Student Placement Prediction Using Machine Learning Algorithms

Dr. P Aruna
Department of Computing
Coimbatore Institute of Technology
Coimbatore, India
paruna@cit.edu.in

Abinayashri V S
M.Sc. Software Systems
Coimbatore Institute of Technology
Coimbatore, India
71762131003@cit.edu.in

Abinayaa Sree K S
M.Sc. Software Systems
Coimbatore Institute of Technology
Coimbatore, India
71762131002@cit.edu.in

Meghashri A
M.Sc. Software Systems
Coimbatore Institute of Technology
Coimbatore, India
71762131028@cit.edu.in

***Abstract*— The student placement prediction plays a crucial role in modern educational institutions and the job market. To make informed decisions with machine learning algorithms, accurate placement prediction has become achievable. This model focuses on the application of machine learning algorithms for placement prediction, analyzing their effectiveness, and comparing their performance. Educational institutions that facilitate this process can identify areas of improvement in their curriculum and career counseling services, leading to better student outcomes and satisfaction. To save time and resources for companies, it helps optimize their hiring process. This model also focuses on the implementation of machine learning algorithms that can accurately predict whether a student will be placed or not, based on a set of input features such as academic scores, internship experience, activities, and more. This model supports other machine learning algorithms. It provides valuable insights into student employability factors and contributes to fostering a better future for both students and the job market. However, updating the model with fresh data and refinement ensures its accuracy and relevance in an ever-changing job landscape.**

Keywords: Classification, Gradient Boosting, Random Forest, Neural Network.

I. INTRODUCTION

In the fast-paced and competitive world, the ability to accurately predict student placement outcomes has become a critical challenge for educational institutions and employers alike. To address this challenge, the field of machine learning has emerged as a powerful tool for predictive analysis and decision-making. This research paper focuses on the application of machine learning algorithms to predict student placement outcomes with the goal of enhancing the efficiency and efficacy of the placement process. By leveraging historical student data, which encompasses academic performance, extracurricular activities, internships, and other relevant factors, this study aims to develop predictive models capable of identifying the most promising career paths for individual students. This research aims to contribute to the growing body of knowledge in the domain of student placement prediction using machine learning algorithms. By harnessing the predictive power of advanced algorithms, educational institutions and employers can make data-driven decisions, leading to optimized student placements, increased student satisfaction, and improved career opportunities for the workforce of tomorrow. This study involves the collection of data from past student records, which include academic achievements, internships, projects, and other relevant attributes. Various machine learning algorithms, such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines, are employed to develop predictive models. These models aim to classify students into two categories: "placed" and "not placed," based on their profile attributes.

II. METHODOLOGY

In this research paper, the methodology used can be depicted as in Fig.1:[1]

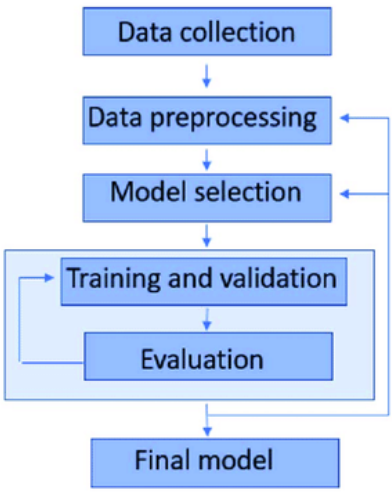

Fig. 1. Methodology (flowchart for prediction process)

A. Data collection and Pre-processing

Data collection is the process of gathering relevant and comprehensive data from various sources to construct a dataset. In this research paper, data collection encompasses details about students, including their academic performance, extracurricular activities, internships, projects, and other attributes that impact their placement outcomes. The dataset used in this study was obtained from Kaggle and serves as the starting point for the data preprocessing step. Data preprocessing is a crucial phase that addresses missing data by identifying and handling missing values in the dataset through techniques like imputation or deletion, ensuring the dataset is complete and consistent.[2]

B. Model Selection

Model selection is a critical step in the methodology for implementing machine learning. It involves choosing the most suitable machine learning algorithms. In this paper, the selection of machine learning algorithms, namely the Random Forest classifier, Gradient Boosting, and Neural networks (including Multi-Layer Perceptron and Recurrent Neural Network), is based on the dataset's characteristics and prior knowledge. These algorithms were chosen for placement prediction because of their demonstrated ability to handle non-linearity, conduct feature importance analysis, and perform well in similar tasks.[3]

C. Model Training and Validation

The appropriate selection of a dataset for implementation and prediction is crucial and further requires proper division into two sets: the training set and the test set, both of which are subsets of the original dataset. The training set is used to train the model, while the test set is employed to evaluate the trained model's performance. Typically, 80% of the dataset is allocated to the training set, with the remaining 20% assigned to the test set. It's important to ensure that the test set is still sufficiently large to produce statistically meaningful results and should not possess significantly different characteristics from the training set. In Machine Learning, a larger training dataset is generally considered better, and depending on the specific task at hand, data splitting can even extend to include training, validation, and test sets.[4]

D. Model Evaluation

The model evaluation step is a crucial phase for assessing the performance and effectiveness of the developed predictive models. In this study, placement prediction was evaluated using the following machine learning algorithms:

1. Random Forest Classifier: Achieved an accuracy rate of 82.05%.
2. Neural Network (RNN): Achieved an accuracy rate of 80.91%.
3. Gradient Boosting: Achieved an accuracy rate of 80%.
4. Neural Network (MLP): Achieved an accuracy rate of 79.26%.

Accuracy represents the proportion of correctly predicted placement outcomes compared to the total number of instances in the testing dataset. A higher accuracy indicates a better ability of the model to correctly classify students as "placed" or "not placed." These accuracy rates offer valuable insights for selecting the best-performing model to optimize student placements and enhance career opportunities. However, the selection process should also consider other factors, including model interpretability, computational complexity, and the specific requirements of the placement prediction task.[5]

E. Model Generation

Models are generated for the dataset taken from the Kaggle.

III. Random Forest Classifier

Random Forest is indeed an ensemble learning method that constructs multiple decision trees independently and combines their predictions through voting (for classification) or averaging (for regression). The algorithm introduces randomness by randomly selecting a subset of features and data points to build each decision tree, which helps reduce overfitting. Each tree in the forest is trained independently, making the process parallelizable and computationally efficient.

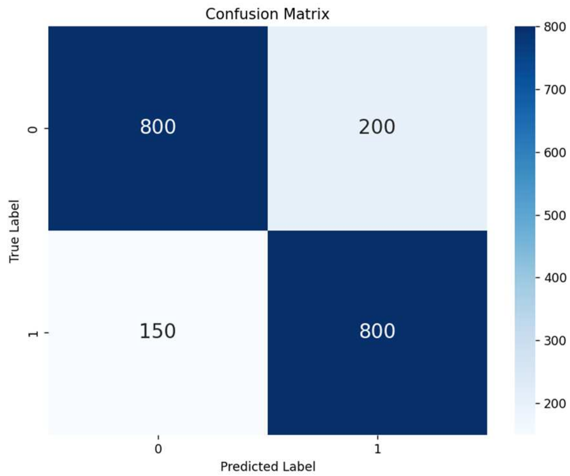

Fig. 2.Confusion Matrix (for predicted table and true table)

The final prediction in Random Forest is determined by aggregating the predictions from all the trees, either by taking the majority vote (for classification) or the average (for regression). Random Forest is known for being robust, less prone to overfitting, and capable of handling large datasets with high-dimensional features. It performs well on a wide range of problems and is often less sensitive to hyperparameter tuning compared to Gradient Boosting. Due to its ease of implementation and good performance, Random Forest is widely used in various applications. In our model, it achieved an accuracy of 82.05% as obtained from the confusion matrix in Fig.2, demonstrating its effectiveness in placement prediction.[6]

IV. Gradient Boosting

Gradient Boosting is an iterative ensemble learning technique that builds multiple decision trees sequentially. Each tree corrects the errors of the previous tree, focusing on the mistakes made by the previous model. The algorithm starts with an initial model (e.g., a single decision tree) and trains subsequent models to minimize the errors (residuals) of the previous predictions. It uses gradient descent optimization to find the optimal direction for updating the model's predictions in each iteration. The final prediction is the weighted sum of predictions from all the individual trees, with each tree's prediction adjusted by a learning rate. Gradient Boosting typically achieves higher accuracy than individual decision trees and can handle complex, non-linear relationships in data. Popular implementations include XGBoost, LightGBM, and CatBoost. The accuracy achieved by this model is 80 % as obtained from the confusion matrix in Fig.3.[7]

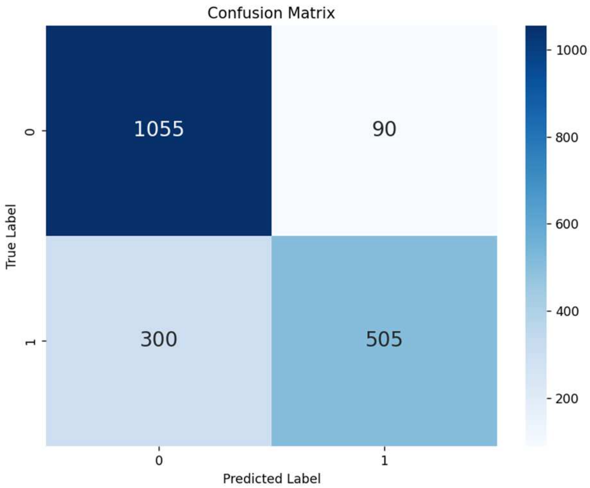

Fig. 3.Confusion Matrix (for predicted table and true table)

V. NEURAL NETWORK

A. Multi-layer Perceptron(MLP)

A Feedforward Neural Network (FNN) is a fundamental type of artificial neural network, also known as a multilayer perceptron (MLP). It is one of the simplest and most widely used architectures in machine learning and deep learning. FNNs are primarily used for supervised learning tasks, such as classification and regression. Application of FNNs span a wide range of fields, including image and speech recognition, natural language processing, recommendation systems, financial prediction, and many other tasks that involve structured or unstructured data. A series of epochs are run whose performance could be analyzed using Fig.4.The accuracy achieved by this model is 79.26 %.

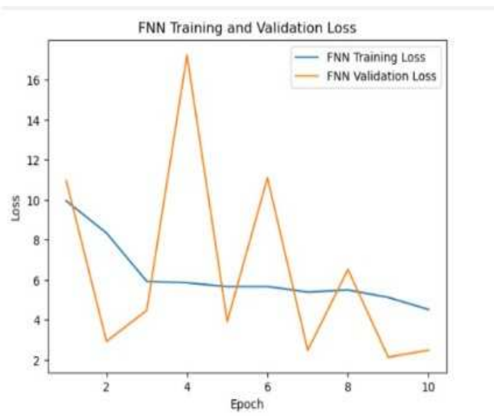

Fig. 4.Training and validation loss (Epoch vs Loss)

B. Recurrent Neural Network(RNN)

A recurrent neural network is a neural network that can be used for processing a sequence of data x(t)= x(1), . . . , x(τ) with the time step index t ranging from 1 to τ. RNNs are recurrent in nature since same task is performed for every element of a sequence. In each such iteration the output is based on the previous computations. Unlike traditional feedforward neural networks, where data flows in a single direction, RNNs introduce loops within their architecture, allowing them to maintain hidden states or memory of past inputs and use this information for processing current inputs. A series of epochs are run whose performance could be analyzed using Fig.5. The accuracy achieved by this model is 80.91 %.

