iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44273

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Model Construction Using ML for Prediction of Student Placement

Preeti Nutipalli[1], Dandupati Priyanka[2], Ladi Yuktha[3], Pollai Prasanth[4], Tankala Ramsai[5], Vakamulla Jagansai[6]
[1]*Assistant Professor, Department of CSE,* [2, 3, 4, 5, 6]*Final Year B.Tech. Students, Aditya Institute of Technology and Management, Tekkali, A.P, India*

Abstract: ***"Model construction using ML for prediction of student placement" aims to predict the placement of a student using various performance metrics on the Machine Learning algorithms. Early prediction makes the institutional growth as well as the student to get placed. It helps the student to prepare all the company requirements at early stage and monitors the student performance. Existed work was done on the algorithms like Logistic Regression (LR), Support Vector Machine (SVM), Naive Bayes. In the proposed work to predict the student placement considered dataset and applied data preprocessing to make the data easier to train the model for prediction using Decision Tree (DT) and XG Boost along with the existing algorithms. Accuracies are calculated using different performance metrics like Accuracy and F1-score, Precision, Recall. The algorithm that worked with the best accuracy is SVM with 91%, and the LR and DT algorithms got 88% accuracy whereas Naïve Bayes got 86% and then the XG Boost stood last with an accuracy of 84%. We are able to make a decision which algorithm is better than other algorithms. Higher accuracy algorithm is mostly preferred to predict the student performance.***
Keywords: Machine Learning, classification Algorithms, Performance Metrics

I. INTRODUCTION

Placements play a crucial role in every educational institution. College reputation is based on students' placement rate. The main goal of institutes is to get their students placed with a better offer. Every educational institution is mainly focused on student placements. Educational institutions are fulfilling student dreams using placements. A placement prediction system can be used to determine the capability of a student for that specific job role. Generally, placement prediction is favorable to students and educational institutions. It will be helpful to the students to get placed early. So that students can prepare based on their company requirements and improve their overall development of a student. A high placement rate is an important factor in establishing institutional standards. It will be helpful to the institutions in the form of admissions and also faculty can look after the students by giving better training. The college contains a large number of student records; it is very difficult to get the specific characteristics when we do manual prediction. Obtaining placement status from these tasks is a tedious task. Manual prediction needs lots of human resources to maintain student records. All these manual implementations bring barriers to educational institutions. With the machine learning algorithms, we overcome the problem in the manual process. Placement prediction system helps in effective filtering of students by considering various factors like CGPA, technical skills, soft skills, coding skills, communication skills.

Using different Machine Learning (ML) algorithms, predict the probability of student placement performance. In this we used different machine learning algorithms like support vector machine (SVM), Logistic Regression (LR), Naive Bayes, XG Boost, Decision Tree to predict the student performance. These classifiers independently predict the results and then compare the accuracy of the algorithms, on the data set. The performance analysis on different metrics like accuracy, precision, F1-score, we able to find out the algorithm which performs better than other on the student data set, so that we make a guideline for future improvement of student placement performance in educational institutions. Machine Learning enables computer intelligent by supplying data to algorithms and constructs forecasting models. Machine learning algorithms estimate new output values using historical data as input. A Machine Learning system trains from past data, constructs the forecast models, and on getting new data, predicts the output for it. The accuracy of expected output is dependent on the amount of data, large amounts of data aid in more precise forecasting.

If we have a complex situation for which we need to make predictions, rather than writing code for it, we may just input the data to generic algorithms, and the machine will develop the logic based on the data and forecast the outcome. Machine learning has shifted our perspective on the issue.Some Aspects of machine learning makes use of data to find patterns in a dataset. It can learn from previous data and improve on its own and it is a data-driven system.

Data mining and machine learning are very similar in that they both deal with large amounts of data.

A. Classification of Machine Learning:

1) *Supervised Machine Learning:* Supervised learning is a type of machine learning in which machines are trained using well-labelled training data and then predict the output using that data. Some of the input data has already been tagged with the correct output. In supervised learning, the training data provided to the machines acts as a supervisor, instructing the machines on how to correctly predict the output. It uses the same concept as when a student learns under the guidance of a teacher. The process of providing input data and correct output data to the machine learning model is known as supervised learning. A supervised learning algorithm's goal is to find a mapping function that will map the input variable(x) to the output variable(y). Supervised learning can be used in the real world for risk assessment, image classification, fraud detection, spam filtering, and so on.
2) *Unsupervised Learning:* The use of artificial intelligence (AI) systems to find patterns in data sets including data points that are neither categorized nor labelled is known as unsupervised learning. Even if no categories are specified, an AI system will categorize unsorted data according to similarities and differences in unsupervised learning. unsupervised learning can be more unpredictable than a supervised learning model. Although, unsupervised learning can be more unpredictable compared with other natural learning methods. Unsupervised learning algorithms include clustering, anomaly detection, neural networks, etc.
3) *Reinforcement:* Reinforcement learning is a type of machine learning method in which an intelligent agent interacts with its surroundings and learns how to act within it. Reinforcement learning is a machine learning training method that rewards desired behaviors while punishing undesirable ones.

B. Life Cycle of Machine Learning

Models are trained using a labelled dataset in supervised learning, where the model learns about each type of data. The model is tested on test data after the training process is completed, and it then predicts the output.

