International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

Machine Learning-Based Method for Predicting Academic Performance of Students

Puneet Matapurkar

Dept. of Mathematical Science & Computer Applications, Bundelkhand University, Jhansi (U.P.), India

***Abstract*— Automatic Student accomplishment forecasts are a crucial task since educational databases include a vast amount of data. EDM takes care of this job. EDM is developing methods to locate data from academic backgrounds. The methods are applied to aid in learning and pupil comprehension. Educational institutions typically care about how many students passed or failed to complete the necessary preparations. Many issues can arise for institutions at any time. One is the student's subpar performance. The second is when a student leaves school because of the monetary difficulties, curriculum's complexity, lack of support and psychological issues, etc. The use of machine learning can address these problems. To enhance the accuracy of our student performance prediction model, we used an ensemble approach based on oversampling, using both the random forest and Xgboost classifier. A method for formulating strategies to address classification and forecasting challenges.**

***Keywords*— education data mining (EDM), Bayesian networks, vector machines, machine learning, voting classification, XGBoost classifier, deep neural network**

I. INTRODUCTION

Each nation must invest in high-quality education to advance. With the aid of e-learning, admissions systems, learning management systems, academic information systems, etc., the amount of data in the education sector is increasing day by day. The information gathered from students is typically utilized to create basic decision-making questionnaires. However, because of their complexity and size, most data sets go underused. Thus, it is of considerable interest to examine this vast amount of educational data in order to forecast student performance.

Predictions of student performance have been made for several reasons, including preventing dropouts, helping struggling students get back on track, optimizing course loads, and more.

Machine learning algorithms, which have been widely used, have recently emerged as a crucial component of educational evaluation. Estimating students' academic achievement with data mining [1]. So, it would be useful for teachers and professors to enhance their lessons.

In addition, the instructor observed the students' work and evaluated two types of models (a Bayesian network and a decision tree) for estimating the students' final grade point average (GPA) in UG and PG courses, respectively. There is a total of 936 and 20,492 student records in the two datasets, respectively.

Technology is being developed with the specific intent of making life easier for people in many aspects. Since computers have advanced so quickly in recent years, people have been able to create even-more complicated systems and assign increasingly difficult tasks to them [2]. Machine learning enables humans to be helped through the automation process and its one application that is employed in daily life is in the area of customer support, where a bot can answer and provide information about a current activity or program [3].

Machine learning is a type of approach where a computer or machine algorithm is able to infer information or decisions and continuously enhances its performance on intelligent tasks similar to those performed by humans.)e used machine learning models such as SVM, RF, J48, and ANN suffer from the overfitting problem[4] Machine learning has recently improved significantly in terms of effectiveness and accessibility, and it is now widely used in a variety of fields, such as business [5].

Despite this, there is mounting evidence that failure prediction models based on machine learning improve the reliability of a variety of systems. Despite extensive work completed as lab prototypes, few industrial implementations, such as true industry data, are published in the literature[6].

II. RELATED WORD

(Pereira et al., 2019) Employ it as a characteristic in ML models. 486 CS1 students successfully used a set of grade-related features they created using online judge log data stored in a database. They used this collection of features in an improved ML pipeline that also included a random search hyperparameter tweaking, together with an automated technique, and an enhanced algorithm.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

They were able to accurately predict the final graduation rate (75.55%) by analyzing data from the first two weeks of the semester. They demonstrate the superior performance of our pipeline in situations like these [7].

(Almasri et al., 2019) Undergraduate (UG) performance at a business school's MIS department may be predicted with the use of a novel dataset, as proposed by (Almasri et al., 2019). In an attempt to forecast the future success of MIS majors in Jordanian colleges, this poll has been designed specifically for them. This article describes the dataset and draws parallels to two similar resources. The results showed that the R Square (R2) values in the MIS Dataset were the most accurate, at 88.5%, followed by Dataset 2 at 19.3% and Dataset 1 at 69.1%. In terms of correlation and dispersion, the results demonstrate that the suggested dataset is both efficient and effective. To aid educational officials in improving future student performance, the MIS Dataset includes additional analytics including classification and clustering techniques[8].

