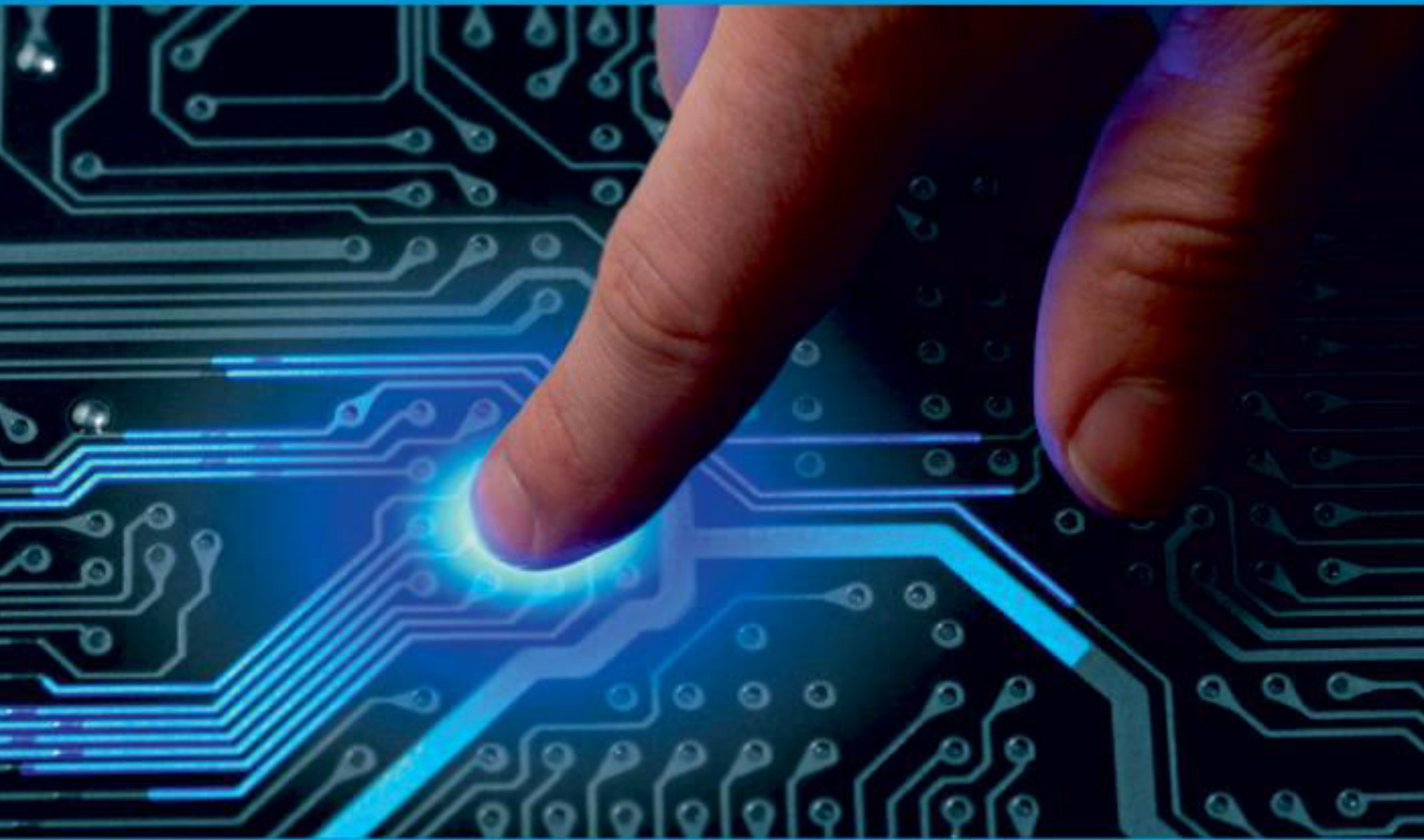
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204052 |

The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure

Gokul M [1], Suthish Kannan T [2], Christon S [3], Prakash R [4], Dr. Charles A [5]

UG Students, Dept. of ECE., Government College of Engineering, Bargur, Tamilnadu, India[1,2,3,4]

Assistant Professor, Dept. of ECE., Government College of Engineering, Bargur, Tamilnadu, India [5]

ABSTRACT: Education is very important for students' future success. The performance of students can be supported by the extra assignments and projects given by the instructors for students with low performance. However, a major problem is that students at-risk cannot be identified early. This situation is being investigated by various researchers using Machine Learning techniques. Machine learning is used in a variety of areas and has also begun to be used to identify students at-risk early and to provide support by instructors. This research paper discusses the performance results found using Machine learning algorithms to identify at-risk students and minimize student failure. The main purpose of this project is to create hybrid model using the ensemble stacking method and to predict at-risk students of each using this model. We used machine learning algorithms such as Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neigbhors, Support Vector Machine, AdaBoost Classifier and Logistic Regression in this project. The performance machine learning algorithm presented in the project was measured with various metrics.

KEYWORDS: Machine Learning Techniques; Early identification; Hybrid Model; Naïve Bayes; Random forest, K-Nearest Neighbor

I. INTRODUCTION

Some students may fail their courses during the semester due to various problems such as psychological reasons, family situation, friend environment or not getting enough support from the teachers. The school success of such students is at risk. Early intervention is required by teachers to identify students at risk and to support the educational status of these students. Early prediction of students' achievement performance can help instructors identify those students who need extra courses, additional assignments, or assistance.

It can be a problem for teachers to analyse the performance of each student in schools with a large student population. In schools with a large student population. If students whose school success is not good can determined early, studies can be started to increase the school success of such students and it can be ensured that such students succeed before it is too late. In our case, the best student group to study for this subject is High School or University students. For now, the best target of the project includes high school students, since school success will also affect the future education life of the students. Thus, with the data set obtained from high school students and containing the academic and demographic information of the students, their failure status can be determined. This project includes high school students studying in Turkey.

Students counted as successful or unsuccessful within the scope of the project were determined according to the rules of the education system. It is determined whether the students are successful or not by looking at the year-end average scores of the students. If the year-end average of the course is 50 and above, the student is considered successful in that course.

If the year-end average of the course is below 50, the student is considered to have failed that course. In addition, according to the education system regulation, students with a year-end general average grade below 50 can pass to the next grade as responsible if they have at most 3 failed courses at the end of the year. According to the specified rules, students who will be unsuccessful at the end of the year must be determined early. However, determining the school performance of all students is a very difficult issue for educators and training places.

II. RELATED WORK

Nowadays, with the development of technology, many studies have been started in the fields of data science and machine learning. Machine learning has become widely used in areas such as predicting students at risk, predicting students' final exam scores, and identifying unsuccessful students early. Identifying students at risk is an important condition for teachers to carry out additional studies to support their performance. Since teachers do not have the appropriate resources to identify such students, machine learning techniques are used. In previous literatures, models were created by using many machine learning algorithms and the accuracy of predictions were discussed. Behr et al. utilized random forest for early pre-diction of university dropouts. Berens et al. used ad- ministrative student data for early detection of students at risk. Lee and Chung worked on improving the performance of dropout prediction. Figueroa-Canas and Sancho Vinuesa checked performance of students in quizzes to predict early dropout. Cano and Leonard concentrated on underrepresented student populations to avoid their dropout by early warning. Finally, dropout from MOOC courses has been tacked by several research groups, Liao et al. developed an approach for predicting low performing students[2] The most important element of this project is the data set. To identify students at risk early before the end of the school, the data of the students with various information should be obtained from a school or platform, as is included in the literature. After the data is obtained, we can train the models created with machine learning techniques and we can compare the performance of the models. The types of data containing the information of students obtained from schools or various educational platforms are also important for the high performance of the models created. Many studies in the literature divided student data as time-varying and timeinvariant. According to research Er, it has been concluded that not using time-invariant 4 data (gender, experience, etc.) obtained from students has no significant effect on overall results [2 results obtained from these studies, we can learn which data of the students will increase the accuracy of the performance in the models created

III. PROPOSED ALGORITHM

In the proposed method, it has been realized that stratified k-fold cross validation and hyper parameter optimization techniques increased the performance of the models. The hybrid ensemble model was tested with a combination of two different datasets to understand the importance of the data features. In first combination, the accuracy of the hybrid model was obtained as 94.8% by using both demographic and academic data. . 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.

Advantage:
Machine learning is used in a variety of areas and has also begun to be used to identify students at-risk early and to provide support by instructors. This research paper discusses the performance results found using Machine learning algorithms to identify at-risk students and minimize student failure. The main purpose of this project is to create a hybrid model using the ensemble stacking method and to predict at-risk students using this model. We used machine learning algorithms such as Naïve Bayes, Random Forest, Decision Tree, K-Nearest Neighbors, Support Vector Machine, AdaBoost Classifier and Logistic Regression in this project.

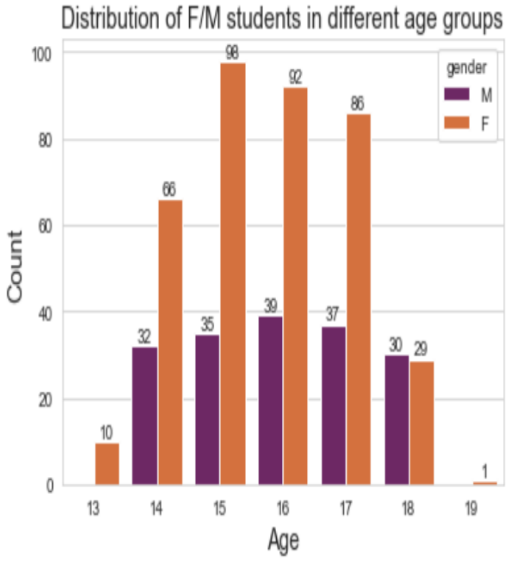
IV. PSEUDO CODE

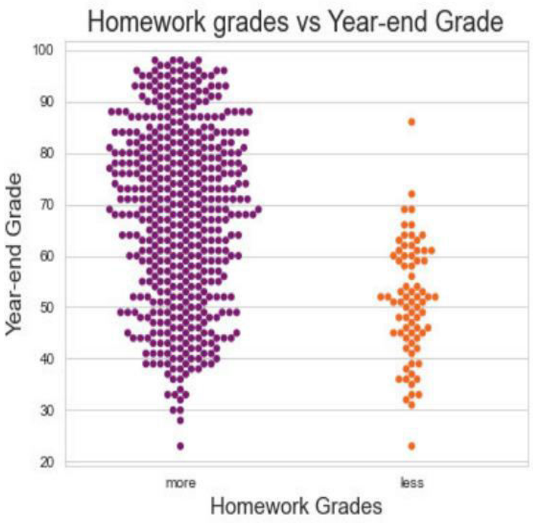
```
Step 1: Gathering  data from third party APIs
Step 2:  Performing feature Engineering
Step 3: Performing feature selection
Step 4: Select the right model
Step 5: Performing Hyperparameter Tunning
Step 6:  Production Deployment
Step 7: End.
```

V. DATA VISUALIZATION

In this project, a hybrid model created using machine learning techniques to predict students at risk was presented. Firstly, the data set was collected from various schools via forms. The data set includes both demographic and academic characteristics of the students. Thus, it can be observed which characteristics contribute to the students' performance. There are a total of 555 students and 38 features in the data set.

In the data visualization section, various graphs and tables are visualized for a better understanding of the data characteristics in the data set. To observe which characteristics of the students affected their grades, various student characteristics were compared and visualized with graphs and tables. If we look at the distribution graph showing gender and age range in Figure 4, female students are more and 15-year-old students are the majority in the dataset. In addition, the age range of the students in the data set is between 13 and 19. The 21 effect of the knowledge of whether students want to go to university in the future or not on the year-end averages can be observed with the boxplot in Figure 5. The box plot shows the minimum, maximum, median, and values in the first 25% and third 75% quartiles of the compared features of a dataset. Data points marked outside the box plot are defined as outliers. Based on the boxplot in Figure 5, it can be said that students who want to go to university in the future have higher year-end average scores. It can be understood that the students' setting such a goal for themselves has a positive effect on their course grades. Assignments given to students are considered important for course work by teachers. For this reason, the effect of homework grades on students' end- ofyear averages can be observed in Figure 6. Students with 50 or more homework grades have higher year-end average scores than students with lower homework grades In addition, another important point is that students with low homework grades have low end-of-year average scores, so homework grades can be an important factor in determining students at risk. Figure 7 shows the histogram plotting of the students' end-ofyear average scores. Most students have a year-end average of between 40 and 80.

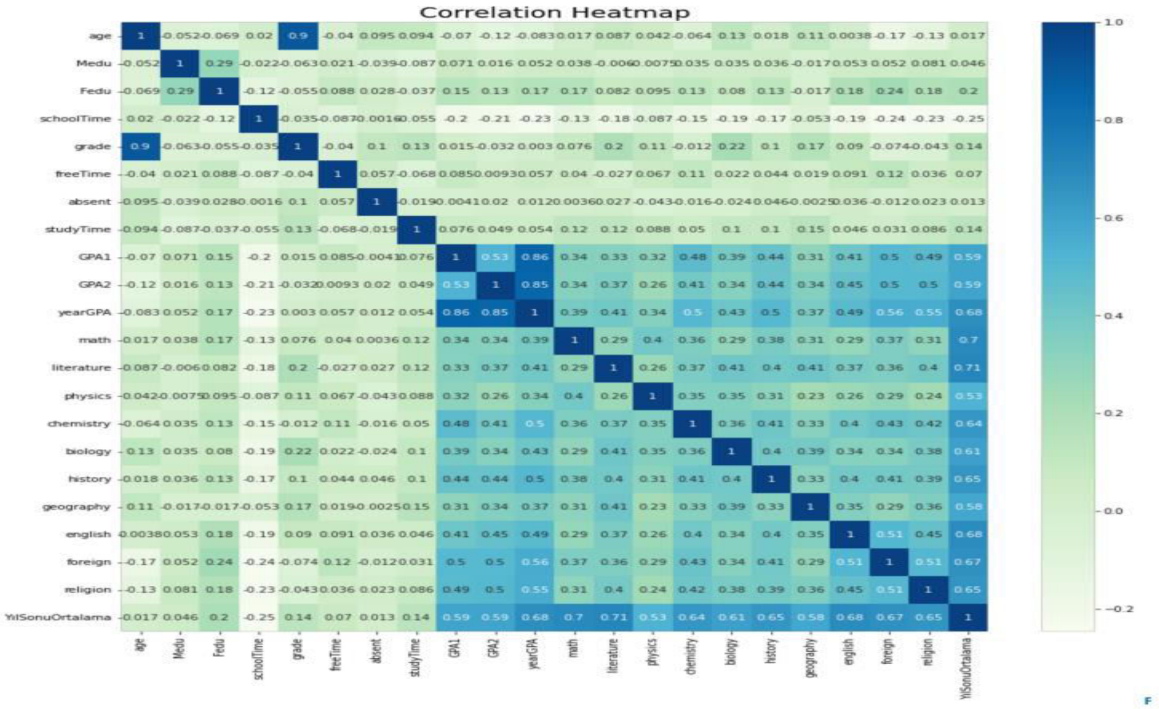




| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204052 |



VI. CONCLUSIO AND FUTURE WORK

For future school success of students to increase positively, the students who will fail should be identified early by the teachers. If the students who will be unsuccessful can be identified early, additional studies can be provided to these students by the teachers. Thus, the school success performance of the students in this situation can be increased. The aim of this project is to predict students who are at risk early before their school term ends. To solve this problem, it has been suggested to use machine learning techniques in the literature. In this project, the creation of a hybrid model with the supervised Machine Learning algorithms is presented as a solution. In the studies in the literature, various Machine Learning algorithms have been applied on data sets and the models have been evaluated individually. However, unlike other studies, a hybrid ensemble model was created with a stacking approach to predict students at risk in this study. The data set containing the high school students' information was obtained through the form. Both demographic and academic characteristics of the students are included in the data set obtained. For a system to work properly in our own education system, Turkey-specific features must be taken into account. For this reason, course grades have been collected according to the education system of Turkey. Firstly, performance results were obtained when models were evaluated individually. According to the results obtained, the use of the Stratified K-Fold Cross Validation method had a positive effect on the performance of the models. It is also observed that performing hyperparameter optimization increases the performance of the models. Thus, it can be said that using stratified 10-fold cv and performing hyperparameter optimization improves the performance of machine learning models. According to the results of the individually evaluated models, the model with the best accuracy value of 94.4% is Logistic Regression. A hybrid model was created according to the proposed method. The hybrid model was created with the stacking method and different supervised algorithms were tried as a meta learner.

REFERENCES

[1] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, ''A machine learning framework to identify students at risk of adverse academic outcomes,'' in Proc. 21st ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Aug. 2015, pp. 1909–1918, doi: 10.1145/2783258.2788620.
[2] H. Agrawal and H. Mavani, ''Student performance prediction using machine learning,'' Int. J. Eng. Res., vol. 4, no. 3, pp. 111–113, Mar. 2015, doi: 10.17577/ijertv4is030127.
[3] L. A. B. Macarini, C. Cechinel, M. F. B. Machado, V. F. C. Ramos, and R. Munoz, ''Predicting students success in blended learning—Evaluating different interactions inside learning management systems,'' Appl. Sci., vol. 9, no. 24, p. 5523, Dec. 2019, doi: 10.3390/app9245523.

[4] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, ‘‘Predicting student performance using personalized analytics,’’ Computer, vol. 49, no. 4, pp. 61–69, Apr. 2016, doi: 10.1109/MC.2016.119.

[5] I. E. Livieris, K. Drakopoulou, V. T. Tampakas, T. A. Mikropoulos, and P. Pintelas, ‘‘Predicting secondary school students' performance utilizing a semi-supervised learning approach,’’ J. Educ. Comput. Res., vol. 57, no. 2, pp. 448–470, Apr. 2019, doi: 10.1177/0735633117752614.

[6] N. Mduma, K. Kalegele, and D. Machuve, ‘‘Machine learning approach for reducing students dropout rates,’’ Int. J. Adv. Comput. Res., vol. 9, no. 42, pp. 156–169, May 2019, doi: 10.19101/IJACR.2018.839045.

[7] H. Yates and C. Chamberlain. (2017). Machine learning and higher education: EDUCAUSE. Educause. [Online]. Available: https://er.educause.edu/ articles/2017/12/machine-Learning-and-higher-education

[8] S. B. Kotsiantis, C. J. Pierrakeas, and P. E. Pintelas, ‘‘Preventing student dropout in distance learning using machine learning techniques,’’ in Proc. Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst., in Lecture Notes Artificial Intelligence: Subseries Lecture Notes Computing Science, vol. 2774, 2003, pp. 267–274, doi: 10.1007/978-3-540-45226-3_37.

[9] S. Rani and N. S. Gill, ‘‘Hybrid model for Twitter data sentiment analysis based on ensemble of dictionary based classifier and stacked machine learning classifiers-SVM, KNN and C5.0,’’ J. Theor. Appl. Inf. Technol., vol. 98, no. 4, pp. 624–635, 2020.

[10] K. T. Chui, R. W. Liu, M. Zhao, and P. O. D. Pablos, ‘‘Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine,’’ IEEE Access, vol. 8, pp. 86745–86752, 2020, doi: 10.1109/ACCESS.2020.2992869.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

☎ 9940 572 462 ✆ 6381 907 438 ✉ ijircce@gmail.com

www.ijircce.com

Scan to save the contact details