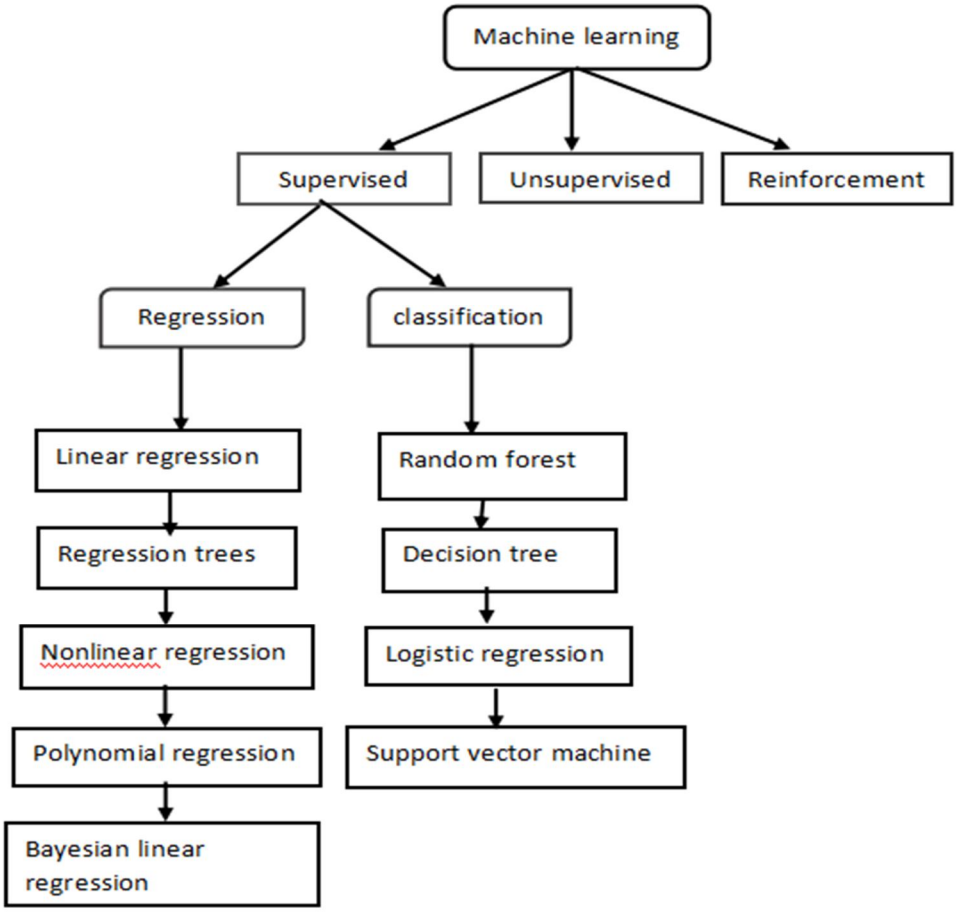


Fig 1: Machine Learning Taxanomy

II. RELATED WORK

Sultana et al.[1] worked on training and modelling of a data set and created a better prediction model with better accuracy rate. They worked on ConvolutionalNeuralNetwork (CNN). They proved that CNN achieves a better accuracy than other traditional methods of accuracy 97.55%. Mondal et al.[2] used Recurrent Neural Network (RNN) to predict the students performance. Their analysis made the faculty to identify the students who are in danger zone and based on that they can give better solution to their problem. They also made the comparison among Artificial Neural Network and Deep Neural Network of accuracy 85.4%. Ramesh et al.[3] presents an approach by using Deep Learning algorithms to create a new placement prediction method for all graduates. Manvitha et al.[4] used to compare their proposed method with some of the traditional classification algorithms like Random forest, Decision tree with respect to precision, accuracy. They also proved that their proposed method performs better in comparison with other traditional classification algorithms for decision tree accuracy is 84% and for random forest accuracy is 86%. Kavipriya et al.[5] worked on new classifier called as Adaptive weight Deep Conventional Neural Network(AWDCNN) which is optimized with the help of Genetic algorithm to predict the student performance. They proved that proposed classifier AWDCNN produces better results than existing prediction methods with the help of training data set of accuracy 92.41%.D. Satish Kumar et al.[6] have conducted the Study on placement prediction they have randomly taken a 250 MBA students records from 5 leading institutions and six-predictor logistic regression model was used for this data they have taken various parameters like CGPA in UG and PG, Specialization in UG and PG, Soft skill score and gender and they have diagnostic the record before applying of Logistic Regression and they measured their accuracy by using Confusion matrix and finally came up with a conclusion Hosmer-Lemeshow Goodness-of-Fit test conforms that the model is best fit for their dataset and based on ROC model it gets accuracy 60%. Irene Treesa Jose et al.[7] have studied the student placement prediction they are used several Machine learning (ML) techniques such as K-nearest neighbor (KNN) , Support Vector Machine (SVM), Logistic Regression(LR), Random Forest(RF) and they tested the performance accuracy of various machine learning algorithms and try to compare the accuracy of machine learning algorithms. They have taken a dataset which consists of various parameters like Quantitative scores, Logical Reasoning scores, Verbal scores, Programming scores, CGPA, No. of hackathons attended, No. of certifications, current backlogs they have test on these parameters and they have mentioned about the detailed description of each of these algorithms. Shreyas Harinath et al.[8] have conduct the study of placement status of the student's prediction is done by using Naïve Bayes, K-nearest neighbor they used the past data of scholars and used to train the model for rule identification and testing the model for classification. They consider the parameters like USN, Tenth and PUC/Diploma results , CGPA, Technical and Aptitude Skills for their prediction. The Naïve Bayes classifier is very effective on many real data applications and the KNN is work on similarity measures and the accuracy of KNN is 95.18%, for Logistic regression 97.59%, for Random forest 96.38%, and for SVM accuracy is 100%. Krishnanshu Agarwal et al.[9] had implemented various data mining techniques for student placement prediction and they have used algorithms like K-nearest neighbor(KNN) which is used to labeled points to learn to label other points and random forest(RF) which works by creating a group of random uncorrelated decision trees(DT) and increase the accuracy by taking grade point average(GPA), cumulative grade point average(CGPA) as in their dataset from final year students of B.tech from Kalinga Institute of Industrial Technology(KIIT). Their analysis gives that KNN gives 93.54% and random forest gives the 83.87% accuracy and they have concluded that KNN gives more accuracy for their dataset. Abhishek S. Rao et al.[10] have conducted a detail study of placement prediction and Educational data mining (EDM) in the field of machine learning and data mining to analyze the data for prediction the data is collected from different institutions and they used different factors like CGPA attained, certifications courses completed and they used machine learning algorithms like KNN(K-nearest neighbor), SVM(support vector machine), and ANN(Artificial neural network) and they have measures the performance metrics such as accuracy precision, sensitivity, F1-score and Area under ROC curve(AUC).

