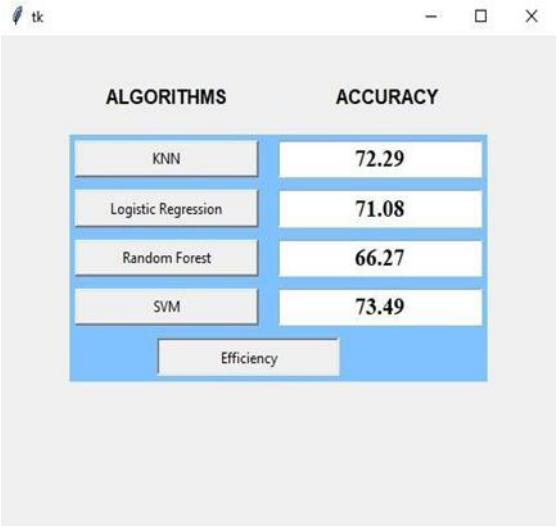
Result and Discussion

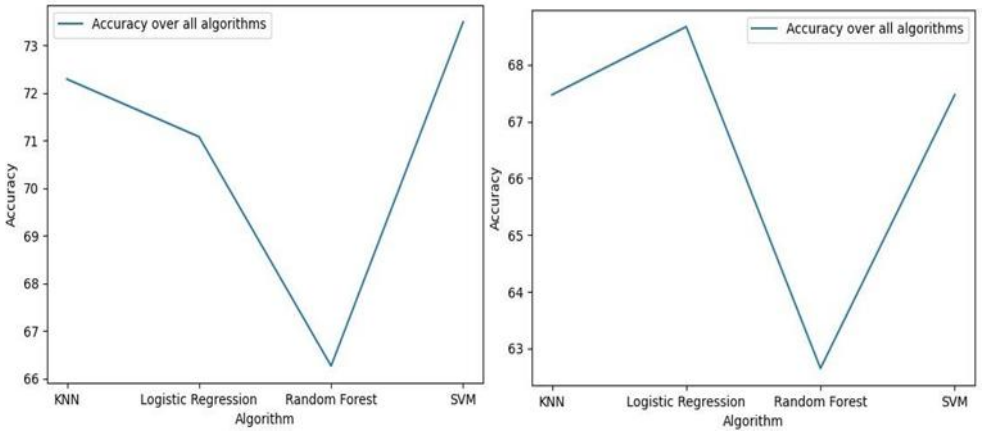
Using the provided algorithms and additional Python packages, we determine the correctness of various methods. Using K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machine, we can predict students' placement information with high accuracy (SVM).
K-Nearest Neighbor : The final accuracy obtained for the test data set using the Random Forest model was 66.27%.

Logistic Regression : The final accuracy obtained for the test data set using the Random Forest model was 68.67%.
Random Forest : The final accuracy obtained for the test data set using the Random Forest model was 62.65%.
SVM : The final accuracy obtained for the test data set using the Random Forest model was 67.47%. This is for one dataset,



K-Nearest Neighbor : The final accuracy obtained for the test data set using the Random Forest model was 72.29%.
Logistic Regression : The final accuracy obtained for the test data set using the Random Forest model was 71.08%.
Random Forest : The final accuracy obtained for the test data set using the Random Forest model was 66.27%.
SVM : The final accuracy obtained for the test data set using the Random Forest model was 73.49%.





Conclusion

Educational data mining is becoming more popular as a result of the large number of students enrolling in higher education, according to this work. Fraternity for students Educational data mining is becoming more popular. It can help with both descriptive and predictive analytics. Precision forecasting and profiling will not only improve prediction accuracy, but they will also improve prediction accuracy. not only to improve educational standards, but also to promote good learning The experience will be beneficial to the student fraternity. The factors in our work are their grades/marks in Quants, verbal, logical reasoning, programming, cgpa, networking, cloud computing, web services, data analytics, quality assurance, AI, and placement. The proposed method predicts student data using K-NN and the logistic regression algorithm based on previous student datasets. KNN, logistic regression, random forest, and svm are used for both training and testing the dataset to determine the accuracy of each method and which algorithm is most efficient for the dataset. After calculating all of the efficiencies of these four algorithms, we will compare their efficiencies to determine which algorithm provides the highestaccuracy.

References

1. A review on student placement chance prediction,Liya Claire Joy,Asha Raj,2019 IEEE.
2. Application of KNN and Decision Tree Classification Algorithms in the Prediction of Education Success from the Edu720 platform ,Omar Dervišević, Emir Žunić, Dženana Đonko, Emir Buza 2019 IEEE.
3. Prediction analysis of student dropout in a Computer Science course using Educational Data Mining , Alexandre G. Costa, Emanuel Queiroga, Tiago T. Primo, Júlio C. B. Mattos, Cristian Cechinel,2020 IEEE.
4. Prediction of placement of candidates using KNN, Logistic Regressionand
5. Random Forest model Antarjita Mandal, Sai Shruthi S, Yogadisha S, 2021 IJCRT. [5]Prediction System for Student Performance Using Data Mining Classification,Rahul Patil, Sagar Salunke, Madhura Kalbhor, Rajesh Lomte 2018 IEEE.

6. Rinartha, K., & Suryasa, W. (2017). Comparative study for better result on query suggestion of article searching with MySQL pattern matching and Jaccard similarity. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)* (pp. 1-4). IEEE.
7. Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.
8. Susilo, C. B., Jayanto, I., & Kusumawaty, I. (2021). Understanding digital technology trends in healthcare and preventive strategy. *International Journal of Health & Medical Sciences*, *4*(3), 347-354. https://doi.org/10.31295/ijhms.v4n3.1769