iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** IV **Month of publication:** April 2024

DOI: https://doi.org/10.22214/ijraset.2024.59731

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 12 Issue III Mar 2024- Available at www.ijraset.com

Academic Performance Prediction System for Polytechnic Students Based on Machine Learning Algorithms

Sandhya S[1], Arya J Nair[2], Sreejakumari S[3]
[1, 2, 3]*Dept. of Computer Engineering, NSS Polytechnic College, Pandalam, Kerala, India*

Abstract: ***Recent years have seen a rise in the use of machine learning techniques for creating predictive models to evaluate academic performance. This academic performance prediction method makes use of student data, including demographics, academic records, and socioeconomic background, to forecast future academic performance outcomes. An academic prediction system for Polytechnic Students based on machine learning model that predicts whether the student pass/ fail in the examination is introduced in this paper. A comparative analysis of eight prediction models and factors affecting the performance of students has been analyzed. Employing a performance prediction system can help teachers promote student achievement, allocate resources more effectively, and incorporate data-driven methods into their instruction.***
Keywords: Machine learning (ML), K-Nearest Neighbors (KNN), Random Forest Classifier, Support Vector Classifier (SVC), Gaussian Naïve Bayes Classifier, Deep Neural Network, Iterative Dichotomiser, (ID3)algorithm

I. INTRODUCTION

The degree of success a learner exhibits in their academic pursuits is referred to as their academic performance. Educational institutions can improve the academic performance and general success of their students by identifying and serving the requirements of struggling students earlier[1] . To promote an inviting learning atmosphere and guarantee that every student has the chance to realize their maximum potential, it is imperative to put measures for observing and helping these children into action. Additionally, educational establishments must pinpoint the elements that could affect students' academic achievement. These variables may include a broad variety of factors like family background, pressure, health, and well-being, teacher's support, etc. By comprehending and tackling these factors, educational establishments may establish a more welcoming and encouraging atmosphere that caters to the varied requirements of learners.

Students with varying economic backgrounds (mostly medium and low level), learning preferences, and professional goals are frequently enrolled in polytechnic colleges. Students' academic performance may vary as a result of this diversity. Students who are unable to enroll directly in a B.Tech program due to financial limitations typically join polytechnic colleges in Kerala[2]. Typically, polytechnic schools emphasize technical skills and practical knowledge pertinent to certain sectors through hands-on, practical learning experiences. Effective educational experiences for polytechnic students are enhanced by engaging teaching strategies, supportive instructors, and well-equipped infrastructure. Here, different machine learning models are used for predicting whether a student will pass or fail in a polytechnic college. These models are cross-validated to find the best one. The factors affecting the pass or fail status of students are analyzed and correlations between the variables are also presented in this paper.

II. REVIEW ON EXISTING APPROACHES

The decision tree algorithm has been used for predicting student's academic performance in [3]. Student data were gathered from the University of Nigeria, Nsukka first-year computer science students enrolled in two courses, STA172 and Cos101. The decision tree analysis for STA172 provided a prediction accuracy of 71.91%, while the study for COS101 produced a prediction accuracy of 96.7%. It is demonstrated that the precision depends on the datasets used to train the model. To develop predictions based on students' changing performance states, a two-layered framework consisting of many base predictors and a cascade of ensemble predictors has been presented in [4]. The University of California, Los Angeles' dataset of undergraduate students was gathered over three years.

Machine learning-based prediction models for analyzing the performance of Chinese University students have been introduced in [5]. Four prediction models were trained and tested, and SVC was found to be the best classifier among the four in performance prediction. The major aspects of the questionnaire were determined by analyzing its contents using the chi-square test. The system provided 80.9% accuracy. Student's performance and pass rate have been determined using a Decision tree and SVM classifier in [6]. The accuracy of the prediction model was enhanced by optimizing DT and SVM using the grid search algorithm. A ten-fold cross-validation has been done to DT and SVM to find the best prediction model. Support Vector Machine (SVM), Random Forest (RF), Neural Networks (NN), and Generative Adversarial Networks (GANs) have been used for student's academic performance analysis in [7]. This research maximized the effectiveness of predictive algorithms by using a five-level grade system. Two Portuguese secondary schools provided the dataset. This study demonstrates how well a regression strategy can forecast academic achievement when compared to a classification approach. The accuracy of forecasting student achievement on the 5-level grading system is greatly increased when the classification strategy is replaced with a regression approach. The quantity and diversity of the dataset are increased by adding synthetic instances produced by statistical methods or machine learning algorithms to the already-existing data.

III. PROPOSED SYSTEM

A machine learning-based system for predicting whether polytechnic students will pass or fail the course has been proposed in this paper.Fig.1 shows the block diagram of proposed system. Their academic performance and the factors affecting their performance have also been studied. The system mainly consists of 3 stages

- Dataset preparation
- Feature Selection and Classification
- Prediction Model training and testing

1) *Dataset Preparation:* The dataset is gathered and ready for analysis in this first phase. To offer pertinent data for analysis, students are asked to complete Google Forms throughout the dataset preparation phase. An effective and user-friendly method for gathering structured data from participants is Google Forms. The form guides students through entering grades, history of attendance, personal information, and other details that can affect their academic achievement. The information gathered using Google Forms is automatically combined into a dataset after it is entered. After that, this dataset is cleaned, processed, and made ready for feature selection and categorization. By making sure the dataset is well-organized and prepared for additional processing, this step establishes the groundwork for the next phases. Missing values are added and error values are corrected.
2) *Feature Selection and Classification:* Feature selection strategies, which identify the most important characteristics that contribute to the predicting job, are used in the second stage. This lessens complexity and concentrates attention on the data's most valuable features. Classification algorithms are employed to create predictive models based on the features that were found to be relevant.

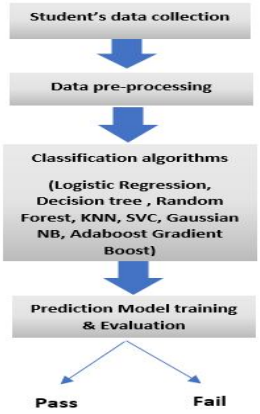


Fig.1 Proposed system

Eight classification algorithms have been cross-validated for finding the best prediction model. They are Logistic Regression, Decision tree, Random Forest, KNN, SVC, Gaussian NB, Adaboost, and Gradient Boost algorithm. Classification algorithms group occurrences into preset categories, which in this case is the pass or fail status of students.

3) *Logistic Regression:* As a popular statistical technique for predicting binary outcomes, logistic regression can be applied to the prediction of academic performance in scenarios where the desired outcome is binary, such as pass/fail or high/low academic performance. A student's likelihood of falling into one of the binary result classes is calculated by the logistic regression model using the value of the variable that predicts the outcome [8]. It seeks to represent the linear relationship that exists between the dependent and the independent variable.
4) *Decision Tree:* The Decision Tree method assesses several characteristics, or features, during the training phase to choose the most appropriate feature for dividing the data variables. As the training data is divided into subsets according to the feature values, decision trees encapsulate the decision criteria that are acquired from the data. Attribute selection measure is a metric that the DT algorithm employs to assess what features are most suited for partitioning a dataset. DT classifiers are suitable for empirical research since they can be constructed without any domain knowledge or parameter settings [9].
5) *KNN Classifier:* A new data input is compared to the values in a specified data collection using the K-NN algorithm. The distance that exists between each occurrence in the specified data collection and the freshly acquired input is calculated by the algorithm. The two prevalent distance measurements used are Euclidean distance and cosine similarity. Next, the data is sorted according to the distance score in ascending order. Then 'K' closest neighbors of the new item based on the calculated distances have been found out. Next, the majority class in the closest neighbors has been allocated to the newly entered data [10].
6) *Random Forest:* The outcomes of several decision trees is combined by a random forest to get a single output. The feature significance scores produced by Random Forest models can be used to determine which academic performance factors are most important. Random forests merely choose a portion of those attributes, whereas decision trees take into account all potential feature splits. The three primary hyperparameters of RF methods are the dimensions of the node, the total number of trees, and the number of features evaluated [11].
7) **SVC:** SVC is a particular use of the SVM algorithm created with classification problems in mind. It is based on optimization and linear algebraic mathematics. Predicting academic performance is a binary classification task where the SVC finds a hyperplane that divides the data values into two groups (pass/fail). More accurate categorization results from a greater separation between the hyperplane and the closest data points from each group. [https://www.linkedin.com/pulse/svc-support-vector-classifier-dishant-salunke [12].
8) *Gaussian NB:* The Gaussian Naive Bayes algorithm relies on a probabilistic methodology and considers that every class has a normal distribution. An ultimate prediction is the sum of predictions made for each variable to determine whether the dependent variable will be categorized into each group. The final classification is given to the group that has the highest probability. Although there are different algorithms for estimating data distribution, the Gaussian distribution is the easiest to utilize because it only requires the training data's mean and standard deviation [13].
9) *Adaboost Algorithm:* AdaBoost is a boosting technique that operates on the same idea as the stage-wise addition technique, which is to obtain strong learners by using numerous weak learners. The weak classifier is trained continuously on the training dataset, with every new classifier assigning greater weight to the incorrectly categorized data values. With this method, boosting can be accomplished at a better precision than with any one base model alone [14].
10) *Gradient Boost Algorithm:* A potent machine learning method called gradient boosting frequently produces cutting-edge results on various tabular data issues. They are the resulting procedure when a decision tree exhibits weak learner behavior. Loss function, Weak learners, and Additive model are the key components of the Gradient Boosting Algorithm [15].
11) *Model training and validation:* The prediction models are cross-validated and find out the best model. The selected prediction model is trained using pre-processed data. The model modifies its parameters during training to lower the prediction error.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

IV. EXPERIMENTAL RESULTS

A questionnaire has been created to forecast polytechnic students' academic performance, taking into account both personal information and several variables that could affect their educational outcomes. A Google form consisting of 25 questions was given to students of NSS Polytechnic College, Pandalam. The dataset comprised responses from 135 students, forming the basis for subsequent data analysis. The dataset has been used for training 8 machine learning-based prediction models. The prediction models are designed to forecast whether students will pass or fail their final Diploma examination. This is a binary classification task. The models have been cross-validated and the best one among 8 models has been identified. After evaluating various models, it was determined that the Support Vector Classifier (SVC) model yielded the highest cross-validation score, indicating its superior performance in predicting student outcomes. Fig. 2 shows the cross-validation score of the 8 models. After evaluating various models, it was determined that the Support Vector Classifier (SVC) model yielded the highest cross-validation score, indicating its superior performance in predicting student outcomes. KNN exhibits the lowest mean score of 79.26% in comparison to the other models evaluated. The Support Vector Classifier (SVC) achieves a testing accuracy of 84.61% and an impressive training accuracy of 97.3% in predicting pass/fail outcomes. Fig 1 shows the accuracy of SVC. The heat map depicted in Figure 4 illustrates the correlation of various parameters with the pass/fail outcome.

```
model trained with SVC()
Model accuracy on train is::  0.9738562091503268
Model accuracy on test is::  0.8461538461538461
Confusion matrix train is::  [[74  3]
 [ 1 75]]
Confusion matrix test is::  [[16  3]
 [ 3 17]]
Wrong Predictions made: 6 / 39
```

Fig 1: Accuracy of SVC

```
cross validation model : LogisticRegression
Mean score :  87.5263157894737
_____________________________________________________
cross validation model : DecisionTreeClassifier
Mean score :  88.57894736842105
_____________________________________________________
cross validation model : RandomForestClassifier
Mean score :  89.1842105263158
_____________________________________________________
cross validation model : KNeighborsClassifier
Mean score :  79.26315789473685
_____________________________________________________
cross validation model : SVC
Mean score :  90.63157894736842
_____________________________________________________
cross validation model : GaussianNB
Mean score :  73.5
_____________________________________________________
cross validation model : AdaBoostClassifier
Mean score :  87.07894736842104
_____________________________________________________
cross validation model : GradientBoostingClassifier
Mean score :  88.07894736842104
_____________________________________________________
```

Fig. 2 : Cross-validation scores of prediction models

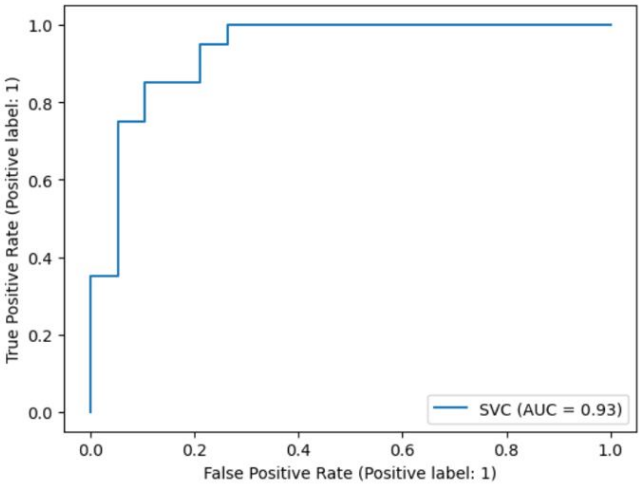


Fig. 3: ROC curve of SVC

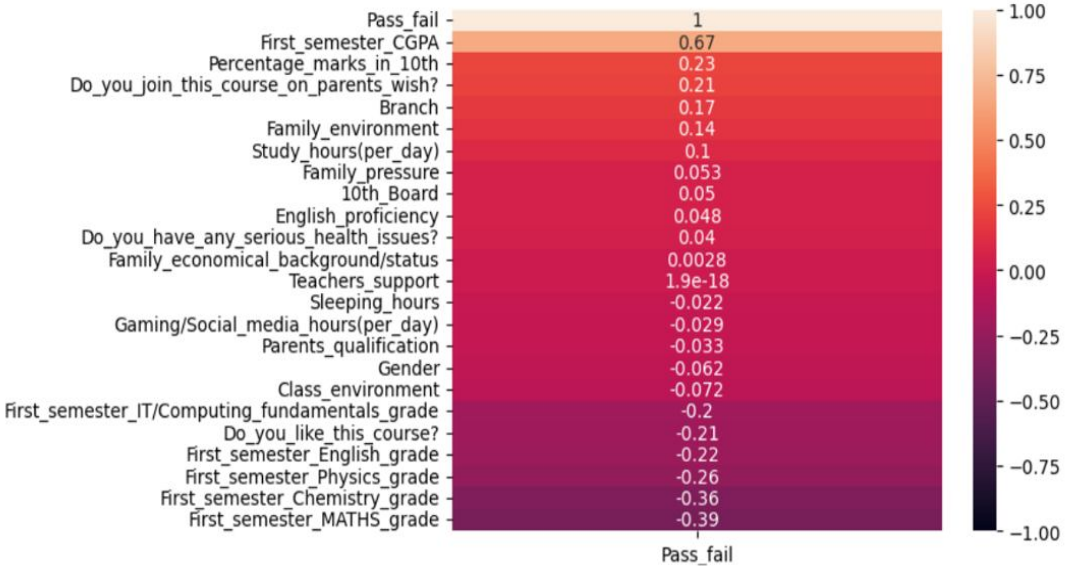


Fig.4: Heat map

V. CONCLUSION

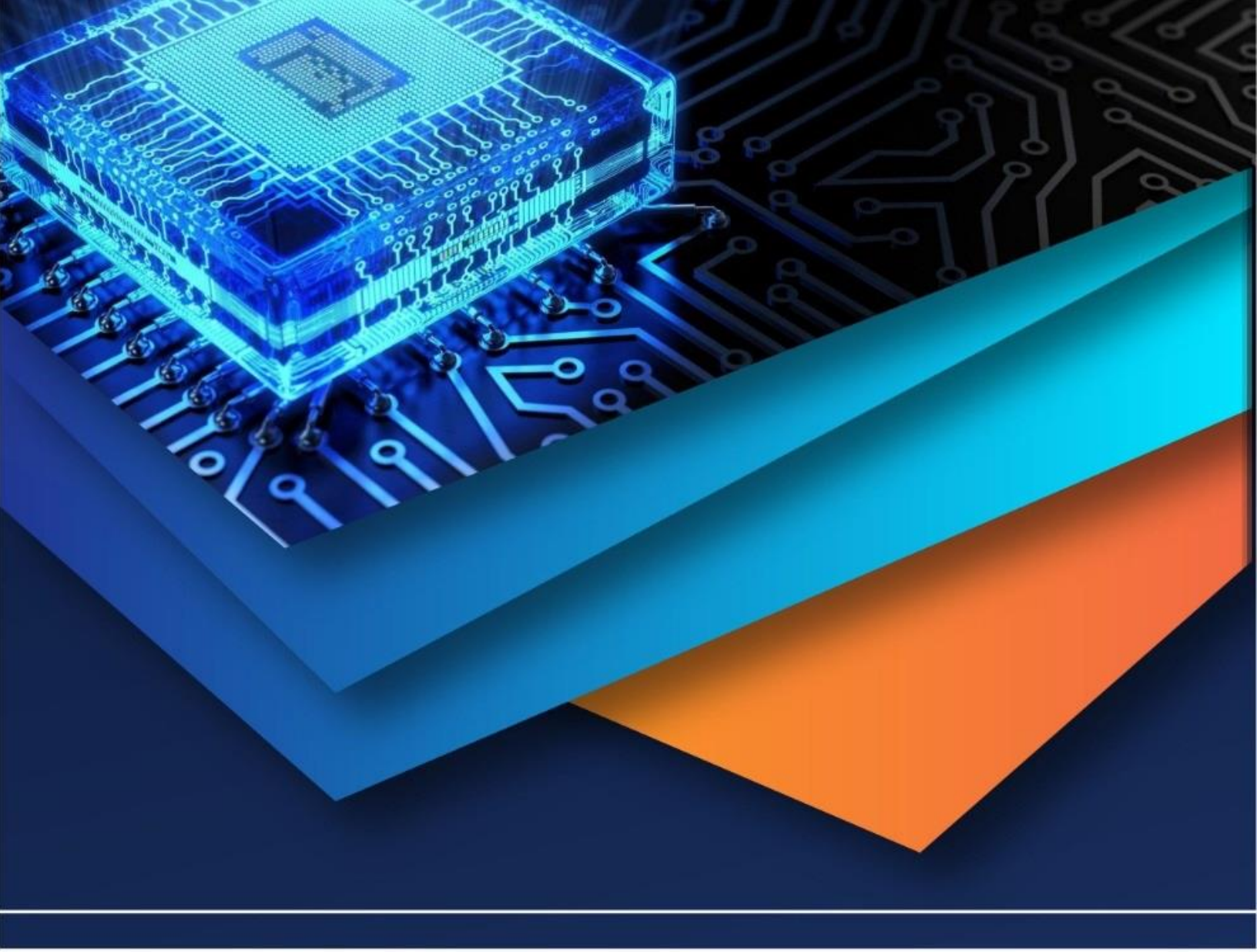
A machine learning-based system for analyzing the academic performance and prediction of pass/ fail outcomes of polytechnic students has been implemented. Data has been collected from Polytechnic students through Google Forms. Analysis of the data revealed that a majority of polytechnic students originate from middle-class or lower-middle-class families. Over fifty percent of students enroll in this institution due to financial strain and parental pressure. Some students enroll in this course without a clear career plan, as indicated by the data. Therefore, the amount of time they invest in their studies tends to be lower. Eight machine learning-based prediction models were employed to assess the academic performance of students, with the Support Vector Classifier (SVC) emerging as the top-performing model, achieving the highest cross-validation score and giving an accuracy of 90.3%.In conclusion, educational institutions may increase student performance, allocate resources more effectively, and promote a data-driven approach to student assistance and intervention by utilizing machine learning for academic performance monitoring.

Conflict of Interest

The authors have no conflicts of interest to declare.

REFERENCES

[1] Joy, L. C., & Raj, A. (2019, March). A review on student placement chance prediction. In 2019 5th International conference on advanced computing & communication systems (ICACCS) (pp. 542-545). IEEE.
[2] Thangavel, S. K., Bkaratki, P. D., & Sankar, A. (2017, January). Student placement analyzer: A recommendation system using machine learning. In 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) (pp. 1-5). IEEE.
[3] GE, O., Mamah, C. H., Ukekwe, E. C., & Nwagwu, H. C. (2020). A machine learning based framework for predicting student's academic performance. Physical Science & Biophysics Journal, 4(2).
[4] J. Xu, K. H. Moon and M. van der Schaar, "A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 5, pp. 742-753, Aug. 2017, doi: 10.1109/JSTSP.2017.2692560.}
[5] Wang, D., Lian, D., Xing, Y., Dong, S., Sun, X., & Yu, J. (2022). Analysis and prediction of influencing factors of college student achievement based on machine learning. Frontiers in Psychology, 13, 881859.
[6] Ma, X., & Zhou, Z. (2018, January). Student pass rates prediction using optimized support vector machine and decision tree. In 2018 IEEE 8th annual computing and communication workshop and conference (CCWC) (pp. 209-215). IEEE.
[7] Ying D, Ma J. Student Performance Prediction with Regression Approach and Data Generation. Applied Sciences. 2024; 14(3):1148. https://doi.org/10.3390/app14031148
[8] https://www.ibm.com/topics/logistic-regression
[9] https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
[10] https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning
[11] https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/
[12] https://www.tutorialspoint.com/scikit_learn/scikit_learn_support_vector_machines.htm
[13] https://www.analyticsvidhya.com/blog/2021/11/implementation-of-gaussian-naive-bayes-in-python-sklearn/
[14] https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/
[15] https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY