IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

OPTIMIZING STUDENT PERFORMANCE: THE INTEGRATION OF MACHINE LEARNING FOR EARLY IDENTIFICATION AND INTERVENTION

[1]Deshpande G. R., [2]Wananje Aditya Vitthal, [3]Kavtikwar Omkar Govindrao, [4]More Shrikant Kailas

[1] Assistance Professor, [2]Student, [3]Student, [4]Student

[1]Department of Computer Engineering

[1]Gramin Technical & Management Campus Vishnupuri, Nanded, India

Abstract: Helping Students Succeed: Using Smart Computers to Spot Those Needing Support Early. In school or college, students sometimes need extra help to do well. Teachers often give additional assignments and projects to support those who are struggling. But a big challenge is figuring out who needs help early on. This research explores how we can use smart computer techniques, like Machine Learning, to find students at risk of struggling. We wanted to create a special computer model using a mix of techniques to predict who might need help. We used different algorithms like Naïve Bayes, Random Forest, Decision Tree, and others in our project. We tested our model using a set of information about students, like where they're from and how they're doing in school or college. We also made a web app to make it easy for teachers to use our model and get predictions. By combining the best techniques, our model was able to predict with high accuracy which students might need extra support. In the end, our goal is to help teachers identify struggling students early so they can give them the support they need.

Index Terms - At-risk students, classification, dropout prediction, hybrid model, machine learning techniques, stacking ensemble model, student performance prediction

I. INTRODUCTION

Helping Students Succeed: Using Smart Computers to Spot Those Needing Support Early. Sometimes, students face challenges that can lead to them struggling in school, whether it's due to personal problems, family situations, or not getting enough support. Identifying and helping these students early on is crucial for their success. Predicting students' performance early can guide teachers in providing the right assistance, like extra courses or assignments. Analyzing the performance of each student in a large school can be tough for teachers due to limited resources. Machine learning techniques have proven useful in identifying at-risk students. In this project, a hybrid model using different machine learning algorithms is created, focusing on high school students in Turkey. The hybrid model combines algorithms like Naive Bayes, Support Vector Machine, Random Forest, and more. The goal is to predict at-risk students with high accuracy. The study uses ensemble methods like Bagging, Boosting, and Stacking to combine different machine learning techniques. The project aims to identify high school students at risk early and provide timely support. The research also focuses on understanding the factors affecting students' school performance. A newly collected dataset from high school students was used, with a focus on academic data. The developed hybrid model contributes to solving the challenge of identifying students at risk, providing higher prediction performance. The research includes creating a web application to help teachers more effectively identify at-risk students. In conclusion, this study presents a hybrid ensemble model that successfully identifies students at risk, offering a promising solution for educators. The paper covers background, methods, performance results, limitations, and concludes with future work considerations

II. BACKGROUND AND LITERATURE REVIEW

A) Strategies for Predicting and Improving Student Success

In addressing the challenge of identifying at-risk students and enhancing their performance, various solutions have been explored.

1) Machine Learning Techniques: Previous studies have highlighted the effectiveness of machine learning techniques [6], [7], [1], [8], [2], [5], [9], and [10]. These techniques involve using algorithms like Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, and AdaBoost.

2) Hybrid Model Innovation: This project introduces a novel approach by creating a hybrid model. Combining different machine learning algorithms, such as Bagging, Boosting, and Stacking, enhances predictive accuracy. The goal is to achieve better performance compared to individual models.

3) Data Analysis: Understanding the factors influencing student performance is crucial. The study uses a newly collected dataset from high school students, focusing on both demographic and academic information. Analyzing these characteristics contributes to a more efficient predictive model.

4) Early Intervention: The ultimate aim is to identify at-risk students early, enabling timely support from teachers. This proactive approach involves providing additional resources, courses, or assignments to help struggling students succeed.

By combining these strategies, the research aims to create a robust solution for predicting at-risk students and improving their academic performance. The subsequent sections will delve into the experimental setup, performance results, limitations, and the conclusion with considerations for future work

B) Understanding Key Factors in Predicting At-Risk Students

To enhance the accuracy of predicting at-risk students, it's crucial to identify key features in datasets that significantly impact the outcomes.

1) Demographic and Academic Information: The study focuses on two main types of data — demographic and academic. These features provide valuable insights into a student's background and their performance in school.

2) Data Analysis: Through careful analysis of the dataset, researchers aim to pinpoint specific characteristics that play a vital role in predicting students' academic success or challenges. This involves looking at factors like attendance, previous academic history, and socio-economic background.

3) Contribution of Academic Data: The research observes that academic data tends to have a more substantial impact on identifying at-risk students. This could include factors such as exam scores, assignment performance, and overall academic achievements.

By understanding and emphasizing these features, the study aims to create a predictive model that accurately identifies at-risk students. The next sections will detail the experimental setup, performance results, limitations, and the conclusion with considerations for future work

C) Hybrid Ensemble Model

In this study, a cutting-edge approach known as the Hybrid Ensemble Model takes center stage. Let's break down this innovative concept.

1) Combining Multiple Models: The hybrid model harnesses the power of ensemble methods, specifically Bagging, Boosting, and Stacking. This involves combining predictions from various machine learning algorithms to create a more robust and accurate model.

2) Two-Phase Architecture**:** The hybrid ensemble model operates in two distinct phases. The first phase involves creating base models, each employing different algorithms. The model that yields the best predictions fitting the training data is selected. In the second phase, a meta-classification model is determined. This model learns to effectively combine predictions from the base models

3) Stratified k-Fold Cross Validation: To prepare the training dataset for the meta model, the base models undergo stratified k-fold cross validation. This technique enhances the accuracy of the overall hybrid model.

4) Improved Prediction Results: By integrating various supervised machine learning algorithms such as Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression, and AdaBoost, the hybrid model aims to produce superior prediction results compared to individual models.

The primary objective of this project is to identify high school students at risk early in their educational journey and provide targeted support. The hybrid ensemble model plays a pivotal role in achieving this goal. The

subsequent sections will delve into the experimental setup, performance results, limitations, and the conclusion with considerations for future work

III. RESEARCH METHODOLOGY

A) DATA COLLECTION

In this project, it was planned to use the data set of Gramin college of engineering students studying in Nanded. The dataset was collected from Gramin college students in Nanded. Permission was obtained from the principle of Gramin college of engineering to collect the data set by distributing the prepared questionnaire form to students. Since students have different backgrounds, social aspects, and talent tendencies, questions in the questionnaire were prepared for students by considering these situations. The data collection tool was determined as a questionnaire. The prepared questionnaire consists of 2 parts. In the first part, questions about demo- graphic characteristics are asked to students, and in the second part, questions about academic characteristics are asked to students. The answer to some of the questions included in the questionnaire consists of two options: Yes/No. The answers to some questions will be taken from the students in the form of text (numbers or words). The survey was created via Google Forms. The collected data set is in line with the research purpose; it includes data features such as students' study times, exam scores, homework scores, future education plans, and social activities. The data set features form an important factor for the success performance of the model. The contribution of the collected data set to the research subject includes the use of current data, creation of the hybrid model by considering current student problems, the features of the education system, and data features that will be useful for identifying students at risk. The data set was collected from Gramin college of engineering. All grades in the data set are average grades evaluated out of 100. Since the grades in the dataset were mostly high, random data was added to the dataset so that the model could also recognize low grades. The academic and demographic characteristics in the data set were not changed, the randomly generated students' course grades were added to the data set.

B) DATA PREPROCESSING AND FEATURE ENGINEERING

The created data set to be ready for analysis, it must first go through the data preprocessing process. Thus, raw data will be transformed into more meaningful data for use in the model. It is necessary to identify missing, inconsistent, outlier and wrong data in the data set. Inaccurate estimation results can be obtained as a result of incomplete and inconsistent data. To prevent this situation, first of all, missing values in the dataset were observed. There are 20 features in total in the data set. There are in Gramin college of engineering that provide education in different fields and the courses are not the same. The courses in the data set are Mathematics, Literature, Physics, Chemistry, Biology, History, Geography, English. The year-end average score was calculated using the grades of these courses. Each lesson has a specific time interval per week. The total weekly course hours of the collected data set are 29. The formula for calculating the year-end grade point averages of the students is given below.

Year-end GPA = Total weight grade/ Total course hours... (1)
Total weight grade = Average grade of the course/ Weekly time of each course... (2)

For the purpose of this project, the ''Pass'' column was added to the dataset to monitor the students' passing the class. For the student to be considered successful in any course at the end of the academic year, the arithmetic average of the two semester average points must be at least 50. For this reason, students with an average of at least 50 or more at the end of the year will be considered successful. However, students with a year-end average below 50 will be considered unsuccessful. Binary values of ''1'' was assigned to students with a grade point average of 50 and above at the end of the year, and ''0'' was assigned to students with a grade below 50. The ''Pass'' data feature has been added to make models predictive targets. After the data preprocessing process, the data was visualized with various graphs to understand the relationship of data features and to see the contribution of these features to the student's success performance.

C) BUILDING THE MODEL FOR INITIAL RESULTS

In this research paper, Random Forest, K-NN, SVM and Logistic Regression were used to obtain the first model results. By comparing the machine learning algorithms used in previous studies described in the literature, the algorithms that provide the best performance were observed and selected for use in this study. Before the hybrid model was created, these algorithms were also individually evaluated and compared. For the first results of the models, both academic and demographic data in the data set were used. In addition, a data set containing only academic data features was used to observe the effect of academic data and

demographic data on the prediction result. The ‘‘Pass’’ column was set as the target label of the model. All algorithms were evaluated using the default model hyperparameters. Looking at the first results, the hyperparameter values of the algorithms that provide the best prediction will be adjusted before the hybrid model is created. In the proposed method, the stratified k- fold cross-validation method was used to split dataset into training and validation folds. Instead of dividing the data set into two parts as train and test sets, models were trained and tested with each data feature using the stratified k-fold cross validation method, and prediction results were obtained. The model built on the train dataset may have used only data containing certain features. This may affect the predictions made by the model on the test dataset. Using the cross-validation method is a pretty good way to avoid such problems. Results can be observed using cross validation to avoid overfitting problem. With this method, it can be observed whether the high performance of the model is random or not.

D) CREATING THE HYBRID MODEL

For this project we created multiple models for better accuracy. The accuracy performance of each model is different due to the errors made by the models considering different points in the data set. Ensemble learning technique is a good way to use to improve the performance of the models. With Ensemble learning, results are combined using multiple best performing models. Thus, a clearer and higher estimation result can be obtained. In this project, the stacking method has been used. It is one of the ensembles learning techniques, to create the hybrid model. With this approach, the performance of the predictive model is increased, and margins of error are reduced. The hybrid model was created with the models used for initial results by the stacking method. Each algorithm was tested for meta-learner, and the algorithm that gave the best performance result was used as a meta-learner. According to the results obtained, Random Forest, Logistic Regression and KNN algorithms were used as base learners. The Support Vector machine algorithm was used as a meta learner. The diagram showing the setup stages of the hybrid ensemble model with the stacking approach is given in Figure 1.

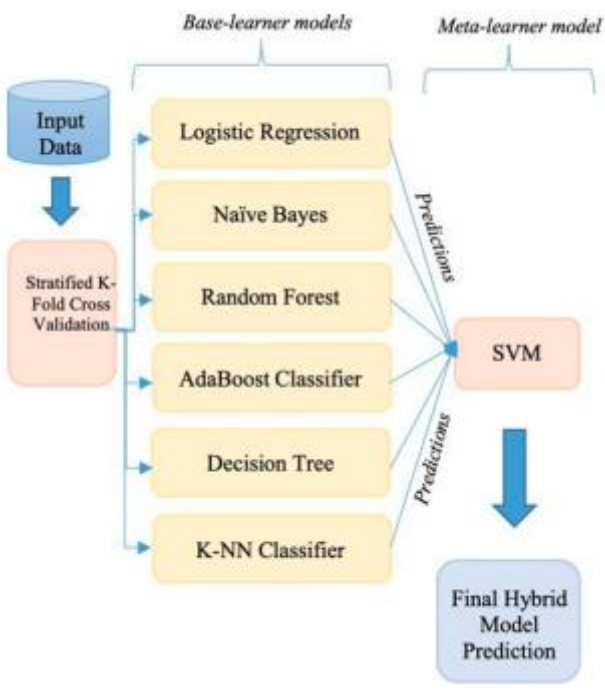

Fig 1. Hybrid ensemble model with stacking approach.

In the method presented in this project, stacking 10-fold cross validation is used to create a new training dataset for meta learner. In this project, SVM was used as the meta learner model and as a result of the meta learner model, the prediction results of the final hybrid model were obtained. The block diagram of the process followed in this study is shown in Figure 2.

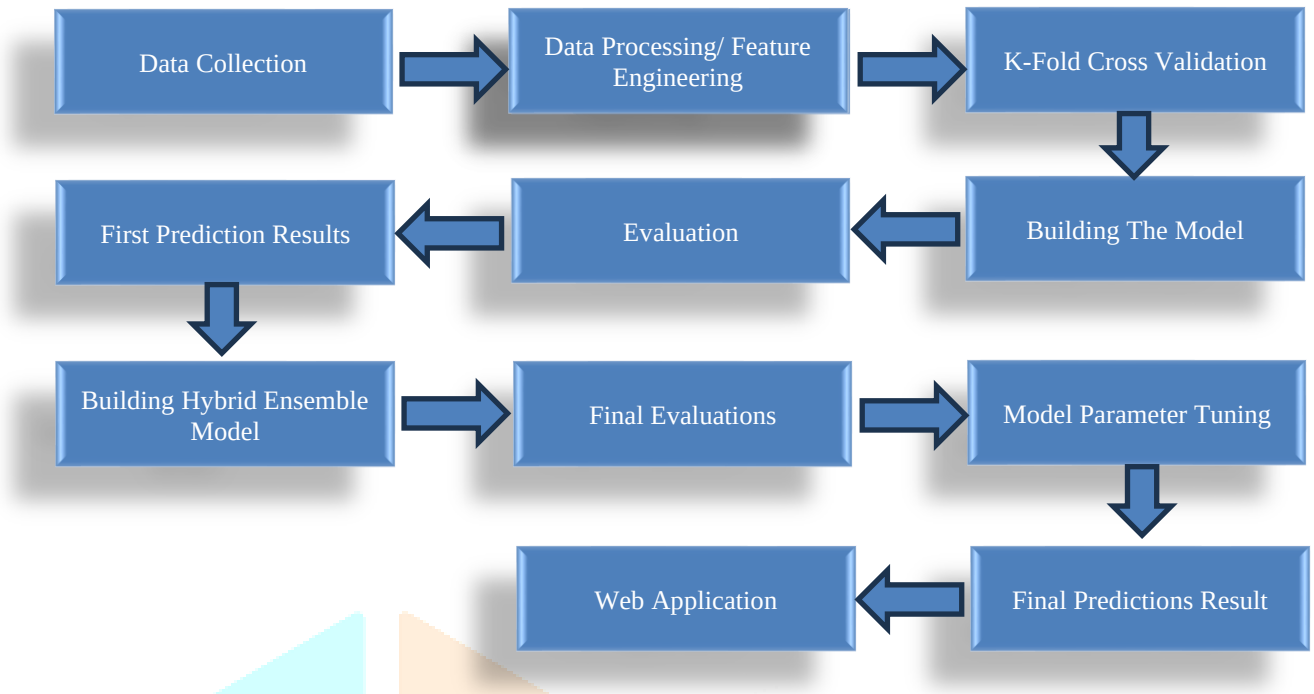

Fig 2. Project Progress Diagram

F) DEVELOPING WEB APPLICATION

The stream lit web application for the student behavior prediction system acts as a bridge between sophisticated machine learning algorithms and practical educational applications. Its user-friendly interface enables educators to effortlessly interact with the predictive model, inputting relevant student data and receiving instant predictions. This real-time capability empowers teachers to make informed decisions about providing additional support to students at risk of academic challenges. The application's design prioritizes simplicity and accessibility, ensuring that even users without extensive technical expertise can leverage the power of machine learning for proactive student intervention. Overall, the Stream lit web application represents a crucial component in translating complex predictive analytics into actionable insights within educational environments.

The python programming language and the Stream lit framework were used for the construction of the web application. Stream lit is a python framework and allows the creation of web applications on topics such as machine learning and data science. Therefore, stream lit was used to create a web application for prediction, and CSS was also used to make changes to the appearance of the application. The main functionality of the web application is the prediction page where the predicted result is obtained.

IV. Discussion

- **Result Interpretation:** The interpretation of results indicates that the hybrid ensemble model effectively leverages the strengths of diverse algorithms, leading to more accurate predictions of student behaviour.
- **Practical Implications:** The practical implications of this research are substantial. Educators can use the predictive model to identify at-risk students early, allowing for targeted interventions and support mechanisms. This proactive approach can potentially mitigate academic challenges and enhance overall student success.
- **Limitations:** Despite the promising results, it's essential to acknowledge the limitations of the study. Factors such as data quality, external influences, and the dynamic nature of student behaviour may impact the model's accuracy.
- **Future Research:** Future research endeavours could explore expanding the dataset, incorporating additional features, and refining the model to address identified limitations. Exploring the model's performance across diverse educational contexts would also be valuable.

V. Conclusion And Framework

In conclusion, this research has delved into the realm of predicting at-risk students and enhancing their academic performance through an innovative Hybrid Ensemble Model.

The journey began by exploring existing literature, emphasizing the importance of early identification and the role of machine learning techniques. The study then introduced a novel approach—the Hybrid Ensemble Model, incorporating diverse algorithms and ensemble methods.

By analyzing a newly collected dataset from high school students, the research identified key features, emphasizing the significance of both demographic and academic information. The model's development involved careful consideration of various supervised machine learning algorithms and the implementation of ensemble techniques like Bagging, Boosting, and Stacking.

Results showcased the Hybrid Ensemble Model's superior predictive performance, particularly in early identification of at-risk students. The integration of the model into a web application further enhances its practical utility for teachers.

However, like any research endeavor, this study has its limitations. Challenges in data collection, model interpretation, and generalizability should be acknowledged. Future work could address these limitations, explore additional features, and expand the applicability of the model.

In essence, this research contributes to the ongoing discourse on student success by providing a practical and effective solution for early identification and support. The Hybrid Ensemble Model stands as a promising tool for educators aiming to proactively address the needs of at-risk students, fostering a more supportive and successful educational environment.

The future work in predicting student behavior using machine learning is not only about improving accuracy but also about addressing ethical considerations, ensuring inclusivity, and creating a positive impact on the overall learning experience. Collaboration and a multidisciplinary approach will be key to advancing the field responsibly.

- **Enhanced Personalization**: Machine learning will enable even more personalized learning experiences. Predictive models will adapt content, pace, and teaching methods to individual student needs, optimizing the learning process.
- **Continuous Assessment and Feedback**: Real-time assessment and feedback systems will become more prevalent. These systems will analyse student behaviour as they interact with content and provide immediate insights to both students and educators.
- **Intervention Strategies**: ML models will not only predict at-risk students but also recommend tailored intervention strategies, enabling educators to provide the right support at the right time.
- **Social and Emotional Learning (SEL)**: ML models can be used to detect emotions and social interactions among students. This can aid in fostering emotional intelligence and improving interpersonal skills.

References

[1] J. Y. Chung and S. Lee, ‘‘Dropout early warning systems for high school students using machine learning,’’ Children Youth Services Rev., vol. 96, pp. 346–353, Jan. 2019.

[2] S. Huang and N. Fang, ‘‘Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models,’’ Computer. Educ., vol. 61, pp. 133–145, Feb. 2013.

[3] G. Sujatha, S. Sindhu, and P. Sacariases, ‘‘Predicting students’ performance using personalized analytics,’’ Int. J. Pure Appl. Math., vol. 119, no. 12, pp. 229–238, 2018.

[4] M. S. A. J. Kumar and D. Handa, ‘‘Literature survey on educational dropout prediction,’’ Int. J. Educ. Manage. Eng., vol. 7, no. 2, pp. 8–19, Mar. 2017, doi: 10.5815/ijeme.2017.02.02.

[5] S. Al-Sarem, ‘‘Predictive and statistical analyses for academic advisory support,’’ J. Eng. Technol., vol. 6, no. 2, pp. 304–315, Dec. 2016, doi: 10.21859/jet-060222.

[6] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, ‘‘Detecting atrisk students with early interventions using machine learning techniques,’’ IEEE Access, vol. 7, pp. 149464–149478, 2019.

[7] B. Sekeroglu, K. Dimililer, and K. Tuncal, ‘‘Student performance prediction and classification using machine learning algorithms,’’ in Proc. 8th Int. Conf. Educ. Inf. Technol., Mar. 2019, pp. 7–11. [25] T. Soffer and A. Cohen, ‘‘Students’ engagement characteristics predict success and completion of online courses,’’ J. Comput. Assist. Learn., vol. 35, no. 3, pp. 378–389, Jun. 2019.

[8] M. Hussain, W. Zhu, W. Zhang, S. M. R. Abidi, and S. Ali, ‘‘Using machine learning to predict student difficulties from learning session data,’’ Artif. Intell. Rev., vol. 52, no. 1, pp. 381–407, Jun. 2019.

[9] A. Cano and J. D. Leonard, ‘‘Interpretable multiview early warning system adapted to underrepresented student populations,’’ IEEE Trans. Learn. Technol., vol. 12, no. 2, pp. 198–211, Apr. 2019.

[10] P. Nair, N. Khatri, and I. Kashyap, ‘‘A novel technique: Ensemble hybrid 1NN model using stacking approach,’’ Int. J. Inf. Technol., vol. 12, no. 3, pp. 683–689, Sep. 2020.

[11] K.-W. Hsu, ‘‘A theoretical analysis of why hybrid ensembles work,’’ Comput. Intell. Neurosci., vol. 2017, pp. 1–12, Jan. 2017

[12] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, ‘‘Models for early prediction of at-risk students in a course using

[13] J. Xu, K. H. Moon, and M. van der Schaar, ‘‘A machine learning approach for tracking and predicting student performance in degree programs,’’ IEEE J. Sel. Topics Signal Process., vol. 11, no. 5, pp. 742–753, Aug. 2017.

[14] A. Behr, M. Giese, H. D. Teguim, and K. Theune, ‘‘Early prediction of university dropouts—A random forest approach,’’ Jahrbücher Nationalökonomie Statistik, vol. 240, no. 6, pp. 743–789, Feb. 2020.