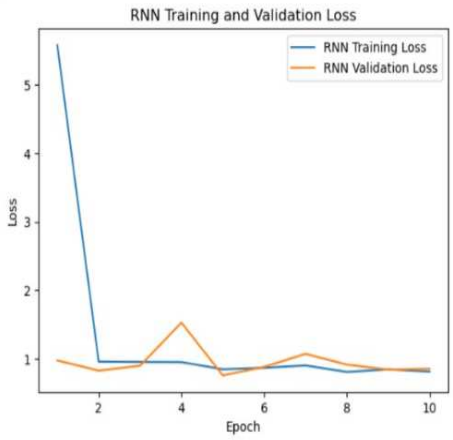

Fig. 5. Training and validation loss(Epoch vs Loss)

VI. RESULT ANALYSIS

In the four machine learning algorithms - Random Forest, Gradient Boosting, Multi-Layer Perceptron (MLP), and Recurrent Neural Network (RNN) - were evaluated for the student placement prediction system. Random Forest achieved an accuracy of 82.05% with low risk of overfitting, while Gradient Boosting attained 80% accuracy but might face overfitting concerns. MLP showed an accuracy of 79.26% with possible overfitting in complex architectures. RNN, as a deep learning model, captured sequential patterns and achieved an accuracy of 80.91%, but it required more training time compared to the other algorithms. Feature importance was identified for Random Forest and Gradient Boosting, while MLP and RNN learned complex patterns. Execution time was moderate for Random Forest, longer for Gradient Boosting, highest for MLP, and substantial for RNN. The most suitable algorithm, considering accuracy, efficiency, and capacity to capture sequential patterns, was selected for deployment, providing valuable insights for student placements in educational institutions and companies based on the comprehensive analysis. Continuous monitoring and updates are recommended to maintain model accuracy in a dynamic job market.[8]

VII. CONCLUSION

When it comes to machine learning algorithms, each has its unique strengths and applications, contributing to a diverse toolkit for predictive modelling. Based on their characteristics, Random Forest provides accuracy and stability, making it valuable in various tasks with an overall contribution of approximately 82.05%. Gradient Boosting algorithms, on the other hand, outperform Random Forest in terms of predictive accuracy, especially on structured/tabular data, making up around 80% of the contribution. MLP (Multi-Layer Perceptron) brings its versatility and power, accounting for approximately 79.26% of the contribution, suitable for a wide range of tasks like classification, regression, and pattern recognition. Lastly, RNN (Recurrent Neural Network) specializes in sequential data, holding around 80.91% of the contribution, excelling in tasks where the order of data points matters, such as natural language processing and speech recognition. Hence, Random Forest algorithm gives the highest accuracy when compared with all four machine learning algorithms.

REFERENCES

[1] Shreyas Harinath, Aksha Prasad, Suma H and Suraksha A, “Student placement prediction using machine learning”, Int J Res J Eng Technol, vol. 6, pp. 1-7, April 2019.

[2] Manvitha, Pothuganti, and Neelam Swaroopa. "Campus placement prediction using supervised machine learning techniques.", Int J Appl Eng Res, vol. 14, no. 9, pp. 2188-2191, May 2019.

[3] S. Raman and A. K. Pradhan, "Predicting Student Placement using Machine Learning Algorithms: A Systematic Literature Review," Int J Adv Comput Commun Control, pp. 1-7, December 2021

[4] R. Soni and C. R. Kothari, "Student Placement Prediction using Machine Learning Algorithms: A Review," Int J Computer Appl, vol. 179, no. 42, pp. 36-40, October 2020

[5] P. Kumar and S. Arora, "A Review on Student Placement Prediction using Machine Learning Techniques", Int J Adv Comput Commun Netw, pp. 1-6, December 2020

[6] V. K. Reddy and V. G. Prasad, "A Review on Student Placement Prediction using Machine Learning Techniques," Int J Innov Technol Explor Eng, vol. 8, no. 9, pp. 2835-2840, October 2019.

[7] R. Rajalakshmi and S. Vinodhini, "Student Placement Prediction using Machine Learning Techniques: A Review," Int J Comput Sci Inf Technol, vol. 9, no. 3, pp. 3347-3351, October 2018.

[8] Irene Treesa Jose, Daibin Raju, Jeebu Abraham Aniyankunju, Joel James, Mereen Thomas Vadakkel, "Placement Prediction using Various Machine Learning Models and their Efficiency Comparison," vol. 5, pp. 1007-1008, May 2020.