III. METHODOLOGY

A. *Implementation of Machine Learning algorithms*

1) *Logistic Regression*

Logistic regression algorithm is used to predict the dependent variable on a given set of independent variables. log(a/1-a) is the link function. It predicts the dependent variable in a categorical form (in the form of binary 0 or 1, or yes/no).

$$\log\left[\frac{a}{1-a}\right] = Y \quad (1)$$

The generalized linear model equation is:

$$g1\big(E1(y1)\big) = \alpha 1 + \beta x1 + \gamma x2 \quad (2)$$

Here, g1() is the link function, E1(y1) is the target variable's expectation and α1+ βx1 + γx2 is the linear predictor (α1, β, γ to be predicted) is the linear predictor (α1, β, γ to be predicted).

2) Naive Bayes Classier - Gaussian Naive Bayes

Bayes' Theorem also called as Bayes' law or Bayes' rule. It says the probability of the occurrence of an event based on the knowledge it has on that event. Mathematical equation for Bayes' Theorem is:

$$P\left(\frac{S}{T}\right) = P\left(\frac{T}{S}\right) \cdot \frac{P(S)}{P(T)} \quad (3)$$

where S and T are events and P(T) ≠ 0. P(S|T) is called as the Posterior probability which means Probability of hypothesis S on the observed event T. P(T|S) is called as the Likelihood probability which means Probability of the evidence given that the probability of a hypothesis is true. P(S) is called as the Prior Probability which means Probability of hypothesis before observing the evidence. P(T) is called as the Marginal Probability which means Probability of Evidence.

3) XG Boost

XGBoost stands for Extreme Gradient Boosting. It is a gradient boosted ML library used to find accurate model based on the data.

$$G2(X) = \sigma(0 + 1 * f1(X) + 1 * f2(X))) \quad (4)$$

where G2(x) value is taken as the prediction from Boost model. G0, the initial model is defined to predict the target variable y. A new model f1 is fit to the residuals from the previous step. Now, G0 and f1 are combined to give G1, the boosted version of G0. The mean squared error from G1 will be lower than that from G0. Now, the deep workings of XGBoost is given below.

$$G1(X) < -G0(X) + f1(X) \quad (5)$$

To improve G1 performance, we could model after the residuals of G1 and create a new model

$$G2(X) < -G1(X) + f2(X) \quad (6)$$

Let this be done for 'n' iterations, until residuals have been decreased as much as possible:

$$Gn(X) < -Gn-1(X) + fm(X) \quad (7)$$

$$G1(X) < -G0(X) + f1(X) \quad (8)$$

4) Decision Tree

Decision Tree is a widely used classification as well as Regression problem. It is a Supervised learning technique**. In** decision tree internal nodes represent the features of a dataset, branches depict the decision rules and each leaf node says the outcome**.**

$$\text{Entropy= } En(S1) = -p1_{(+)} \log p\,1_{(+)} - p1_{(-)} \log p\,1_{(-)} \quad (9)$$

where $p1_+$ is the probability of positive class, $_{p1-}$ is the probability of negative class, S1 is the subset of the training example.

5) Information Gain

Information gain nothing but the reduction of uncertainty and it also says which attribute should be selected.

$$Info.\ Gain = En(Y1) - En(Y1|X1|) \quad (10)$$

B. Proposed Work

Here, in the proposed work to predict the student placement considered dataset and applied data pre-processing to make the data easier to train the model for prediction.