(Ko & Leu, 2018) Using machine learning (ML) techniques, identify the salient characteristics of a successful student course that are often displayed through computers. Five ML algorithms—vector support machines, naive infants, various perspectives, logistic regression, and decision-making, —have been tested for performance, precision, and aesthetics. The results show that naive Bayes is best suited to forecast students' final performance. Scores for this category are 77.53% and 88.68%, respectively. The evaluation of sensitivity and precision. The classification model is discussed in this paper along with a plan to increase classification accuracy [9].

(Deo and others, 2020) Using the data on student performance gathered, the ELM model is benchmarked against the rival models, as well as the Random Forest (RF) and Volterra models. Six years' worth of instruction spanning introductory to expert levels across many modalities. By analyzing and researching marks as crucial WS variables, ELM (RF and Volterra) was able to outperform its ONC and ONL competitors, resulting in an intermediate ranking. As a consequence, the relative error during the testing phase decreased to 0.74% from 3.12% and 1.06%, and the estimations for the ONL range decreased to 0.51% from 3.05% and 0.70%.[10].

(Mngadi et al., 2020) The goal of this approach is to forecast a student's performance attrition so that at-risk pupils may be identified early in the school year and given the help they need to succeed via targeted interventions.

The four risk profiles served as the basis for the instruction of predictive algorithms for the equipment learning process. This study demonstrates that it is more challenging than previously thought to predict which students, based on their personal characteristics, family history, and school environment, would be at risk for academic failure [11].

III. Problem Statement

Every school should make improving student outcomes and the quality of instruction a top priority. One possible component of providing high-quality education is a thorough review of the student's background. Accurately predicting a student's performance is still difficult for a variety of reasons. Most performance prediction methods suffer from inefficiency and the introduction of unnecessary or irrelevant factors. Predicting student achievement is notoriously difficult, thus methods like neural networks, decision trees, naive Bayes, voting classification, & KNN technology are often utilized as a stumbling block. These methods are used to collect student records, supplement decision-making algorithms, aid in pattern extraction, etc. The inspection period was decreased from about 120 hours to 5 minutes for the identical procedure. [12] Instead of using a questionnaire's self-efficacy subscale, for example, a single item may be used to construct a robust predictor of students' performance. Instead of ignoring these problems, we may use the voice classification tool to fix them and extract the crucial features from the subscales.

IV. Proposed Methodology

To determine the crucial characteristics a successful student often exhibits over the course of a computer, a framework is created in this paper that draws on supervised and unsupervised ML techniques. We zero in on four possible traits shared by successful students: metacognitive self-regulation, self-efficacy, study habits, and computational efficiency. The Computer Self-Efficacy Scale and the MSLQ's Efficiency, Metacognition, and Time Management Scales are selected from relevant research topics, respectively. A Measure of Computer Confidence Both supervised and unsupervised learning options, including a voting classifier and a deep neural network, are provided. Successful children's defining characteristics are isolated by a combination of FP growth and gauze mixing. Choose a Kbest strategy to choose features.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

Therefore, researchers should include the effects of these factors on student achievement while developing their learning models. This research demonstrates how ML algorithms may be used to identify key characteristics of high-achieving kids. Two paper-and-pencil examinations measuring students' computer-related concepts and one online test measuring their computer-related knowledge gained in the laboratories will be used to determine each student's final grade.[13].

A. Data Pri-processing

Data pre-processing is an essential part of building a machine-learning model. The first step is to import the data set in the form of a CSV file. Since then, null values have been validated. The sum of a numerical variable is then transformed into a nominal one. At the start of the last test, 215 surveys were given out to students. Teachers have access to all students' final transcript information with their permission. The real data were carefully checked (vote classifier) before being fed into an ML system.

B. Attributes Selection

The discovery of important factors that affect the prediction result via exploratory data analysis. The seaborn python package is used to create the heat map. Between -1 and 1, the range is variable. The negative characteristics are correlated negatively with other variables, while the positive characteristics are correlated positively with other factors. An example from our data set demonstrates the favorable correlation between a mother's education and her child's final grade. Weaknesses in our data set might explain negative correlations. For instance, if a student receives a failing grade, their overall performance is almost certainly worse than it was. Near-zero numbers, whether positive or negative, show little to no correlation with other variables. These characteristics may be eliminated from databases without impacting performance.

- Select k best
- FP growth algorithm
- Deep neural network

C. Machine learning Techniques (Voting Classifier)

The field of educational data mining often employs predictive models to foretell learner outcomes. Predictive modeling requires a wide variety of work, including but not limited to classification, regression, and categorization. Students most often engage in classification-based prediction tasks.

Predicting a student's performance is possible with certain algorithms. As a result of the difficulty humans have in selecting the best option among competing hypotheses or points of view, many different approaches have been proposed for building voting systems, including parametric, non-parametric, probabilistic, heuristic, logical, etc. One tactic for using voting systems with various learning methods is to choose a classification method (e.g., decision trees, kNN, etc.) that uses the same training datasets. Three distinct varieties of machine learning, namely feedforward NNs, SVR, and LR, will be examined. [14] According to the World Health Organization, the following variants are of concern: [15]. classifier yields a unique set of predictions. If you want an accurate prediction, you may use a voting system where the top students all receive votes[16].

D. Voting Classifier

It is a machine learning model that trains several individual models and then predicts outcomes using the models' best probabilities of a given class. The results of every classifier that was submitted for voting are added together, and the outcomes are predicted by the classifier with the most votes. The idea is to train and anticipate outcomes using a unified model based on the aggregate majority of votes across all classes, rather than developing and validating individual specialized models.

The voting Classifier supports two distinct voting methods.

- Hard voting (HV)
- Sort voting (SV)

V. PROPOSED ALGORITHM

Step 1: Show the training data.
Step 2: D: C-class training data set with labels.
Step 3: L: The learning algorithm is voted on by the classifier voters.
Step 4: N: no. of L utilized.
Step 5: Do n=1 to N

1. Eliminate L from a classification using Dn.
2. Update voting contrasts Wn with Cn made by Ln.
3. Aggregate vote to an ensemble

Step 6: The majority performance class is determined based on a number of gathered data.
Step 7: The evaluation of the performance matrix.
Step 8: Performance prediction for students.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

VI. FLOWCHART OF METHODOLOGY

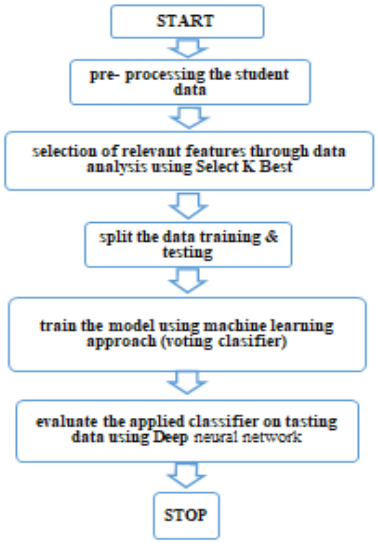

Figure 1: Methodology Flowchart

Students who have received online technical training may now evaluate the efficacy of their learning by using voting categories thanks to the research (Minaei-Bidgoli, Kortemeyer, & Punch, 2004). Their findings show that a wide variety of categorization combinations significantly enhances the dataset's accuracy. Since the decision depends on the collective output of several models, classification combinations frequently have greater predictive performance than a single classification.

An improvement approach driven by combined vote classifications is used to predict the final score[17]. The great accuracy achieved in earlier investigations suggests a technological approach to localized voting of low classifiers, whereby neither local uniqueness nor global learning approaches are ignored.

VII. PROPOSED MODEL

Based on students' demographic information (gender, location), pre-university information (gap year between high school graduation and entering university, university entrance score, university entrance categories), and learning results from the first and second semesters, this study applies the aforementioned techniques to predict economic students' final grade point averages. The factors that influence students' academic success were then identified.

Living location, sex, gap year, university entrance rates, first- and second-year average ratings, and so on are used in the early phases of the education program to predict final student accomplishment. The proposed research prototype is divided into three stages: gathering and incorporating data, gathering data, and evaluating and developing the model. Due to the vast quantities of information available in institutional repositories of education, the automated prediction of student performance is a formidable task. This task is addressed by educational data mining (EDM). EDM creates methods for mining the informational richness of the classroom. Students and their learning environments may be better understood with the help of these methods. The UCI Machine Learning Repository has been a valuable resource for us during this study. The percentage of students who study hard and yet fail is a common area of interest for educational institutions. Researchers have been found to focus excessively on the choice of appropriate algorithms for a particular classification while overlooking the need of addressing the enormous dimensionality of data, class imbalances, and categorization mistakes throughout the data mining process. In order to improve the accuracy of the student prediction model, this study employs the ensemble approach used as an over-sample technique. Problems with classification and making accurate predictions are among the targets of an approach based on ensemble techniques. The study found that all four locations were viable options for large-scale production. [18].

Machine learning now plays a crucial role in every technological area. Even they may admit that innovation in fields like hotel management, train travel, public transit, healthcare, manufacturing, and other related fields is an integral part of their daily life. Image processing, pattern recognition, medical diagnosis, predictive analytics, and product recommendations are just some of the fields where ML has been put to use during the last several decades. The future value of a building must account for the fact that housing costs fluctuate annually. They use a prediction-oriented Kaggle Dataset. A home price projection may be used to analyze consumer pricing and assist set housing prices. Using local UK schools, they want to predict home values using a variety of machine learning techniques. Because of the positive effect that proximity to a good school has on property values in the United Kingdom, we propose using Machine Learning models to predict how changes in those values would occur. The K-best techniques to feature selection will be used.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

The study's findings will verify that the ridge regression or the light GBM is used as the most effective ML model of a CNN. These models would favorably affect home values by expanding on a wide range of input characteristics. Finally, this study's impact was planned to encourage and help future academics build a genuine model that could readily and reliably anticipate house price movements.

VIII. RESULT

A Windows 10 and Windows 4GB HP PC ran the trial. UCI Machine Learning Repositor evaluates our prototype utilizing this research's student efficiency dataset. Two Alentejo schools in Portugal provided 2005-2006 data. It has 1044 examples with 33 criteria, including student degrees, demographics, society, and schools. Long-term economic growth requires education. Portuguese education has improved in recent decades. Portugal leads the EU despite high student failure and dropout rates. In 2006, 40% of 18-24-year-old Portuguese left school early, compared to 15% throughout the EU (Eurostat 2007). Failure of the main math and Portuguese courses, which are critical to school achievement, is particularly harsh (e.g., physics or history). Due to information technology advances, corporate information was a significant emphasis, and commercial and organizational databases developed dramatically. Success and decision-making might be enhanced by analyzing trends and patterns in this massive collection. Experts are just human; therefore, they can only do so much, and sometimes they overlook things. Intriguing, high-quality data may be provided through automated data analysis to decision-makers.

A. Comparison Chart

Student performance data sets have been compiled using the UCI ML Repository. The information was collected from two Portuguese schools in the Alentejo region during the 2005-2006 school year. There are 1044 total examples included, each with its own unique set of 33 attributes including student grades, demographics, background info, and more. After completing 9 years of elementary school, students in Portugal enter the 3-year secondary school system. The vast majority of pupils take advantage of the public school system since it is free of charge. Key areas like Portuguese and mathematics are covered in a wide variety of courses (including those in the sciences, the arts, and the humanities). Similar to many other countries (including France and Venezuela), a grading scale of 0–20 is used, with 0 being the lowest and 20 being the highest possible score.

Students are graded based on their performance on three separate assessments administered during the school year.

To forecast students' academic success in college, we turned to machine learning techniques. Metrics from each training iteration are saved. The outcomes of the suggested model are shown. After finishing data preparations, we used simple plots of distributions to get a feel for the data set's structure. Next, five important machine learning technologies were evaluated, and the factors used in the evaluation were discussed. The whole 480-characteristics data set was used for the analysis. During the pre-processing phase, certain traits were removed since they did not provide enough differentiation. While almost all students live at home with their parents and have access to their own computers, very few respondents disclosed their families' incomes (perhaps due to privacy concerns).

B. Comparison Chart

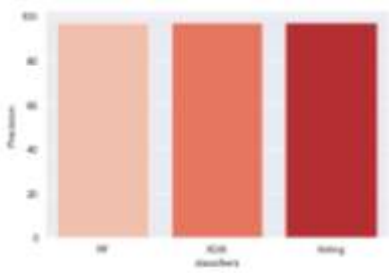

Figure 2: The comparison plot of Accuracy (RF, XGB, and Voting)

Above, you can see a plot of accuracy against contrast. We may repeat the first experiment using the same data set and run classification algorithms (Vote for Random Forest and XGBoost) to assess the efficacy of the suggested model phases. When compared to the existing literature and the lower level of precision achieved by other approaches, the accuracy achieved by voting with Random Forest and XGBoost was considered inadequate.

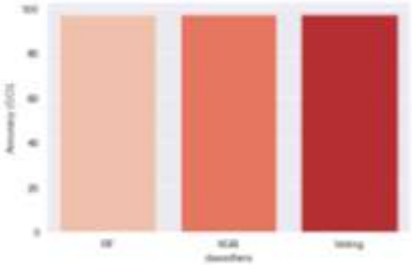

Figure 3: The comparison plot of Precision (RF, XGB, and Voting)

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 12, Issue 04, April 2023)

The Precision comparison plot is shown in the preceding picture. We find that the first experiment may be run independently of the specified model phases, meaning that the classification algorithms (Voting with Random Forest and XGBoost) can be used on their own. Random forest is one of the precious techniques for classification and prediction problems [19]

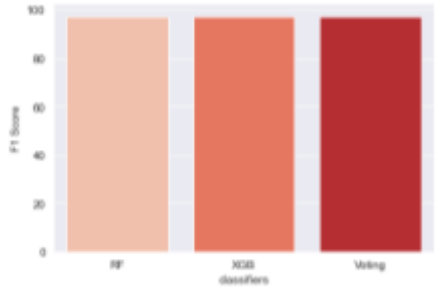

Figure 4: The comparison plot of Recall (RF, XGB, and Voting)

The Recall comparison plot is shown in the preceding picture. We find that the first experiment may be run independently of the specified model phases, meaning that the classification algorithms (Voting with Random Forest and XGBoost) can be used on their own.

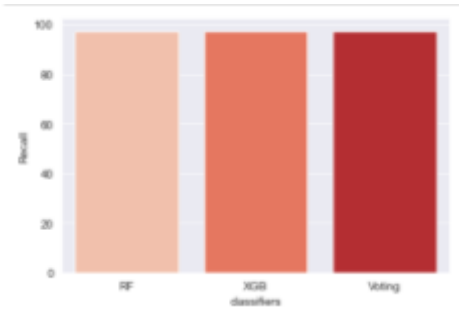

Figure 5: The contrasted plot of F1 Score (RF, XGB, and Voting)

The Recall comparison plot is shown in the preceding picture. We find that the first experiment may be run independently of the specified model phases, meaning that the classification algorithms (Voting with Random Forest and XGBoost) can be used on their own.

IX. Conclusion

As a means of early student failure identification, this study compares and contrasts data mining and preprocessing techniques. Support vector machines were shown to be inferior to other approaches and models (Random Forest and xgboost voting) employed in this study.

There were two types of data collectors used. Students' academic performance was shown to be negatively impacted by the number of days they were absent from school, whereas the number of hours spent in class had no effect. We want to broaden the study by including other factors, such as the facilitators' and instructors' use of encouraging and motivating approaches and the inclusion of additional resources in an online learning environment. Functions like psychological characteristics that affect pupils' performance are also taken into account. Future studies aiming to predict students' performance in higher education should also make use of more engaging and extensive data collection.

Although several well-known classification techniques are used in this area, this paper proposes that supervised decision tree categorization is essential to a model for predicting students' success. Additionally, classification performance is improved by using an ensemble technique. Ensemble techniques are a methodology developed to solve classification and prediction problems. This study underlines the necessity of data pre-processing and technique tweaking in fixing data quality issues. The Alentejo, Portugal, experimental dataset used here is publicly accessible from the UCI Machine Learning Repository. Three supervised algorithms (Voting, random forest, and xgboost) are employed experimentally in this study. The results showed that, among other things, J48's accuracy hit 99%.

This research used random forest and xgboost voting to create a model that can predict a student's efficiency. The accuracy of student prediction models like Random Forest and xgboost is measured by a categorization of votes. An ensemble method is also used to improve the accuracy of these sorts of classifications. High accuracy is achieved by the recommended ensemble model consisting of the Random Forest and xgboost Categorization in 95.78 percent of cases.

REFERENCES

[1] B. Kumar and S. Pal, "Mining Educational Data to Analyze Students Performance," Int. J. Adv. Comput. Sci. Appl., 2011, doi 10.14569/ijacsa.2011.020609.

[2] E. Adamopoulou and L. Moussiades, An Overview of Chatbot Technology, vol. 584 IFIP. Springer International Publishing, 2020. doi 10.1007/978-3-030-49186-4_31.

[3] A. M. Rahman, A. Al Mamun, and A. Islam, "Programming challenges of chatbot: Current and future prospective," 2018. doi 10.1109/R10-HTC.2017.8288910.

[4] V. Rohilla, S. Chakraborty, and M. Kaur, "Artificial Intelligence and Metaheuristic-Based Location-Based Advertising," Sci. Program., vol. 2022, p. 7518823, 2022, doi 10.1155/2022/7518823.

[5] Y. S. Reshi and R. A. Khan, "Creating Business Intelligence through Machine Learning: An Effective Business Decision Making Tool," Inf. Knowl. Manag., 2014.

[6] S. Proto et al., "PREMISES, a Scalable Data-Driven Service to Predict Alarms in Slowly-Degrading Multi-Cycle Industrial Processes," 2019. doi 10.1109/BigDataCongress.2019.00032.

[7] F. D. Pereira, E. H. T. Oliveira, D. Fernandes, and A. Cristea, "Early performance prediction for cs1 course students using a combination of machine learning and an evolutionary algorithm," 2019. doi: 10.1109/ICALT.2019.00066.

[8] A. Almasri, E. Celebi, and R. Alkhawaldeh, "MISDataset: Management information systems dataset for predicting undergraduate students' performance," 2019. doi: 10.1109/ICCIA.2019.00017.

[9] C. Y. Ko and F. Y. Leu, "Analyzing attributes of successful learners by using machine learning in an undergraduate computer course," 2018. doi: 10.1109/AINA.2018.00119.

[10] R. C. Deo, Z. M. Yaseen, N. Al-Ansari, T. Nguyen-Huy, T. A. M. P. Langlands, and L. Galligan, "Modern Artificial Intelligence Model Development for Undergraduate Student Performance Prediction: An Investigation on Engineering Mathematics Courses," IEEE Access, 2020, doi: 10.1109/ACCESS.2020.3010938.

[11] N. Mngadi, R. Ajoodha, and A. Jadhav, "A Conceptual Model to Identify Vulnerable Undergraduate Learners at Higher-Education Institutions," 2020. doi 10.1109/IMITEC50163.2020.9334103.

[12] S. U. Ahmed, M. Affan, M. I. Raza, and M. Harris Hashmi, "Inspecting Mega Solar Plants through Computer Vision and Drone Technologies," 2022. doi 10.1109/FIT57066.2022.00014.

[13] P. R. Pintrich, D. A. F. Smith, T. Garcia, and W. J. Mckeachie, "Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq)," Educ. Psychol. Meas., 1993, doi: 10.1177/0013164493053003024.

[14] K. K. Hasan, M. A. Hairuddin, R. F. Mustapa, S. A. Nordin, and N. D. K. Ashar, "Machine Learning Approach of Optimal Frequency Tuning for Capacitive Wireless Power Transfer System," Int. J. Emerg. Technol. Adv. Eng., 2022, doi: 10.46338/ijetae1122_07.

[15] M. A. Alamri et al., "Molecular and Structural Analysis of Specific Mutations from Saudi Isolates of SARS-CoV-2 RNA-Dependent RNA Polymerase and their Implications on Protein Structure and Drug–Protein Binding," Molecules, vol. 27, no. 19, 2022, doi 10.3390/molecules27196475.

[16] "ML | Voting Classifier using Sklearn."

[17] W. Zang and F. Lin, "Investigation of web-based teaching and learning by boosting algorithms," 2003. doi: 10.1109/ITRE.2003.1270655.

[18] R. Asghar et al., "Wind Energy Potential in Pakistan: A Feasibility Study in Sindh Province," Energies, 2022, doi: 10.3390/en15228333.

[19] V. Rohilla, D. S. Chakraborty, and D. R. kumar, "Random Forest with Harmony Search Optimization for Location Based Advertising," Int. J. Innov. Technol. Explor. Eng., 2019, doi: 10.35940/ijitee.i7761.078919.