IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138
INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG
An International Open Access, Peer-reviewed, Refereed Journal

College Student Placement Prediction Using Machine Learning Classifier

[1]Ashwini Virulkar,[2] Dr. Ashish Sasankar
1Research Scholar, , RTMNU, Nagpur, India
2Principal, Indraprastha New Arts Commerce and Science College, Wardha, India
1 Department of Electronics and Computer Science, RTMNU, Nagpur, India
2Indraprastha New Arts Commerce and Science College, Wardha, India

Abstract: Education is important in every field. It improves the skills of the learners. There are many challenges faced in this field nowadays due to reformations in education system. Once completion of education there are opportunities for the students for Placement. There are many factors responsible for the placement of the students. Factors such as Academic performance, Extracurricular activities, Internships, Economical status of the students etc. are responsible for students. Generally, the students are placed considering their academic performance and their skills. But there is an important factor Intelligent Quotient need to be taken into consideration. This research paper focuses on the different factors and their correlation with each other responsible for the student's placement Using supervised learning algorithm like Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Support Vector Machine. Confusion matrix and correlation coefficient is used to evaluate the results. The dataset used is secondary dataset College student Placement from Kaggle. It consists of 10,000 records. The results show Support vector Machine algorithm performed well as compared to others with 94% accuracy.

IndexTerms - Placement, Intelligent Quotient, Confusion Matrix, Machine learning, Corelation

I. INTRODUCTION

In the educational system the students are assessed and their performance is evaluated based on different metrics. Teachers play an important role in shaping the student's performance for better carrier opportunities. They put an effort for the holistic development of the students. It is necessary to identify the early dropout rate of the student and maximize their performance accordingly. Different activities need to be conducted to improve the performance and make them capable of placement. Students are placed based on their skills. Not merely the skills are important but for their successful career IQ is equally important. The future of student depends on the placement and therefore placement is very important for them [1]. The reputation of the educational institution depends on the students placed and the package offered to them by the recruiters. Identifying the factors is the most challenging job [2]. Machine learning algorithm helps to identify the accuracy of different models for predicting the student's performance and their placement. Today, we follow the outcome-based education. The outcome mainly focuses on the placement rather than the results. Every educational institution strives to improve the placement of the students. The research paper emphases on predicting the placement of college students using ML algorithms considering the parameters like CGPA, Internship, Communication Skill, IQ, Projects etc. The correlation between the factors affecting the placement is analyzed.

II. LITERATURE REVIEW

In the conference Paper [3] by Kajal Rai focused on the machine learning algorithms like Random Forest, Naïve Bayes Classifier and Decision Tree to predict the student's placement on the basis of accuracy gained. The Random Forest performed well with 86% accuracy as compared to other algorithms.

In the conference Paper [4] by Abdulwahhab at. el., emphasis on the learning Analytics (LA) for predicting the students' performance behavior. It uses the digital metrics of learners like activity and student's marks. Compact decision tree (CPT+) has been used to prove the efficiency where accuracy was assessed in comparison with Dependency Graph (DG) and Prediction by Partial_ Matching (PMP). The accuracy of CPT+ is found to be better.

G. M. Spandana in the research paper [5] focuses on different machine learning algorithm logistic regression, Random Forest, K Nearest Neighbor, Support vector machine for predicting student's placement.

The "Comprehensive Career Placement Predictor" is a predictive tool that leverages machine learning to evaluate student information (such as grades, skills, and experience) to forecast their employability and suggest areas for enhancement. These systems facilitate the connection between students' abilities and employer requirements by providing insights into job preparedness, tailored skill recommendations, and alignment with market needs, ultimately improving placement success [6].

The Research article [7] by Ashwin Muralidharan uses an exploratory analysis to predict the student's placement using machine learning classifiers. The classifier used are Logistic Regression, Naïve Bayes classifier and Support Vector Machine classifier. Based on the accuracy the best model has been evaluated [7].

The research paper [8] by Pathak A. at. el., introduces a strong Machine Learning (ML) model for predicting student placement outcomes, using a large publicly available dataset from Kaggle. Algorithms of machine learning are trained and comparison of performances are done which includes Decision Trees (DT), Voting Classifier (VC), Logistic Regression (LR), and other classifiers. Hyperparameter tuning and cross-validation techniques improves predictive performance of model. The Voting Classifier achieved the best results with the highest precision at 94.29%, a competitive accuracy of 89.23%, and a robust F1- score of 0.898. AUC founds to be of 0.943.

The project titled [9] focuses on predicting student placement based on different performance metrics from machine learning algorithms. In this proposed work, a dataset used to predict student placements, with data preprocessing applied to make model training easier for predictions using Decision Tree and XG Boost, along with the traditional established classifiers. The algorithm SVM achieved the highest accuracy with 91%, followed by LR and DT, which both scored 88%. Naïve Bayes had an accuracy of 86%, while XG Boost recorded the lowest at 84%. This analysis helped to determine which algorithm performs better, with those achieving higher accuracy being preferred for predicting student performance.

Ruparel, Milind in the research paper [10] aimed to evaluate how well different machine learning algorithms could predict student placement. A thorough assessment included random forests, decision trees, Naive Bayes, linear discriminant analysis (LDA), gradient boosting, support vector machines (SVM), and k-nearest neighbours (KNN). The research found that K-Nearest Neighbours (KNN) and Support Vector Machines (SVM) both reached an accuracy rate of 94%.

III. PROPOSED METHODOLOGY

A. Dataset

Secondary dataset from Kaggle "College student Placement" is used in this research paper which consists of 10,000 records. The dataset is composed of 10 attributes categorised as semester end results, Extracurricular activities, student centric activities, Intelligent Quotient, Placement. It explores on factors that influence placement outcomes.

B. Data Preparation

The dataset is modified by replacing the values of the attribute internship and placement as yes and no with 0 and 1. The preprocessing is performed to identify the missing values and data validation is done. Our input parameter is the factors affecting the placement and the output parameter is Placement. Correlation of different factors with placement is calculated. Splitting of dataset is performed into training (80%) and testing (20%) i.e., with the ratio of (80:20)

Attributes	Attribute Scale
Student ID	1, 2….10,000
IQ	41-158 (Low - high)
Prev_Sem_Result	5-10 (Low - high)
CGPA	4.54-10.46 (Low - high)
Academic_Performance	1-10 (Low - high)
Internship_Experience	Yes, no
Extra_Curricular_Score1	0-10 (Low - high)
Communication_Skills1	1-10 (Low - high)
Projects_Completed	0-5 (low - high)
Placement	Yes, no

C. Evaluation of Training (Tr) and Testing (Ts)

Preprocessing of the data is accomplished. The data is ready for training(tr) and testing(te) once split. The model is evaluated using the performance measures Accuracy (ACCR), F1- Score (F1-S), Precision (Pr) and Recall (Re). A Confusion Matrix helps to evaluate performance metrics with parameters True_ Positive (TrP), True_ Negative (TrN), False _Positive (FaP) and False_ Negative (FaN).

D. Algorithm

Step1: Import Dataset
Step 2: Perform Data preprocessing along with splitting the data
Step 3: Perform training and testing
Step 4: Model the dataset using LR, NB, KNN, SVM
Step 5: Evaluate the model using confusion Matrix
Step6: Compare the values of Accuracy of different models and analyze it.
Step 7: Find the correlation between different attributes of dataset
Step 8: Analyze the correlation of attributes with respect to placement and record it
Stop9: Stop

E. Flowchart of the Model

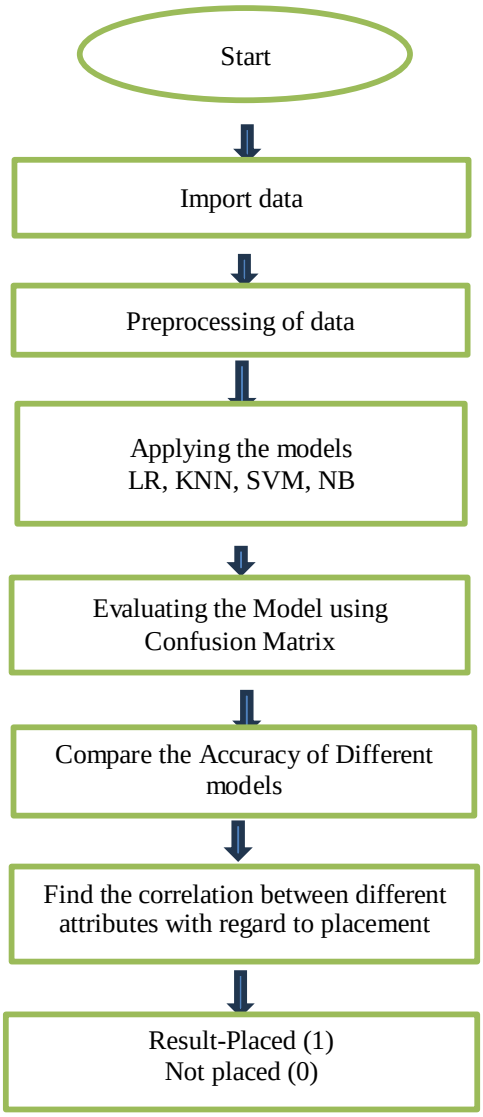


Figure1 : Flowchart of the Model

IV. EXPERIMENTAL METHODOLOGY

A. Machine_ Learning Classifiers

1) ***Logistic_ Regression***

It is the supervised machine learning classifier used for Predicting the probability of binary outcome. It uses the binary values like yes /no or 0/1. It is used for the purpose of classification Problems using Sigmoid function for converting the inputs into the likelihood of binary values.

2) ***Support Vector Machine***

Classification and regression problems both uses Support vector Machine. Basically, the it is used as classifier. The aim of the SVM is the division of data sample into classes to find the hyperplane. The line separating the datapoints is called the hyperplane and the datapoints nearby the hyperplane are support vectors.

3) ***K-Nearest Neighbour***

KNN is classify datapoints based on similarity. It is one of the modest supervised learning techniques. KNN algorithm stores the data set at the training phase and classifies data into a similar category to the new data set.

4) ***Naïve Bayes classifier***

Using Bayes theorem, Naïve Bayes algorithm solves classification problems. It predicts the output on the basis of probability; hence it is also known as probabilistic classifier.

B. Correlation_ Coefficient

It is the statistical Metric to find the strength between the two variables. It describes the association between two continuous variables from -1 to +1. It finds the correlation between predicted and the observed values.
-1 indicates there is a negative association between two variables, i.e., one variable increases other decreases or vice versa. They are inversely proportional to each other.
+1 indicates there is a positive relation within two variables, i.e., one increases/ decreases the other too increase/ decrease. They are directly proportional to each other.
0 indicates there in no relationship between the two variables. Karl Pearson's correlation coefficient is used to measure linear relationship between two continuous variables.

Karl Pearson's Correlation_ Coefficient formula

$$K= \frac{\sum(U-\underline{U})(V-\underline{V})}{\sqrt{\sum(U-\underline{U})^2}\sqrt{\Sigma(V-\underline{V})^2}} \quad \text{----------(1)}$$

Where, $\underline{U}$ = mean of U sample points

$\underline{V}$ = mean of V sample units

C. **Performance_Metrics**

The performance metrics used for evaluation are accuracy(ACCR), precision(Pr), recall(Re), and F1-Score(F1-S). Thes metrics uses the parameters like True_Positive, True_Negative, False_Positive, False_Negative. Assume True_Positive = L, True_Negative = M, False_Positive = N, False_Negative = O. They are calculated as follows

$$Accuracy = \frac{L+M}{L+M+N+O} \text{----------(2)}$$

$$Precision\ (Pr) = \frac{(L)}{(L+N)} \text{----------(3)}$$

$$Recall\ (Re) = \frac{(L)}{(L+O)} \text{----------(4)}$$

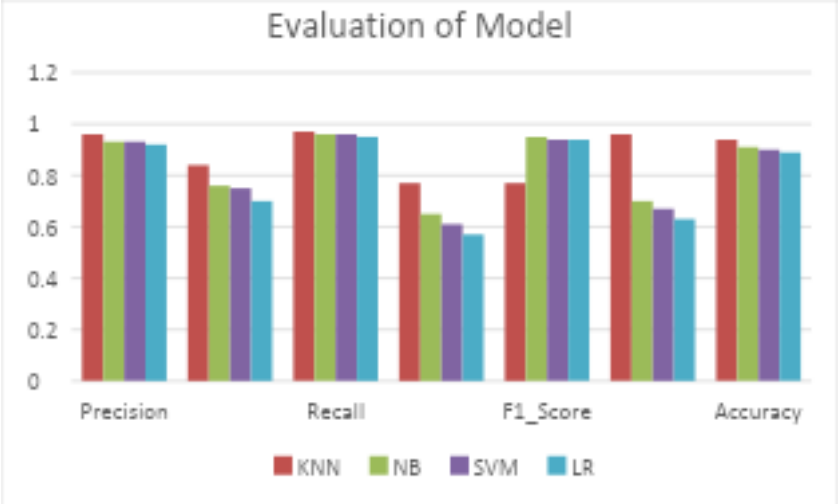
$$F1\text{-}Score = 2 * \frac{Pr*Re}{Pr+Re} \text{----------(5)}$$

V. ***Result Analysis***

The evaluation of the Machine learning models applied on the College Student placement cell dataset is given in the below table. KNN, NB, SVM, LR models are applied on the dataset. Using Confusion Matrix, it gives Accuracy (ACCR), Precision (Pr), Recall (Re)and F1 Score(F1-S).

Table I: Evaluation of model

Classifier	Precision (Pr)		Recall (Re)		F1_Score(F1-S)		Accuracy (ACCR)
	0	1	0	1	0	1	
KNN	0.96	0.84	0.97	0.77	0.96	0.8	0.94
NB	0.93	0.76	0.96	0.65	0.95	0.7	0.91
SVM	0.93	0.75	0.96	0.61	0.94	0.67	0.9
LR	0.92	0.7	0.95	0.57	0.94	0.63	0.89



=

Assessment of Model Evaluation

The table provides a performance assessment of 4 different ML classifiers on a binary classification problem (classes 0 and 1)

It represents that the K-Nearest Neighbours (KNN) classifier generally outperforms the others based on the metrics presented, particularly in terms of overall Accuracy and performance on class 1.

- KNN has the highest Accuracy of 94 %
- The overall performance ranking by Accuracy is: KNN 94% > Naïve Bayes (NB) 91% > Support Vector Machine (SVM) 90% >Logistic Regression (LR) 89%

Class 0 (Presumably the Majority Class)

All models executed very well, with metrics consistently in the range of 0.92 to 0.97

KNN leads in Precision (0.96) and Recall (0.97), and ties for the highest F1-Score (0.96).

Class 1 (Presumably the Minority/Positive Class)

This class is more challenging for all models.

Precision (Class 1): KNN is best (0.84), indicating that for class 1 prediction, 84% of the time its correct. LR is worst (0.70).

Recall (Class 1): KNN is best (0.77). LR is worst (0.57), meaning Logistic Regression misses 43% of the actual class 1 instances (high False Negatives).

F1-Score (Class 1): KNN is clearly best (0.80), providing the best balance between Precision and Recall for the minority class. The F1-Score drops sharply for the other models: NB (0.7) □ SVM (0.67) □ LR (0.63).

Interpretation: The KNN classifier is the most effective model among the four, demonstrating superior overall Accuracy (0.94) and the best performance metrics for the minority class (1), particularly in Recall (0.77) and F1-Score (0.80).

Evaluation of Correlation Matrix

The image displays a Correlation Matrix which assesses the linear relationships between various features and the target variable, Placement. The values range from -1.00 (perfect negative correlation) to +1.00 (perfect positive correlation).

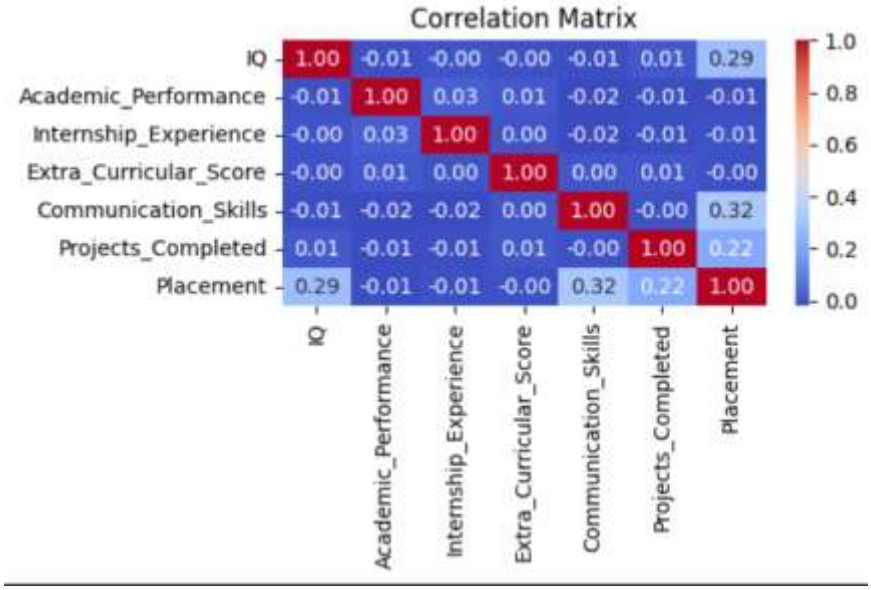


Figure: Correlation Matrix of Student Placement with other factors

Assessment of Correlation Matrix

1. Most Relevant Features: Communication Skills, IQ, and Projects Completed are the most relevant factors, though their impact is minor.

2. Irrelevant Features: The dataset's model development may benefit from considering the low correlation of Academic Performance, Internship Experience, and Extra_Curricular_Score with the target variable; these features contribute very little linear predictive power.

3. Model Benefit: The low inter-feature correlation is beneficial, as it simplifies model interpretation and prevents common issues like multicollinearity.

Interpretation: The most critical part of the matrix is the "Placement", as it indicates which features are linearly related to the final outcome.

- Communication Skills (+0.32): This feature has the strongest positive correlation with Placement. This indicates that better communication skills scores are associated with a slightly higher likelihood of securing a placement.

- IQ (+0.29): This is the second strongest positive correlation, suggesting a slight tendency for higher IQ scores to correspond with higher placement rates.

- Projects Completed (+0.22): This also shows a weak positive relationship with Placement.

- Academic Performance (-0.01), Internship Experience (-0.01), and Extra_Curricular_Score (-0.00): These features exhibit correlations extremely close to zero, meaning they have virtually no linear relationship with the likelihood of Placement.

All features are weak predictors in a linear model, as the maximum correlation is only +0.32. The majority of the features are linearly irrelevant to predicting placement.

VI. CONCLUSION

Placement plays crucial role in every educational institution. The quality of educational institutions nowadays is judged with the number of students placed. The machine learning model K-nearest neighbor performed well with 94% accuracy as compared to Naïve Bayes, Logistic regression and support vector machine. There is high precision, recall and F1-score for KNN classifier with respect to class 0 The Karl Pearsons correlation coefficient gives the relationship between different attributes and it is found that IQ, communication skill and Project completion has positive relationship with respect to placement. Educational institutions should mainly focus on communication skill, include project-based learning, student centric activities as well. IQ play an important role in the students' career which brings better stability and helps to improve the performance.

REFERENCES

[1] S.B, PujhaShree & R, Lekhasree & P, Logasri & Jemima.D, Darling. (2022). STUDENT PLACEMENT PREDICTION. BSSS Journal of Computer. volume 13. 10.51767/jc1303.

[2] Deevi, Sateesh & Siri, Zailan & Rao, D.S. & Anusha, Seetha. (2019). Predicting Student's Campus Placement Probability using Binary Logistic Regression. International Journal of Innovative Technology and Exploring Engineering.ISSN: 2278-3075, Volume-8 Issue-9, July, 2019 8. 2633-2635. 10.35940/ijitee.I8984.078919.

[3] Saraswat, Kajal. (2022). Students Placement Prediction Using Machine Learning Algorithms, South Asia Journal of Multidisciplinary Studies SAJMS June 2022, Vol. 8, No.-5 ISSN:2395-1079 pp 54-60

[4] Abdulwahhab Al Jassim, Rasha & Abdulwahab, Shaqran. (2017). Integrating learning analytics to predict student performance behavior. 1-6. 10.1109/ICTA.2017.8336060.

[5] G. M. Spandana and L. Pallavi, "Placement Prediction System using Machine Learning," *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, Namakkal, India, 2023, pp. 903-907, doi: 10.1109/ICECAA58104.2023.10212409.

[6] M. K. J, V. M, T. P. Rani and S. S. S, "Comprehensive Career Placement Predictor: An Analytical Tool for Optimizing Job Placement Outcomes," *2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS)*, Chennai, India, 2024, pp. 1-6, doi: 1109/ICPECTS62210.2024.10780012.

[7] Ashwin Muralidharan a) , Rijo Benny b) , Keerthivasan Ramalingam Pillai Saravanan c) , Vasanthakumar Mohan d) , Anandh Sivaguru Paramesh e) , Jayashree Kanniappan ,6TH INTERNATIONAL ONFERENCE ON INTELLIGENT COMPUTING: IConIC2K23,28–29 April 2023, Chennai, India Volume 3175, Issue 1, https://doi.org/10.1063/5.0256837

[8] Pathak, A., Matcha, M., Gopisetti, M., Joshi, S. (2025). A machine learning framework for predicting student placement outcomes.International Information and Engineering Technological Associations Ingénierie des Systèmes d'Information, Vol. 30, No. 7, pp. 1715-1721. https://doi.org/10.18280/isi.300704

[9] Preeti Nutipalli1 , Dandupati Priyanka2 , Ladi Yuktha3 , Pollai Prasanth4 , Tankala Ramsai5 , Vakamulla Jagansai6 "Model Construction Using ML for Prediction of Student Placement." International Journal for Research in Applied Science & Engineering Technology (IJRASET), ISSN: 2321-9653,vol. 10, no. VI, 2022, pp. 2213–2219 https://doi.org/10.22214/IJRASET.2022.44273

[10] Ruparel, Milind. (2024). Enhancing Student Placement Predictions with Advanced Machine Learning Techniques. Journal of Information Systems Engineering and Management. 10. 275-288 10.52783/jisem.v10i1s.121.