1) Dataset

Data set is taken from the website https://www.kaggle.com/benroshan/factors-affecting-campus-placement. Data on total contains 215 student records with 15 features with target label.

Attribute details

Attribute	Description	Data type
Sl.no	Serial number	integer
gender	Gender of student	String
ssc_p	Percentage in intermediate	Double
ssc_b	Board of intermediate	string
hsc_p	Percentage of high school	Double

hsc_b	Board of high school	string
hsc_s	Stream in high school	string
degree_p	Percentage in degree	Double
degree_t	Stream in degree	string
workex	Work experience	string
etest_p	Percentage in etest	Double
specialization	Specialisation in mba	string
mba_p	Percentage in mba	Double
status	Status of student (placed/not placed)	string
salary	Salary offered	Integer

Table 1: Dataset for student placement prediction

2) Data Preprocessing

Data pre-processing is a very important step, which works on the meaningless data and converts into clean data, and then trains the Machine Learning algorithms. The below mentioned are the basic steps involved in the Data Pre-Processing.

Data collection which discussed in methodology section 3.

a) *Import all the Required Libraries:* For the language we used, if it need to identify any function we worked, we need to import all the required libraries into it. All the necessary libraries need to be imported into the working environment like Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn.

b) *Data set Importing:* The collected dataset which is in CSV format, need to be imported into the workspace. As the columns of sl. No and salary are not required to predict student placement status, those two columns are dropped.

c) *Dealing with the Missing Values:* Missing values are those whose data is not present in the respective column. Dataset with more number of missing values may lead to less computational power. The dataset need to be checked for any missing values, if found then it need to be replaced by considering mean and mode of the attribute or may be either by deleting entire row or column if more number of missing values are found. In the considered dataset, found no missing values and hence not required dealing with missing values in preprocessing steps.

d) *Label Encoding the Data:* Label encoding is mainly done to make sure that all the data is in numerical format**.** The categories in the ‘hsc_s’ and ‘degree_t’ are splitted into individual columns on applying the dummy encoding. It uses ‘1’indicating ‘YES’ and ‘0’ indicating ‘NO’. Here, the number of newly created columns equals to the number of categories. The following columns ['gender','ssc_b', 'hsc_b','workex','specialisation','status'] are also converted to ‘0’ and ‘1’s. Outlier is an object which completely differs from the rest of the objects in the data. They cause problem for the statistical result. Outliers are checked for the data and some outliers from the columns ‘hsc_p’ and ‘degree_p’ are removed to make it more fit for the statistical analysis. Correlation is mainly checked to know the relationship between the variables. ‘+1’ indicates that if one variable is increasing simultaneously the other is also increasing. ‘-1’indicates that if one variable is increasing the other is decreasing. ‘0’ indicates that correlation is not present between the variables. Two variables x and y are considered which indicates independent and dependent variables respectively.

e) *Dataset Splitting:* Data set need to be divided into two sections, Training and Testing data. Training data is the major part of the dataset which is used to feed the Machine learning model to recognise the patterns. Testing data is data which is used to compute the accurate result of the model. Data can be splitted in the ratio of 70:30, 60:40, 80:20. The dataset is divided into 80:20 ratio of Training and testing respectively.

f) *Scaling the Features:* It is the last step in any preprocessing. It is used on independent variable to limit their range using the fit transform function, so that comparison becomes easy. Standardization method is used on the dataset to limit their features.

Formula for standardization:

$$x` = \frac{(x - mean(x))}{sd} \quad (11)$$

Where, x` indicates new value got, x indicates actual value, Sd indicates standard deviation.

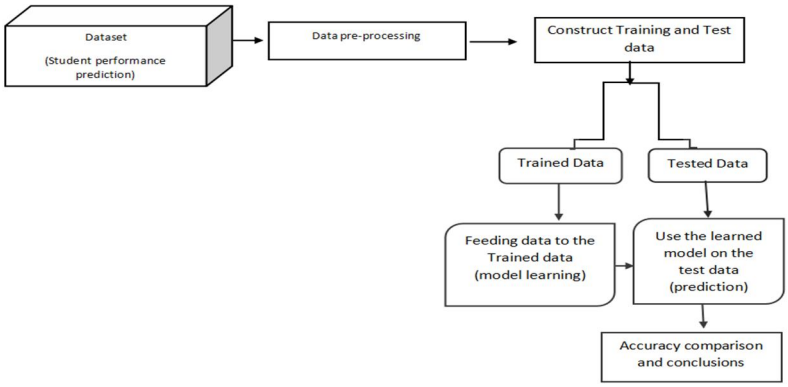


Fig 2: System Architecture for student placement prediction

IV. EXPERIMENTAL SETUP

The hardware configurations used are 8.00GB RAM, 64 Bit Operating system, core i5 intel processor. The software configurations used include Jupyter Notebook. Python 3.6 version is used, which is a high-level programming language that emphasizes code readability. Most of the python machine learning libraries like Pandas (to analyze the data), NumPy (to do operations on arrays), Matplotlib (used to plot graph), Seaborn (to draw statistical graphs), Scikit-learn (provides efficient tools) are used.

A. Result and Analysis

The overall analysis made us to differentiate which factors helped directly and indirectly in predicting the placement of the student and also the comparative study on the various Machine learning algorithms made us to know the most accurate algorithm which works on the student placement data.

To result out the performance of classification we are required to consider the parameters such as Precision(P), Recall (R), Accuracy (Acc) and F1-score (F1_S) which confirms whether the classification is good or bad on the dataset.

The classification metrics purely depends on parameters of True Positive (TP), False Positive (FP), False Negative (FN), True Negative (TN) of confusion matrix. The performance metrics are calculated as follows:

Let X=True Positive, Y=True Negative, S=False Positive, T=False Negative

Recall (Rec)

$$\boldsymbol{Rec} = \frac{\mathbf{X}}{\mathbf{X+T}} \tag{12}$$

Accuracy

$$Accuracy = \frac{(X+Y)}{(X+Y+S+T)} = \frac{(No.of.Coorect\ prediction)}{Total\ no\ of\ predictions} \tag{13}$$

Precision (Pre)

$$\Pr e\ cision = \frac{(X)}{(X+S)} = \frac{(No.of.Coorectly\ predicted\ positives)}{Total\ no\ of\ positive\ predictions} \tag{14}$$

F1-Score Formula

$$F1\_Score = 2 \cdot \Pr e \cdot \frac{rec}{\Pr e+rec} \tag{15}$$

The observation table shows the data of precision, accuracy, recall and f1-score given by various machine learning algorithms. The SVM algorithm gave the best result with precision 92%, recall 92%, accuracy 91% and f1-score 92%.

	SVM	NAIVE BAYES	LOGISTIC REGRESSION	DECISION TREE	XgBoost
Precision	92%	86%	89%	92%	88%
Recall	92%	92%	92%	89%	85%
Accuracy	91%	86%	88%	88%	84%
F1-Score	92%	89%	91%	90%	86%

Table 2. performance comparison of ML Models

The Logistic regression algorithm worked with an accuracy of 88%, precision 89%, recall 92%, f1-score 91%. The Naïve Bayes algorithm worked with an accuracy of 86%, precision 86%, recall 92% and f1-score of 89%, whereas the decision tree algorithm worked with a precision of 92%, recall 89%, accuracy 88% and f1-score of 90%. Among all these algorithms, the XG Boost algorithm worked least with that of 84% accuracy, 88% precision, 85% recall and 86% f1-score.

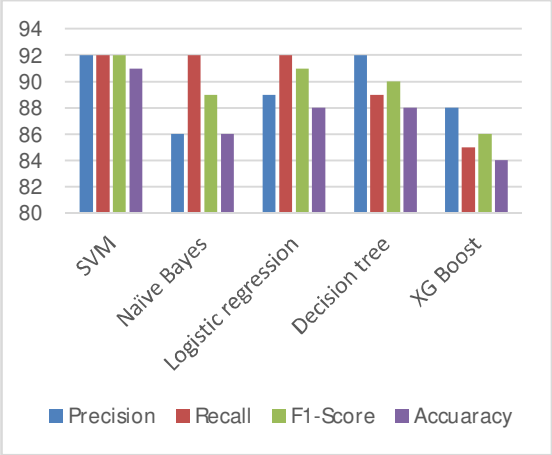


Fig 3: performance metrics graph on Different ML models

V. CONCLUSION

Any institute or university in the world mainly focus on a student's academic achievement. It would be essential to use a variety of methodologies to accurately predict the student's performance. Predicting performance would also allow institutions to focus more on students who are more likely to perform poorly in order to improve further. Predicting the student performance is a dual beneficial for both students and educational institutions as students for getting early placements through pre-hand coaching and for educational institutions to make them stand better at every placement aspect with higher accuracy.

Different machine learning techniques are used namely SupportVectorMachine, NaiveBayes, LogisticRegression, Decision Tree and XGBoost and evaluated perform analysis on different metrics like accuracy, precision, recall, F1-score. Support vector machine (SVM) achieves a better accuracy (91%) than remaining other ML Models.

REFERENCES

[1] Sultana, Jabeen, M. Usha Rani, and M. A. H. Farquad. "Student's performance prediction using deep learning and data mining methods." Int. J. Recent Technol. Eng 8.1S4 (2019): 1018-1021.

[2] Mondal, Arindam, and Joydeep Mukherjee. "An Approach to predict a student's academic performance using Recurrent Neural Network (RNN)." Int. J. Comput. Appl 181.6 (2018): 1-5.

[3] Ramesh, V., P. Parkavi, and P. Yasodha. "Performance analysis of data mining techniques for placement chance prediction." International Journal of Scientific & Engineering Research 2.8 (2011) .

[4] Manvitha, Pothuganti, and Neelam Swaroopa. "Campus placement prediction using supervised machine learning techniques." International Journal of Applied Engineering Research 14.9 (2019): 2188-2191 .

[5] Kavipriya, T., and M. Sengaliappan. "Adaptive Weight Deep Convolutional Neural Network (AWDCNN) Classifier for Predicting Student's Performance in Job Placement Process." Annals of the Romanian Society for Cell Biology (2021): 5494-5590 .

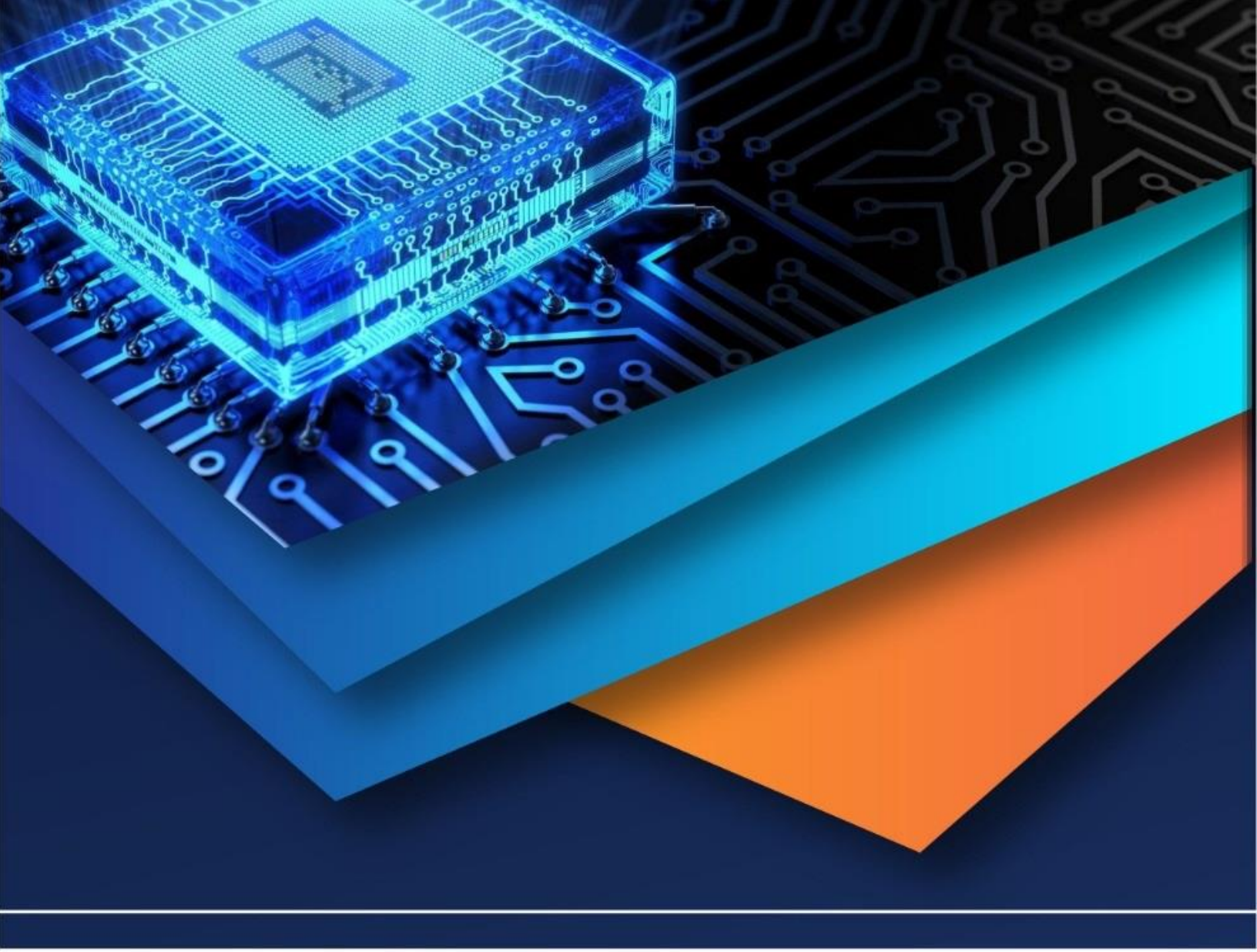
[6] Kumar, D.S., Siri, Z., Rao, D.S. and Anusha, S., 2019. Predicting student's campus placement probability using binary logistic regression. International Journal of Innovative Technology and Exploring Engineering, 8(9), pp.2633-2635.

[7] Jose, I.T., Raju, D., Aniyankunju, J.A., James, J. and Vadakkel, M.T., Placement Prediction using Various Machine Learning Models and their Efficiency Comparison.

[8] Harinath, S., Prasad, A., Suma, H.S., Suraksha, A. and Mathew, T., 2019. Student placement prediction using machine learning. Int. Res. J. Eng. Technol, 6(4), pp.4577-4579.

[9] Agarwal, K., Maheshwari, E., Roy, C., Pandey, M. and Rautray, S.S., 2019. Analyzing student performance in engineering placement using data mining. In Proceedings of International Conference on Computational Intelligence and Data Engineering (pp. 171-181). Springer, Singapore.

[10] Rao, A.S., Aruna Kumar, S.V., Jogi, P., Chinthan Bhat, K., Kuladeep Kumar, B. and Gouda, P., 2019. Student placement prediction model: a data mining perspective for outcome-based education system. International Journal of Recent Technology and Engineering (IJRTE), 8, pp.2497-2507.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY