IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 December 2022, accepted 24 December 2022, date of publication 28 December 2022, date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232984

RESEARCH ARTICLE

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

REYHAN ZEYNEP PEK[1], SIBEL TARIYAN ÖZYER[2], TAREK ELHAGE[3], TANSEL ÖZYER[2], AND REDA ALHAJJ[1,4,5], (Senior Member, IEEE)

[1]Department of Computer Engineering, Istanbul Medipol University, 34810 Istanbul, Turkey
[2]Department of Computer Engineering, Ankara Medipol University, 06050 Ankara, Turkey
[3]ABC Private School, Abu Dhabi, United Arab Emirates
[4]Department of Computer Science, University of Calgary, Calgary, AB T2N 1N4, Canada
[5]Department of Heath Informatics, University of Southern Denmark, 5230 Odense, Denmark

Corresponding author: Reda Alhajj (alhajj@ucalgary.ca)

This work was supported in part by The Scientific and Technological Research Institution of Turkey (TUBITAK) under Program Grant 2209-A.

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 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. The performance of each machine learning algorithm presented in the project was measured with various metrics. Thus, the hybrid model by combining algorithms that give the best prediction results is presented in this study. The data set containing the demographic and academic information of the students was used to train and test the model. In addition, a web application developed for the effective use of the hybrid model and for obtaining prediction results is presented in the report. In the proposed method, it has been realized that stratified k-fold cross validation and hyperparameter 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 second combination, when only academic data was used, the accuracy of the hybrid model increased to 98.4%. This study focuses on predicting the performance of at-risk students early. Thus, teachers will be able to provide extra assistance to students with low performance.

INDEX TERMS At-risk students, classification, dropout prediction, hybrid model, machine learning techniques, stacking ensemble model, student performance prediction.

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 [3].

It can be a problem for teachers to analyze the performance of each student in schools with a large student population. If students whose school success is not good can be

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

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 [1], [2], [3], [4]. 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. The reason for this difficulty may be due to the large student population and the lack of sufficient resources [5]. For this reason, it is necessary to use different techniques to identify students who are in a risk situation.

To solve this problem, many researchers in the literature have stated that machine learning techniques give useful results, e.g., [6], [7], [1], [8], [2], [5], [9], and [10]. Machine learning is based on the idea of modeling by identifying useful information from the data to help with problems [12]. Machine learning techniques, which have been widely used in different fields, have been used in studies in the literature within the subject of identifying students at risk, e.g., [3], [11], [5], [9], and [12]. However, unlike the methods followed to solve this problem in previous studies, it was planned to create a hybrid model with machine learning techniques in this project. Supervised learning algorithms have been considered for the creation of the hybrid model. It is intended to use Naive Bayes, Support Vector Machine, Random Forest, Decision Tree, K-Nearest Neighbor, Logistic Regression and AdaBoost algorithms to obtain the initial results. The accuracy values of the algorithms to be used were measured with various metrics and the algorithms that provide the best performance were determined. With this new method developed for this project topic, at-risk students will be identified with high accuracy. As a result, it enables the hybrid model to produce better prediction results compared to the individual models. Bagging, Boosting and Stacking techniques are ensemble methods used to combine several machine learning techniques. Bagging refers to bootstrap aggregation. It is an ensemble method used to increase the accuracy of classification models. Boosting is another ensemble method that uses a weighted average to produce strong learners from weak learners. Stacking is an ensemble method that combines multiple models of different types and uses a meta classifier [13]. One of the purposes of this project is to create a hybrid model by combining multiple supervised machine learning algorithms. For this purpose, the stacking ensemble learning approach is quite suitable to be used. Stacking can be used to combine multiple classification and regression models. The architecture of the stacking method consists of two phases. Phase 1 is the base model. From the base models, the one which give the best prediction results that fit the training data are determined. In Phase 2, the meta-classification model is decided; it is the model that learns to best combine the predictions of the base models. Stratified k-fold cross validation of the base models is an effective way to prepare the training dataset for the meta model. The final estimation results are obtained as output from the developed hybrid model.

The main aim of this project is to identify high school students at risk before the end of the education period and to support the education of high school students. The purpose of our research is to increase the success performance of students, as well as to identify students who may fail in the class before the end of the semester and to provide timely support by the teacher to such students. Teachers can be informed about students at risk (students who may fail) as determined by the hybrid model and additional study material may be provided to these students by the teachers. By analyzing the characteristics of the data set obtained from the students, the characteristics that affect the school performance of the students can be determined. Thus, more efficient results can be obtained by using these data while creating a hybrid model. The data set used in this research was newly collected from high school students by having them completing a questionnaire. With the created form, a data set containing demographic and academic (school marks) data features was obtained from participated students. According to the conducted analysis, it was observed that academic data contributed more to the identification of students at risk. Before the hybrid ensemble model was created, individual models were created and evaluated with supervised learning algorithms. In this step, algorithms that give better results were observed and their results were compared in the paper. Depending on the results obtained, the hybrid ensemble model was developed. The stacking approach was used to create a hybrid model and the contribution of the obtained results to the research was presented.

The contribution of this research includes the developing of the hybrid ensemble model with the stacking approach, creating and evaluating various supervised learning algorithms, collecting a new dataset from high school students, discovering important data features that affect students' school performance by analyzing the newly collected data set, developing the web application, and integrating the hybrid model into the developed web application. In line with the aim of the

research, with the integration of the developed hybrid model into the web application, it can help teachers to identify students at risk more effectively. Consequently, the developed hybrid ensemble model can help to solve the problem of identifying students at risk. At the end of the study, the hybrid model gave higher prediction performance and successfully identified students at risk.

The rest of this paper is organized as follows. Background and literature review is presented in Section II. The operational stages of the project are explained in the experimental setup which is included together with methods in Section III. In Section IV, the performance results obtained from the applied methods are presented and discussed. Section V includes limitations of the work. Section VI is the conclusion and future work.

II. BACKGROUND AND LITERATURE REVIEW

A. SOLUTIONS FOR PREDICTING THE AT-RISK STUDENTS AND THEIR PERFORMANCE

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 [12], [5], [9]. In previous literatures, models were created by using many machine learning algorithms and the accuracy of predictions were discussed.

Identifying students at-risk and handling them properly has received considerable attention from the research community, e.g., [18], [19], [7], [20], [21], [22], [23], [24], [25], and [26]. For instance, Adnan et al. [15] developed an approach which allows for early intervention by predicting at-risk students during the course offering. Agrawal and Mavani [6] described a method which employs machine learning techniques for student performance analysis. Akçapçnar et al. [27] analyzed eBook interaction logs to identify at-risk students as early as possible.

Behr et al. [28] utilized random forest for early pre-diction of university dropouts. Berens et al. [29] used ad- ministrative student data for early detection of students at risk. Lee and Chung [30] worked on improving the performance of dropout prediction. Figueroa-Canas and Sancho-Vinuesa [31] checked performance of students in quizzes to predict early dropout.

Burgos et al. [32] developed a dropout preventive approach by employing data mining techniques for modeling students' performance. Cortez and Silva [1] applied data mining techniques to predict secondary school student performance. Costa et al. [33] evaluated the effectiveness of data mining techniques for early prediction of students' academic failure in introductory programming courses. Mueen, Zafar and Manzoor [34] applied data mining techniques for modeling and predicting students' academic performance. Gupta and Sabitha [35] studied student retention by applying data mining techniques. Wolff et al. [36] analyzed clicking behavior to predict students at risk and improve retention. Park et al. [37] also analyzed click streams for behavior detection.

Chung and Lee [38], [39], [40] handled dropout among high school students. Huang and Fang [41] compared our types of predictive mathematical models for predicting student academic performance in an engineering dynamics course. Hung et al. [42] employed time-series clustering method for identifying at-risk students which provides the opportunity for early interventions. Hussain et al. [43] utilized machine learning techniques to analyze learning sessions data to predict student difficulties leading to dropout.

Cano and Leonard [44] concentrated on underrepresented student populations to avoid their dropout by early warning. Finally, dropout from MOOC courses has been tacked by several research groups, e.g., [45], [46], [47], [48], and [49]. Liao et al. [50] developed an approach for predicting low performing students.

Chui et al. [14] proposed the use of the improved conditional generative adversarial network based deep support vector machine algorithm. With the developed method, researchers tried to predict the performance of students under supportive learning. They generated new training data with improved CGAN and emphasized the importance of generating new data for the models. As a result, 0.968, 0.971 and 0.954 values from specificity, sensitivity and AUC metrics were obtained, respectively, with the model they developed using 10-fold cross validation.

The success rate of predicting students at risk early depends on the characteristics of the data set used and the algorithms with high performance. Macarini et al. [7], aimed to find students at risk by obtaining the best combination using various data sets and classification algorithms. In this study, models developed with classification algorithms (Naive Bayes, Random Forest, AdaBoost, Multilayer Perceptron, K-Nearest Neighbor, Decision Tree) were tested on student data from "Introductory Programming Courses" taken from the Moodle platform. Among the algorithms tested on 13 data sets with various data features, the AdaBoost algorithm has yielded very good results, especially in the 2nd, 5th, and 12th data sets. To identify students at risk before the end of the term, they obtained very suitable results to determine students' performance before the end of the school with the student data used until the 8th week. In addition, the results of the questionnaires made to the students to measure the effect on the models were also included in the data set, but no increase in the performance of these models was observed. During the experiment, they removed the Decision Tree algorithm due to the over-fitting problem. When the cognitive, social, and teaching presence element were used, it has been observed that these features did not contribute to the performance of the models.

The AdaBoost algorithm was also used by Lakkaraju et al., in his studies [5]. The subject discussed in this study is to

identify students who most need the programs developed by teachers in order for students to graduate on time. Since some schools do not have enough budget and resources, assistance cannot be provided to students in every risk situation. Considering this situation of the schools, it is necessary to ensure that the students in the risk situation determined by the models developed by the researchers are ranked from the highest risk to the least risk situation. The collected data work on a very large data set consisting of 200,000 students from 2 different schools. The Random Forest algorithm performed better in identifying at-risk students early. In addition, it has been observed that AdaBoost algorithm and Decision tree algorithm show poor performance in models developed with data from District A. In addition, researchers have observed that classification models were particularly mistakes on data points where some aspects of students were positive, and others were negative. Therefore, we proposed using cross validation as it is an important robust method for training models with different data points.

Another method, the Naive Bayes algorithm, has been used by many studies by developing models on this subject. Agrawal et al., states that the Naive Bayes algorithm performs well in predicting the performance of students, and this will be the same for Neural Networks [6]. The model created in this study was tested using the data of 80 students from the 3rd semester to the 6th semester. As a result, it has been observed that the Neural Networks algorithm performs better than the Bayes classifier. As the size of the trained data increases, the performance accuracy of the neural networks algorithm also increases. When the size of the trained data set consisted of 70 people, the accuracy of the performance of the Neural Networks algorithm increased to 70%. Researcher Er who developed a model using the Naive Bayes algorithm and other machine learning algorithms, showed in his study that combining many algorithms and evaluating their performance would yield better results than individual algorithms [2]. In addition to the K-star, Naive Bayes and C4.5 algorithms, these algorithms were combined to create 3 decision schemes and the performance of the algorithms was evaluated with overall accuracy, sensitivity, and precision. In the first step, decision scheme 3 reached the best performance with 65% prediction rate. In step 2, decision scheme 3 yielded a very good 75% performance. However, decision scheme 2 showed a high performance of 78% at this step. In the last step, decision scheme 1 showed the best performance at 85%. The K-star algorithm, on the other hand, yielded a good 82% among the other algorithms with individual performance. As can be seen from these results, it is understood that combining algorithms and creating models will provide a better performance for predicting at-risk students. It was observed that the results of the individually used algorithms were lower than the results of the combined model. Therefore, we proposed to predict students at risk by combining more than one algorithm.

Another study that creates decision schemes by combining algorithms has been done by Lykourentzou et al. Support Vector Machine (SVM), Feed-forward neural networks (FFNN), probabilistic ensemble simplified fuzzy ARTMAP (PESFAM) and three decision schemes created by combining these algorithms were used in their studies [4]. During the e-learning course, it was planned to estimate the specific dropout and the records of the daily actions of the students were considered. As it was previously obtained in the literature, in this study, it was concluded that decision schemes performed well when the performance of the models was measured. When the performance of the model created using decision scheme 1 was measured with 3 methods, it came to a very high result, on average between 95% and 100%. As can be seen from this research, single machine learning techniques do not provide accurate predictions for each different data set.

One of the algorithms used to predict students' performance and determine the factors affecting their success is Random Forest (RF). This technique is used with Decision Trees, Neural Networks and SVM algorithms in the study by Cortez and Silva to measure their performance to find the best algorithm [1]. The aim of this study is to predict students' success and determine the factors that affect their success positively or negatively. The data of the students obtained for 2 different courses were used. As a result, the RF algorithm and Decision Tree algorithm performed better than other algorithms, and it was concluded that the SVM algorithm and the Neural Network algorithm were more affected by irrelevant data. In the study conducted by Sujatha, Sindhu and Savaridassan, another research using Random Forest, SVM and Decision Tree algorithms, Multi-Linear Regression algorithms were also used additionally [16]. In this study, it was aimed to predict the performance of the students by using the personal data of the students obtained from the education database. 70% of the data set containing 3-year data of 3000 students was used for train and the rest was used for testing algorithms. As a result, the RMSE (error) values of the Multi- Linear and SVM algorithms were obtained as 4.8 and 4.3, respectively. As can be seen from this result, it is concluded that the SVM algorithm is the best performing algorithm with the most appropriate error result. However, when compared with the other article [1], in this study, the RMSE value of the RF algorithm was 5.1 and the RMSE value of the Decision Tree algorithm was 5.2. In other words, error values are higher than other algorithms.

Another study using Random Forest algorithm was done by Elbadrawy et al. Factorization Machine (FM), Personalized linear multi-regression (PLMR) algorithms are also used in their research [8]. They applied the models they created on the data sets obtained from 4 different schools and platforms. Then the performance of each algorithm was measured. As a result, they stated that the grades of the students in the next semester can be predicted with low error rates by using PLMR and MF techniques in addition to the existing information such as transcript data of the students with the models they developed. Also, the Course-specific regression (CSpR) used performs better than other methods for most courses, with a

RMSE ratio of 0.632. In another study, the Random Forest algorithm was used, and very good performance results were obtained. Adnan et al., aimed to analyze the problems faced by the students at-risk and to provide a prediction model for educators [15]. This study was conducted on data obtained from online learning platforms. They trained and tested the predictive model using various machine learning and deep learning algorithms to characterize students' learning behaviors. Along with the Random Forest algorithm, SVM, K-NN, Extra Tree Classifier, AdaBoost Classifier, Gradient Boosting Classifier, and ANN algorithms are also used to train demographic data.

As a result, even at 20% of the course length, the Random Forest model exhibited an accuracy performance of 79%. As the course length increased towards 100%, the accuracy per- centage increased to 91%. Overall, in this study, the results of the RF prediction model made it possible to predict the performance of students at risk early with high accuracy [15].

Machine Learning has an important role in extracting information from data sets and analyzing this information. It has 3 basic and common uses. These are Supervised, Semi supervised and Unsupervised learning. In the literature, there are quite a few studies using these methods. According to Livieris et al., one of these studies, they examine the effectiveness of semi-supervised algorithms to predict students' performance in final exams [9]. It is concluded that the classification accuracy can be significantly improved with the data used to develop reliable and performance-accurate prediction models of semi-supervised techniques according to their numerical values. The performance of the C4.5 algorithm, which is one of the techniques used, has shown the best classification performance with two data sets obtained by YATSI (Yet Another Two Stage Idea) method. In addition, the YATSI method has further increased the performance of Naive Bayes and Multilayer Perceptron algorithms. A very high increase from 1.5% to 4.0% was observed. But they stated that the YATSI method cannot always improve the performance of the Naive Bayes algorithm if unlabeled data and labeled data are used together.

Another study using supervised learning classification techniques, evaluated the performance of 4 algorithms containing 18 features and used on 61340 samples [10]. In short, the purpose of this study is to focus on the dropout problem of students using machine learning techniques and to estimate this rate. They used Logistic Regression (LR), KNN, Random Forest and Multilayer Perceptron (MLP) algorithms. As a result, they observe that the LR algorithm has the best performance as 0.3375 when compared to other algorithms and when looking at the RMSE value.

B. FEATURES IN DATASETS THAT AFFECT PREDICTING STUDENTS' AT-RISK

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 time-invariant. According to research Er, it has been concluded that not using time-invariant data (gender, experience, etc.) obtained from students has no significant effect on overall results [2]. Likewise, according to Lykourentzou et al., it was observed that student data that did not change over time (time-invariant) had less accuracy in predicting student performance compared to data that changed over time (time-varying) [4]. According to the results obtained from these studies, we can learn which data of the students will increase the accuracy of the performance in the models created. It is known that models created with data sets containing student's past exam results are more efficient [1].

The ability to predict the performance of students at risk before the end of the semester is great importance for additional studies to be created by instructors and for students to improve their success performance. However, if students in a risk situation can be identified early, they can be intervened, and the failure rate of students will decrease. To determine such students before the end of the term, the use of data up to half of the term will affect the prediction model. For this reason, the model can show a good prediction performance by using the demographic data of the students since the beginning of the academic year [11]. To increase the accuracy of students' performance prediction, social and cultural characteristics, past exam results, project and home- work results can be added to the data set features [9]. In another literature study, the features that positively affect the performance of the model were determined as students' grades, gender, family structure and family occupation [17].

Machine learning algorithms and their results, which are used by many researchers working on this subject, have been discussed in the literature review. Naive Bayes, Neural Networks, Decision Trees and Random Forest algorithms are observed as some of the Machine Learning algorithms used by the researchers and as a result, it was found that they performed best. In a study, it was observed that the Naive Bayes algorithm performed 87% well in predicting student performance [18]. Another algorithm that provides good performance results is Neural Networks. Neural Networks can perform several classification tasks simultaneously. In studies on the subject of predicting the performance of students, it has been observed that it performed 70% and above. In another study, compared to the Naïve Bayes algorithm, it was concluded that the performance accuracy was the same or slightly better [6]. Decision Tree and Random Forest are among the machine learning algorithms used in the solution of many problems. It was also used by some researchers in the problem of predicting students' achievement performance. Random

Forest application has shown a good performance in determining risky students [5].

As can be observed from the literature review, the subject of predicting students at risk situation gives satisfactory results with machine learning algorithms. The performance of the model to be created depends on both the accuracy of the algorithms in predicting and the characteristics of the data set. In particular, the data set containing the exam results of the students plays a huge role in predicting the students in risk, as observed in the literature. The aim of this study is to identify students at risk and to do this before the end of their academic term. If this happens, support can be provided to students in this situation by the instructors. With the hybrid model created, not only will this problem be answered, but also the instructors will be given opportunities to provide assistance to such students, and the students will be provided with studies that will increase their own success performance.

C. HYBRID ENSEMBLE MODEL

Individual evaluation of machine learning techniques is widely used by researchers in their studies. The determined algorithms are trained and tested using the same data set. The algorithm that provides the highest accuracy performance is selected and they continue to work with the model created with the individual algorithm for final studies. However, it has been observed that the hybrid model created by combining more than one machine learning algorithm provides higher accuracy performance than the individual model [52]. Ensemble learning technique provides better performance in various machine learning algorithms. It uses a combined combination of multiple classifier algorithms to improve classification accuracy [13]. Thus, ensemble learning improves the performance results of machine learning algorithms by combining multiple models.

In this project, it was planned to create a hybrid ensemble model using the stacking method [51], unlike other studies on the same topic. Thus, hybrid models are expected to give better accuracy than individual models. The results of the models created with individual algorithms and the performance accuracies of the hybrid ensemble models created by combining the same algorithms are also presented in the report for comparison.

III. EXPERIMENTAL SETUP AND METHODS

A. DATA COLLECTION

In this project, it was planned to use the data set of high school students studying in Turkey. The dataset was collected from high school students in Turkey. Permission was obtained from the ethics committee at Istanbul Medipol University 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 data set was collected by considering the rules of the Turkish education system and the different characteristics of the students. 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 various high schools. 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. The academic and demographic characteristics of the students, detailed descriptions of the data features and data types included in the data set are listed in Table 1. While obtaining the actual data of the students, the features suitable for the education system of Turkey were taken into account. A system that existed before or was made for another country cannot work properly in Turkey's education system. This is due to the fact that each country or culture has its own characteristics. For the system to work properly, it is important that these features are adapted to the educational system.

B. DATA PREPROCESSING AND FEATURE ENGINEERING

For 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 38 features in total in the data set.

There are high schools that provide education in different fields and the courses given by each high school are not the same. The courses in the data set are Mathematics, Literature, Physics, Chemistry, Biology, History, Geography, 1st foreign language, 2nd foreign language and Religion. 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

TABLE 1. Dataset features, descriptions and data types.

Dataset Features		
Students' academic feature	**Description and data type**	**Information contained in the data**
Name of the school	The name of the school where the student is studying (categorical)	Names of the high schools
Status of attending kindergarten	Information on whether the student attends kindergarten (binary)	Yes or No
Travel time to school	Students' travel time from home to school (numerical)	Less than 1 hour, 1 hour, 2 hour, 3 hour or more than 3 hour
University plan	The student's desire to go to university in the future (binary)	Yes or No
Class grade	Information of which class the student is studying (numerical)	9, 10, 11 or 12
Additional support for education	The student's information of whether to take additional education support or not (binary)	Yes or No
Failed a course before	Information whether the student has failed a course before (binary)	Yes or No
Number of absences	The total number of absences the student has made at school in a semester (numerical)	0 to 15
Study time	Student's study time at home (numerical)	1 to 4
GPA of semesters	Student's end of the year average scores for 1st, 2nd semesters and overall average grades (numerical)	0 to 100
Course scores	Student's course exam grades includes math, literature, physics, chemistry, biology, history, geography, english, second foreign language, and religion culture courses. (numerical)	0 to 100
Average scores of assignments (or projects)	Student's average homework/project scores (categorical)	Less than 50 or Equal to or more than 50
Is there a project being developed?	Information whether the student has been involved in an important project before (binary)	Yes or No
Future job	Profession that the student wants to be in the future (categorical)	Job names
Students' demographic feature	**Description and data type**	**Information contained in the data**
Gender	Students' gender (categorical)	Female or male
Age	Students' age (numerical)	13 to 19
City	City name where the student lives (categorical)	City names
District	District name where the student lives (categorical)	District names
Number of siblings	Number of siblings the student has (categorical)	Less than 3 or More than or equal to 3
Education status of mother	Education level of the students' mother (categorical)	Primary school, middle school, high school or university
Education status of father	Education level of the students' father (categorical)	Primary school, middle school, high school or university
Occupation of mother	Occupation of the students' mother (categorical)	Job names
Occupation of father	Occupation of the students' father (categorical)	Job names
Social activity	Information of the social activity in which the student participates (binary)	Yes or No
Internet accessibility	Information of the student's ability to access the internet (binary)	Yes or No
Free time	Students' free time after school (numerical)	1 to 4

averages of the students is given below.

$$\text{Year-end GPA} = \frac{\text{Total weight grade}}{\text{Total course hours}} \quad (1)$$

$$\text{Total weight grade} = \frac{\text{Average grade of the course}}{\text{Weekly time of the each course}} \quad (2)$$

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.

Data visualization is one of the necessary steps to better understand the data in the data set. The Seaborn data visualization library in the Python programming language was used for this purpose. The purpose of the data analysis in this section is to determine the data properties that should be used to train the created model. The results and evaluation of the graphics obtained as a consequence of data visualization are included in Section IV.

Future engineering is used to organize variables in the data set and derive new variables in the machine learning process. Feature engineering improves the prediction performance of the model. It has been observed that there are both numerical and non-numeric features in the data set. Non-numeric properties can affect the prediction performance of the model to be created. Therefore, non-numeric values should be converted to numeric values. This situation can be solved in 2 ways. First, the Label Encoding method can be used. The second method is One Hot Encoding. Label Encoding method assigns a unique number to each data. But this can create a problem. When a data is assigned a high number, it can be given high priority when training the model. Another method, One Hot Encoding method, can be used to solve this problem. This method creates a new row for each category and uses binary values (0 or 1) when assigning values. To avoid any problems later, some non-numeric data were converted to numeric values using One Hot Encoding method.

Thus, the properties of categorical data were transformed into numerical values without changing. In the dataset, which contains the data of 555 students, there are categorical features as yes/no and otherwise. As a result of these processes, all non-numeric data were also converted to numeric values.

There are different kinds of numerical values in the dataset. Examples of these values are age, class information, family education status and course grades.

These data have different values from each other. Therefore, the model may give more importance to the highest number when making predictions, which may lead to an incorrect result. Therefore, feature scaling was applied to prepare the dataset for use in the model. The purpose of using Standardization is to ensure that each feature is equally important.

C. BUILDING THE MODEL FOR INITIAL RESULTS

In this study, Random Forest, K-NN, Decision Tree, SVM, Naive Bayes, Logistic Regression and AdaBoost supervised machine learning algorithms 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. Stratified K-Fold Cross Validation is a variation of K-Fold Cross Validation. It performs stratified sampling rather than random sampling. Which means that it splits the data into k-folds like k-fold cv but ensuring that each fold is appropriate representative features from the original data. In this project, the k value was set to 10. It is also expected to give a more accurate prediction performance.

1) LOGISTIC REGRESSION

Logistic regression does binary classification. In our case, the target feature is the ''pass'' column. Here, there are two possibilities, (1) student will be successful (represented by 1) and (2) student will fail (represented by 0). The logistic regression function is given below. In Equation 3,

(i) p is the probability of the target event, (ii) the set $\{x_1, x_2, \ldots x_n\}$ represents the independent variables, (iii) the set $\{b_1, b_2, \ldots, b_n\}$ represents coefficients of the logistic regression, (iv) b_0 represents the bias (or intercept) term. The output value will be modeled as binary value (1 or 0).

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + \ldots + b_nx_n \tag{3}$$

$$p = \frac{e^{(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}}{1+e^{(b_0+b_1x_1+b_2x_2+\ldots+b_nx_n)}} \tag{4}$$

2) K-NEAREST NEIGHBORS

After hyperparameter adjustments, we used Euclidean as the distance metric parameter for KNN algorithm. We can write the Euclidean distance formula as follows.

$$d(x,y) = \sqrt{(x_1-y_1)^2 + \ldots + (x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{5}$$

According to the Euclidean distance, x and y represent the two points in the coordinate plane. Consequently, we can write the K-Nearest Neighbors as follows. In Equation 6, (i) X is our dataset features, (ii) k is the number of neighbors, (iii) Y is our target class, (iv) R represents the set of observations of the k nearest points, (v) we can write indicator variable as $I(y_i = j)$.

$$P(Y=j|X=x) = \frac{1}{k}\sum_{where\ i\ \in R} I(y_i=j) \tag{6}$$

3) DECISION TREE

This algorithm is one of the supervised learning algorithms. It is a model that we use to classify students according to their school performance status (successful or unsuccessful). After making the parameter adjustments, entropy criterion was used as parameter. Entropy is the impurity measurement and calculated based on feature Y. In equation 7, X represents the target variable (pass column), Y represents the feature in the dataset and p_i represents the probability of target i at the node. It can be mathematically expressed as follows.

$$E(X,Y) = \sum_{i=1}^{n} -p_i log_2(p_i) \tag{7}$$

4) NAÏVE BAYES

We can use Bayes theorem along with the conditional probability formula to calculate the probability of an event to occur. In our case, this event can represent whether students are successful or not. In equation 8, X is our vector of features and Y is class variable. The way this algorithm works is by calculating the probability of each event for a variable and classify that variable according to the highest probability outcome. The following expression of Bayes theorem calculates the probability of event Y occurring when event X occurs.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{8}$$

P(Y|X) and P(X|Y) = Conditional probability.
P(Y) = Prior probability of the class variable.
P(X) = Prior probability of the predictor.

5) RANDOM FOREST

Depending on the features in the data set, classification is conducted to predict whether the student will be successful or not. The random forest algorithm classifies by employing a combination of multiple decision trees and chooses the decision tree that gives the best outcome as the prediction result. The entropy function (Equation 7) was chosen to measure the probability of a specific outcome. Furthermore, the random forest model for our problem may be expressed as follows.

$$D = \begin{bmatrix} x_{1,1} & x_{2,1} & \ldots & x_{m,1} & y_1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_1,n & x_{2,n} & \ldots & x_{m,n} & y_n \end{bmatrix}$$

Here, (i) D represents the training dataset, (ii) the variables $\{x_1, x_2, \ldots, x_m\}$ represent data features in the dataset, (iii) There are n samples of each feature, (iv) the variables $\{y_1, \ldots, y_n\}$ represent the class (target) label. From this set, we can get sample random subsets including data samples. The random forest algorithm performs decision tree combinations with each created subset. As a result, the best output is considered based on the classification outcome.

6) ADABOOST

The AdaBoost classifier was used as another machine learning algorithm. This algorithm combines the weak learners with importance weights to get a strong learner. Each time, the training data is updated based on the data points and it is used in the next learner model. In Equation 9, where (i) α is the classifier coefficient (applied weight), (ii) M is weak classifiers, (iii) $y_m(x)$ is weak classifier outputs.

$$Y(x) = sign\left(\sum_{m=1}^{M}\alpha_m y_m(x)\right)$$
$$where\ \alpha_m = \frac{1}{2}\ln\left(\frac{1-\epsilon_m}{\epsilon_m}\right) \tag{9}$$

7) SUPPORT VECTOR MACHINE

SVM is a supervised learning algorithm which can be used to solve the classification problems. To conduct the classification, a linear line is drawn that separates the sample classes and then the optimal hyperplane is found. It classifies the sample data points by deciding which class they belong to. The SVM model can be expressed mathematically as in Equation 10, where (i) f(x) is the main function, (ii) $\Phi(x)$ is the feature map, (iii) b is the bias term, (iv) w is the weight vector. In Equation 11, function y can be defined according to the f(x) function. If the sample point is under the hyperplane, it is evaluated as -1. If the sample point is above or directly on the hyperplane, it is evaluated as 1.

$$f(x) = w^T\Phi(x) + b \tag{10}$$

$$y = \begin{cases} 1\ if\ f(x) \geq 0 \\ -1\ if\ f(x) < 0 \end{cases} \tag{11}$$

D. CREATING THE HYBRID ENSEMBLE MODEL

More than one model is created to determine the predictive model that provides the best accuracy performance. 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 ensemble 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, Naive Bayes, Random Forest, Decision Tree, AdaBoost, 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.

The general procedure of the stacking method proceeds as follows. First-level learning algorithms are represented as base learner models, and the second-level learning algorithm is represented as meta-learner model. These learners are combined to create the stacking model. First, base learners are trained with the training part of the data set, and to train the meta learner, it is necessary to create a different training set than the data set used to train the base learners [53]. The results of the predictions are obtained by testing the base learners with the test set. The prediction outputs obtained from the base learners are used as the input of the meta learner. Final prediction results are obtained after training the meta learner with the new created data set. 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.

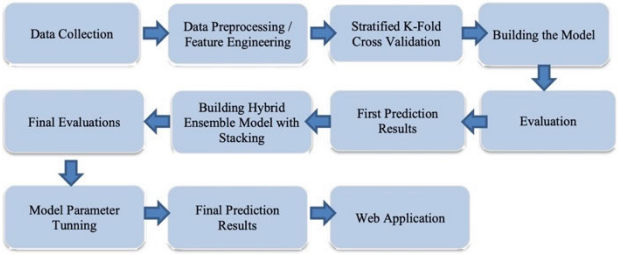


FIGURE 2. Project progress diagram.

Figure 3 shows the block diagram of the proposed method. It summarizes all the steps from the preprocessing of the dataset to the prediction result.

In this paper, the stacking ensemble method was proposed for predicting at-risk students. First, base learner models were trained with the part of the data set created for training. Subsequently, a new training set is created by the model based on the prediction results obtained from the base-learner models. The original data labels are still considered the target class when creating the new dataset. In the following first expression, D is the data set; it includes data features. The second expression is the generated new data set which is expressed as D'. This new generated data set will be used

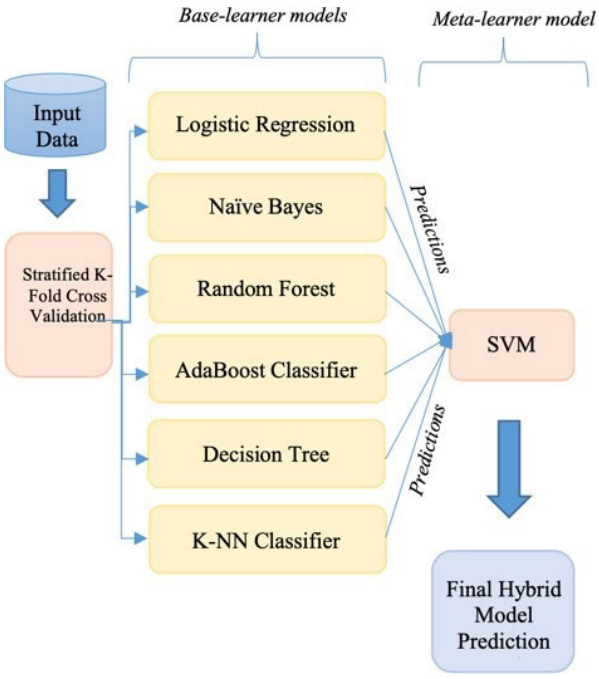


FIGURE 1. Hybrid ensemble model with stacking approach.

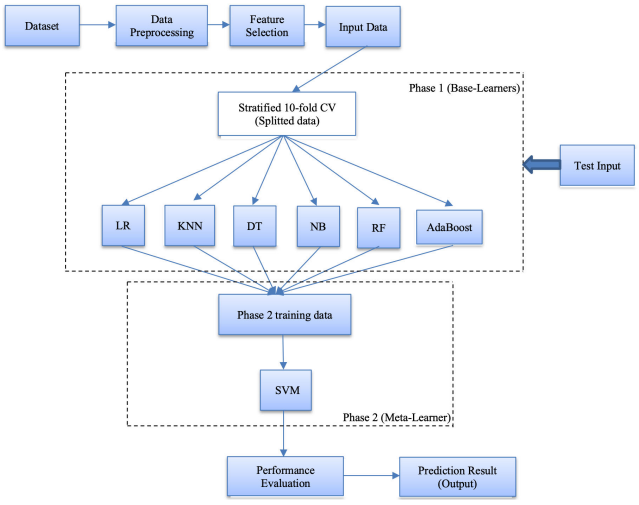


FIGURE 3. Block diagram of the proposed method.

for meta model training. D' includes predictions of base learners and features. Finally, after training the meta learner, we can get predictions of the ensemble model. In the proposed method, stacking 10-fold cross validation was used.

$$D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\} \quad (12)$$

$$D' = \left\{\left(z_{i,1}, \ldots, z_{i,n}\right), y_i\right\} when\ i = 1, 2, \ldots, n \quad (13)$$

E. PERFORMANCE EVALUATIONS

After the final hyper parameter adjustments were made in the hybrid model, the performance of the model was measured with various metrics. There are several metrics for evaluating machine learning models. Since classification models were used in this case, the performance of the hybrid model was evaluated using accuracy, recall, precision, AUC-ROC curve and F1 score metrics. The formulas of the accuracy, recall, and precision metrics are given in Equation 14, 15, and 16.

$$\text{Recall} = \frac{\text{True Positive}}{True\ Positive + False\ Negative} \quad (14)$$

$$Precision = \frac{\text{True Positive}}{True\ Positive + False\ Positive} \quad (15)$$

$$Accuracy = bb\frac{True\ Positive + True\ Negative}{Total\ number\ of\ samples\ predicted} \quad (16)$$

F. DEVELOPING WEB APPLICATION

A web application has been developed to be able to observe the prediction results according to the inputs effectively from the created hybrid model. The target audience of this web application has been identified as instructors and school administration. "Student Performance Prediction Web Application" has been developed for teachers to use the created model. The python programming language and the streamlit framework were used for the construction of the web application. Streamlit is a python framework and allows the creation of web applications on topics such as machine learning and data science. Therefore, streamlit 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. After filling out the form on the Prediction page, information about the student performance can be obtained as a result. The information entered in the form is received as input for use in the model, and the result that the model predicts depending on these inputs is displayed on the screen. De- pending on the prediction result produced by the model, the teacher is informed about whether the student is in a risk situation or not. Other functionalities of the web application are visualization and comparison pages. When a dataset containing specific information about students is uploaded to the Visualization page, it shows the various graphs. These graphs provide visualization of the information contained in the dataset and allow teachers to easily observe the dataset. On the Comparison page, by selecting two different schools in the settings menu, the students' school GPA grades can be compared with the graphs.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

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.

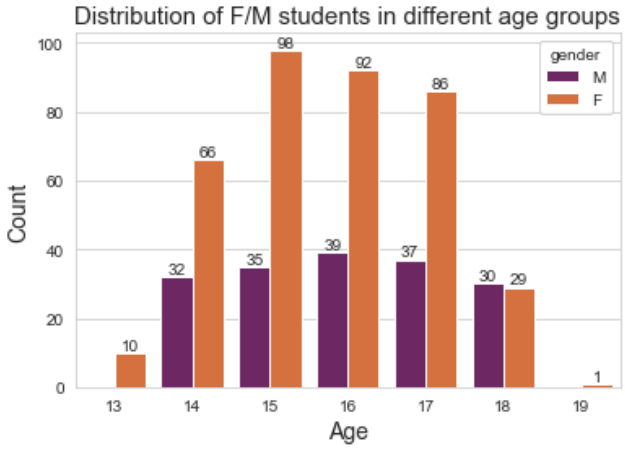


FIGURE 4. Distribution of gender and age of the students.

A. DATA VISUALIZATION

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 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- of-year 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 other words, it can be said that the homework or additional work given

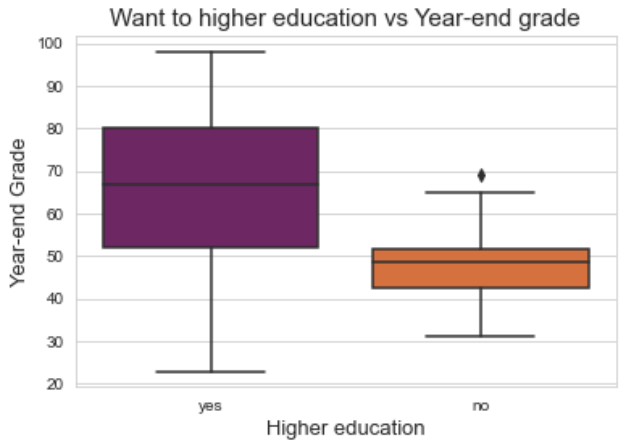


FIGURE 5. Comparison graph of students who want to go to university and their year-end grade averages.

by the teachers positively affects the success of the students. 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-of-year average scores. Most students have a year-end average of between 40 and 80. By visualizing the data in this way, the features that affect the end-of-year average of the students can be observed.

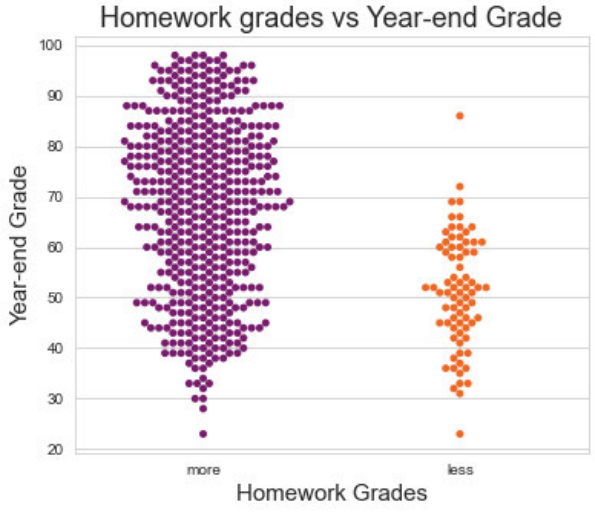


FIGURE 6. Distribution of homework grades and year-end grades. (more: more than 50 / less: less than 50).

Correlation heatmap was created to be able to observe data properties that are useful for prediction. This heatmap is included in Figure 8. With the Correlation heatmap, we can extract the important features and whether there is a correlation between the features. In this project, the important property for prediction is the Year-End Average column. Year-end average grades were used as labels for the predictions. For this reason, the data related to the Year-End Average feature

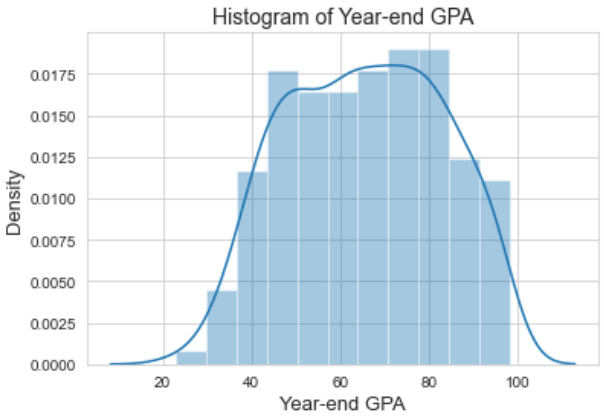


FIGURE 7. Histogram of year-end average grades.

appear in dark color. According to these results, important data properties for prediction are GPA1 (last year), GPA2 (last year), year GPA (last year) and course grades (math, literature, physics, chemistry, biology, history, geography, English, foreign, religion).

These data properties are students' academic data. In other words, academic data affects prediction results more than demographic data. Therefore, students at risk can be predicted efficiently using academic data features. However, in this project, predictions were made with the data set containing both demographic and academic data for the first step. Then, the data set containing the academic data for the hybrid model was used for comparison.

B. RESULTS AND COMPARISON OF THE PREDICTION MODELS

For the student to be considered successful in any course at the end of the academic year, the arithmetic average of the two semester scores must be at least 50. For this reason, year-end average scores were calculated with the course grades in the data set and added to the data set. According to the notes in this column, it was decided to add a target column so that the model could actually predict. If the grades in the year-end average column are 50 and above, a binary value of "1" is assigned. If the grades are below 50, the binary value "0" is assigned. Table 2 shows the percentages of students who passed and failed in the data set. More than half of the students seem to have passed the class.

Logistic Regression, KNN, AdaBoost, SVM, Naive Bayes, Random Forest and Decision Tree supervised learning algorithms were used to create the first models. In the proposed method, the Stratified K-Fold Cross Validation method was used for the models. To observe the effect of using the cross-validation method, the performance results of the models were obtained both using Stratified 10-Fold Cross-Validation and without using cross validation. The reason for using the Stratified 10-Fold Cross-validation method is to prevent deviations and errors when separating the data set into train and test data sets. A comparison of the performance scores of

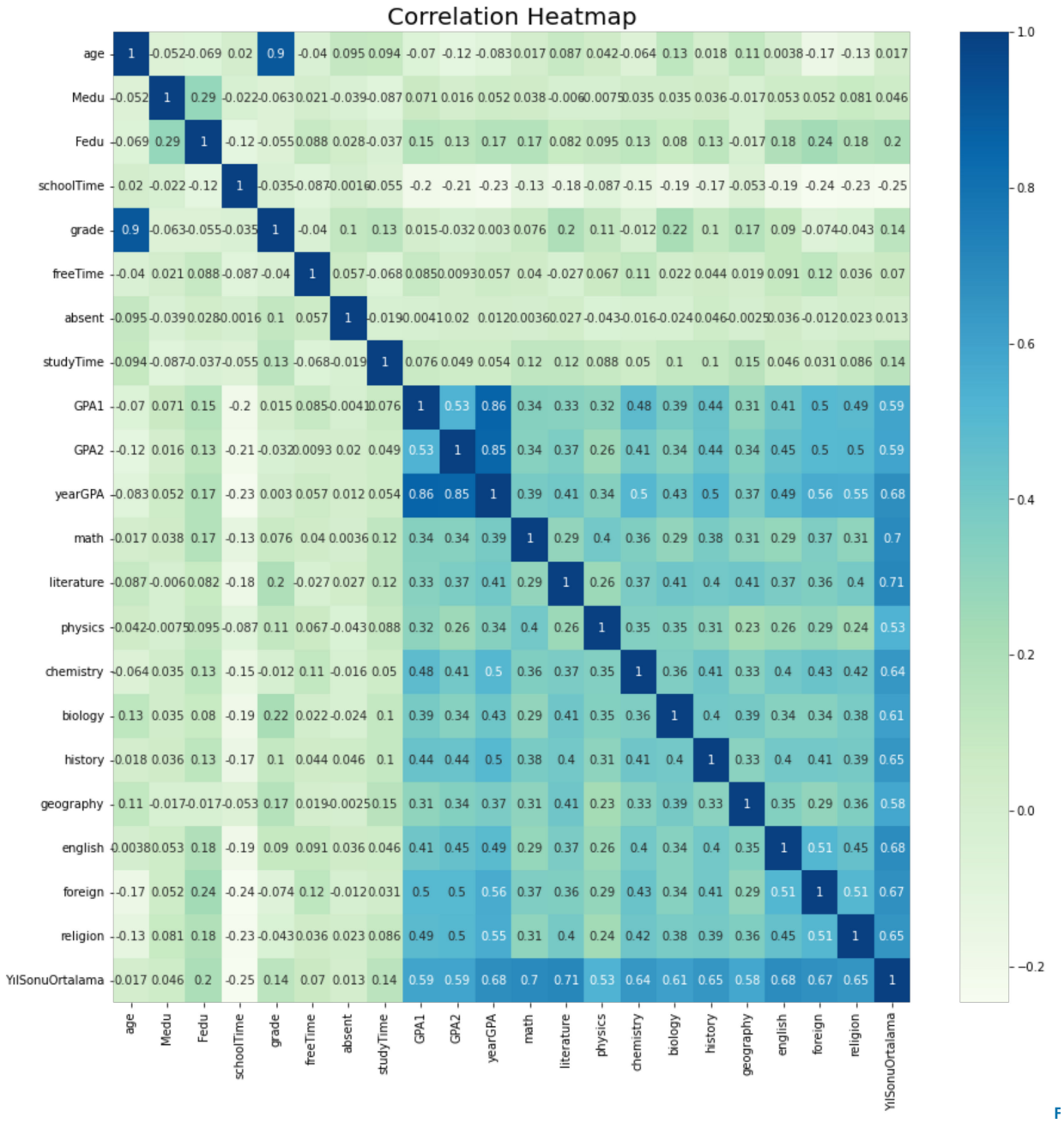


FIGURE 8. Heatmap shows that correlation matrix between features and year-end grade.

TABLE 2. Pass and failure rates of the total students.

Pass rate of the total students	77.48%	430 students
Failure rate of the total students	22.52%	125 students

the models is shown in Table 3. According to the results, the stratified 10-fold cross validation method positively affects the performance of models. Generally, the performances of the models improved with the use of stratified 10-fold cross validation. These results were obtained with the default parameters. The model with the highest performance before parameter optimization is AdaBoost with an accuracy of 93.2%. Table 4 was created to observe and compare the effects of Hyperparameter optimization on models. As observed from this Table 4, the accuracy values of all

TABLE 3. Comparison of the performances scores of the models with using stratified 10-fold cross validation and without using cross-validation.

	Mean Accuracy Results of Models	
Algorithms	With Stratified 10-Fold Cross Validation	Without Cross Validation (75% training / 25% testing set)
Logistic Regression	**92.8%**	92%
K-NN	79.8%	**85%**
Decision Tree	85.2%	**85.6%**
Support Vector Machine	**91.2%**	87.1%
Naïve Bayes Classifier	**45%**	43.1%
Random Forest	**91.4%**	91.3%
AdaBoost	**93.2%**	92.8%

TABLE 4. Comparison of the mean accuracy scores before and after applying hyperparameter optimization to the individual models and hybrid model.

Algorithms	Before Hyperparameter Optimization	After Hyperparameter Optimization
Logistic Regression	92.8%	**94.4%**
Naïve Bayes	45%	**80.7%**
Support Vector Machine	91.2%	**93.9%**
Decision Tree	85.2%	**87.6%**
Random Forest	91.4%	**93.0%**
AdaBoost	93.2%	**94.1%**
K-NN	79.8%	**84.5%**
Hybrid Model (Meta learner: SVM)	87.7%	**94.8%**

models increased after the parameters of the models were adjusted.

Result of the hybrid model can be seen in Table 4. A summary of the best performing parameters obtained after hyperparameter optimization is shown in Table 5. The parameters used in the Stacking Classifier model are also included in the table. For the hybrid model, the meta learner was tested for each algorithm and a hybrid model was created with the meta learner that showed the highest performance. When the models are evaluated individually, Logistic Regression, SVM, AdaBoost and Random Forest have good performance values. When the Logistic Regression model is evaluated individually after parameter optimization, it has the best performance by achieving an accuracy score of 94.4%. After parameter optimization, the accuracy value of the Hybrid model (when meta learner was SVM) increased to 94.8%.

In the proposed method, the Hybrid model consists of two levels. The first level is for base learners and the second level is for meta learners. A hybrid model was created by using each of the seven algorithms as meta-learners. The performance of hybrid models created with different meta-learners was measured with various metrics. Accuracy, precision, recall, F1 score and AUC score values of each hybrid model were obtained. The comparison of these values is given in Table 6.

By using different metrics, the correct predictive performance of the models is observed more precisely. When the SVM algorithm is used as a meta-learner while creating the hybrid model, it achieves the highest accuracy and precision. When the SVM model is used as a meta learner, the accuracy value of the hybrid model is 94.8% and the precision value is 96.8%. It has a higher value than hybrid models created with other meta learners. When the Recall metric is observed,

TABLE 5. Summary of the hyperparameter settings in each of the algorithm.

Models	Parameters
LR	{C= 100.99999447026379, penalty= 'l2', solver= 'liblinear' }
KNN	{weights= 'uniform', n_neighbors= 11, metric= 'euclidean', algorithm= 'auto'}
DT	{criterion='entropy', max_depth=10}
NB	{var_smoothing=1.0}
RF	{n_estimators= 1000, max_features= 'sqrt', criterion= 'entropy'}
AdaBoost	{n_estimators= 100, algorithm= 'SAMME'}
SVM	{kernel= 'linear', gamma= 'scale', decision_function_shape= 'ovo', C= 1.0}
StackingClassifier	{estimators = baseLearners, final_estimator = metaLearner, cv=10} "baseLearners" includes LR, KNN, DT, NB, RF, AdaBoost "metaLearner" includes SVM

TABLE 6. Comparison of the performance values of different meta-learners used in creating a hybrid model with various metrics.

Hybrid Model (Meta Learners)	SVM	LR	K-NN	DT	NB	RF	AdaBoost
Accuracy	**94.8%**	92.6%	93.3%	93.1%	93.5%	93.9%	91.0%
Precision	**96.8%**	94.2%	95.2%	95.0%	94.0%	95.9%	94.2%
Recall	96.5%	96.5%	**97.0%**	94.4%	93.7%	96.0%	94.9%
F1 score	**96.5%**	95.3%	95.9%	95.3%	95.7%	96.1%	94.6%
AUC score	98.2%	98.1%	97.6%	97.4%	**98.4%**	98.3%	97.6%
Total Time Measures (s)	76.20	76.62	75.83	45.54	44.04	11.46	36.27

the hybrid model with the highest degree achieves a performance of 97.0% when KNN used as a meta learner. When the F1 score is observed, the hybrid model with the highest degree is obtained when the SVM is used as a meta learner. F1 score obtained as 96,5%. Finally, when we look at the AUC score, it achieves 98.4% when Naïve Bayes used as a meta learner while creating a hybrid model. Furthermore, when the hybrid models were created using different meta-learners, in addition to performance evaluations, the total training and prediction time measures (in seconds) of the models are included in Table 6. When the performances of the models are measured with different metrics, the SVM model shows very good prediction performance when used as a meta learner. Therefore, in the next work, results were obtained by using the hybrid model with the SVM model used as a meta learner. To better observe the hybrid model result, ROC-AUC curve results were also examined. ROC curve is shown in Figure 9 and AUC values are reported in Table 8. The bar plots in Figure 10 show the performance values measured by various metrics of hybrid models created using different meta-learners.

These results were obtained using demographic and academic data. However, as observed in the data visualization section, it was concluded that academic data was an important factor in predicting students at risk. For this reason, the result of the hybrid model obtained using only academic data is included in Table 7. When demographic and academic data are used, the performance of the hybrid model is 94.8%. However, when only academic data were used, the performance result of the hybrid model increased to 98.4%. As observed in this result, the model achieves a very high performance in predicting the student at risk using academic data. Plotting showing the AUC value and ROC curve for each fold of the hybrid model created when the SVM model is used as a meta

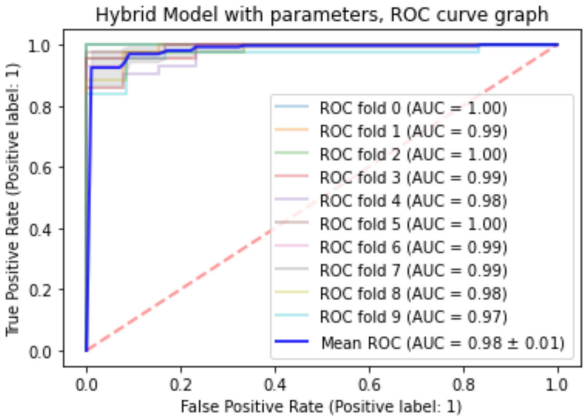


FIGURE 9. ROC curve of the hybrid model (when meta learner is SVM) for each fold and AUC scores.

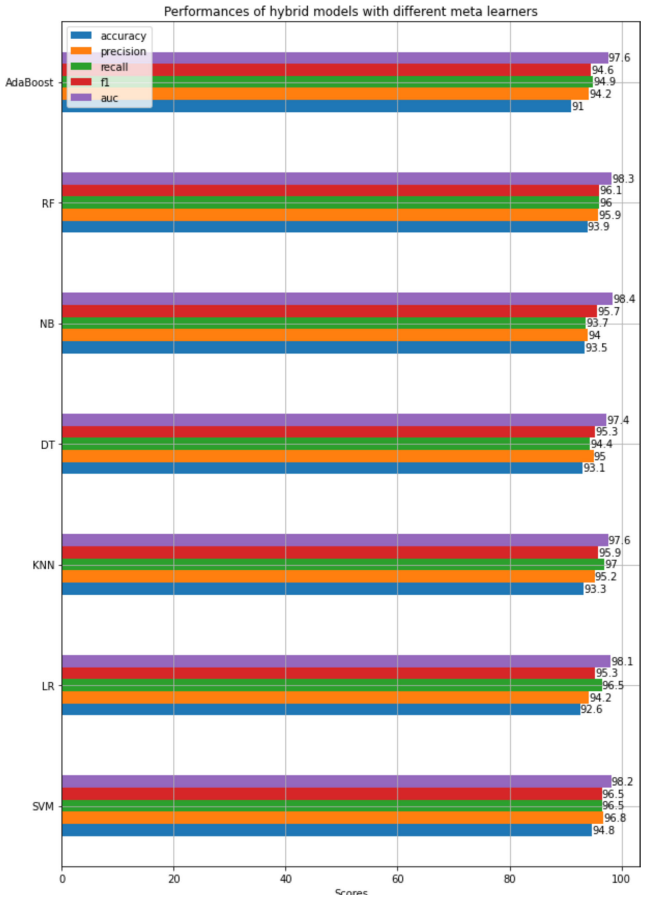


FIGURE 10. Visualizing various metrics values of different meta-learners used in creating a hybrid model.

learner is in Figure 9. The number of folds was set to 10 when using the Stratified K-Fold Cross Validation method. As can be observed from this plotting, the AUC values obtained in each fold and the average AUC value are included. The mean AUC value was obtained as 0.98. In addition, the AUC value for each fold can be observed from Table 8. Considering all the results obtained, the hybrid model has better performance than the individual models.

TABLE 7. The measurement of the performance of the best performing meta learner (SVM) on both academic data and academic/demographic data.

Base Learners	Meta Learner	Demographic and Academic Features	Only Academic Features
LR KNN Decision Tree Naïve Bayes RF AdaBoost	SVM	94.8%	**98.4%**

TABLE 8. AUC values for each fold of the hybrid model.

Folds	AUC scores
Fold 0	1.00
Fold 1	0.99
Fold 2	1.00
Fold 3	0.99
Fold 4	0.98
Fold 5	1.00
Fold 6	0.99
Fold 7	0.99
Fold 8	0.98
Fold 9	0.97
Mean	**0.98**

Therefore, the use of the hybrid model is useful for the prediction results in identifying the students who are at risk. Also, data sets were analyzed and used in the hybrid model. As a result of these analyzes, it was observed that using only academic data increased the prediction performance of the hybrid model. Furthermore, the SVM algorithm was chosen as meta learner and 98.4% performance result was obtained using only academic data. With this result, it is understood that when the model is trained and tested using the academic data features of the students at risk, it makes a good prediction with an accuracy rate of 98.4%.

C. WEB APPLICATION

A web application has been created so that this hybrid model can be used efficiently by educators. The hybrid model created has been integrated into the web application, so that prediction results can be obtained effectively. On the prediction page in this application, depending on the input values entered by the user, it gives a prediction result about whether the student is in a risk situation or not. According to this result, information such as the success of the student and the

course table with low grades are shown to the user by the application. In this way, it is aimed to provide convenience to teachers to identify students who are in a risk situation.

D. COMPARISON BETWEEN HYBRID MODEL RESULTS AND EXISTING WORKS

In existing studies, it has been observed that researchers mostly created models using individual machine learning algorithms for subjects such as estimating students at risk or predicting student performance. The method we present for predicting students at risk includes the development of a hybrid ensemble model with the stacking method. Unlike the methods created by the studies described in the literature, a hybrid model was created by combining multiple supervised learning algorithms with the stacking method. We get our final prediction results with the meta learner, which is determined depending on the prediction results obtained from each algorithm. This increases the prediction performance.

Comparing the results of the studies described in the literature with the hybrid ensemble model presented in this paper, we can clearly observe how the studies that use individual machine learning algorithms donot perform as good as the developed ensemble model. For instance, the Random Forest algorithm, which is one of the popular methods used to predict the performance of students was reported to have an initial accuracy of 79% and 91% accuracy value was obtained in the final step [15]. Another approach developed in previous studies utilized Neural Network as the learning model; the reported performance result was 70% [6]. The performance results of the models developed using C4.5, Naïve Bayes and K-Nearest Neighbor algorithms were 87.55%, 87.95% and 86.3%, respectively [11]. In another study described in the literature, when all data set variables were used, it was observed that 91.9%, 86.3% and 90.7% PCC performance values are obtained, respectively; here, the prediction was made for the mathematics lesson and the study used models created with the naïve predictor (NV), SVM and DT algorithms. When the same models were used for another course, 89.7%, 91.4% and 93.0% PCC values were obtained from the NV, SVM and DT models, respectively [1].

One of the studies described in the literature argues that combining many algorithms will give better results. For this purpose, the authors combined three decision schemes, namely K-star, Naïve Bayes and C4.5 algorithms. As a result, 85% performance was obtained from decision scheme 1 in the last step [2]. The hybrid ensemble model presented in this paper with 98.4% accuracy reported by utilizing only academic data outperforms the studies described in the literature. The hybrid ensemble model presented in this paper combined 7 different supervised learning algorithms leading to higher performance results. The reason for using the hybrid ensemble model presented in this paper to identify students at risk is that it learns different data points by training each algorithm and experiencing it with prediction results. The conducted comparative study demonstrates how it is worth investing the extra effort of combining ensemble models. The same approach has been also demonstrated effective in various other domains where ensemble models outperformed single learning models.

V. LIMITATIONS OF THE WORK

This study focuses on utilizing machine learning techniques for predicting students at risk. For this purpose, we development a hybrid ensemble model with stacking method. Although the research achieved its main target, some limitations were encountered during the project. Since the data contains demographic information, ethics committee approval was first obtained to collect the data set. Afterwards, permission was obtained from high schools to collect the data. The size of the data set is limited due to the scarcity of accessible high schools. Another limitation encountered during the project was that the high school teachers did not find it appropriate to fill in the form created by the authors in the digital environment to collect the data set, so the form was distributed to the students in paper form. Then all the data collected in this process were transferred to the computer environment. This extended the time of the data collection process and required careful handling of the data entry to avoid erroneous data.

When the collected data set was analyzed, it was observed that small number of students were reported with low marks. For this reason, a random data set was additionally created and added to the original data set so that the model could smoothly recognize low marks. Another limitation is the large number of data features. After the categorical data was converted to numerical values with one-hot encoding method, the properties of the data set increased. Therefore, feature engineering was performed so that this situation does not affect the prediction performance of the model. According, we determined data features deemed useful in predicting students at risk. In the latest hybrid model, only academic data was used, and the performance of the model increased. Another limitation of this study is that only certain courses have been considered. The lessons taught in every school are not the same.

In their high school education, students choose a field such as quantitative, verbal or a balance of both. The courses taken by students enrolled in these fields are not the same. However, since it was not possible to collect data from each field and school, students were considered taking most of the common courses when the data set was collected. In future studies, it is important for the development of the project to adopt the education system of each school and to collect data from as many schools as possible. The general limitation of this study is that the education culture and rules which are considered valid in Turkey may be different from those associated with the existing works described in the literature. In other words, the culture and the applicable rules may differ from one country to another, even from one region to another. Hence, systems that are effective for one dataset may not be effective and may not work in the same way in every country. Depending on Turkish education guidelines, approaches for

identifying students at risk are slightly different. Because the passing grade and the style of the exams are different. In future studies, the project can be developed by considering the limitations mentioned.

VI. CONCLUSION 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. The hybrid model with the highest performance was achieved when the SVM was used as a meta-learner. When SVM is used as meta learner, 94.8% accuracy value and 96.8% precision value is obtained. When hybrid models created with different meta learners are compared, the best performance is obtained when SVM meta learner is used.

Accuracy, precision, recall, F1 score and AUC score metrics were used to compare the performance of hybrid models. The performance of models to make accurate predictions can be compared using various metrics. Furthermore, a hybrid model comparison was made using academic data and using both demographic and academic data. According to this comparison result, when both demographic and academic data are used together, the performance of the hybrid model is 94.8%, but when only academic data is used, the performance of the hybrid model increases to 98.4%. In other words, it has been observed that using only academic data is very useful in predicting students who are at risk. According to the results obtained, it is observed that the hybrid model provides better performance than the individual models. A web application has been developed for teachers to use the hybrid model. Depending on the inputs they enter this application, teachers can get information about the students' school success performance. In this way, students who are in a risk situation can be identified and assistance can be provided to them by teachers. In accordance with the purpose of the project, the hybrid ensemble model was created to identify students at risk using the stacking method. As can be seen from the results, the use of a hybrid model gives useful results on this issue. This project was made for high school students studying in Turkey. In future studies, the target audience of the project can be developed, and this method can be planned to include university students. In addition, only certain courses are included in the dataset used, because the dataset was collected assuming that students are taking these courses. In advanced studies, the course list can be expanded, and the same methods can be used after collecting data from different schools.

DECLARATIONS

AVAILABILITY OF DATA AND MATERIALS

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

COMPETING INTERESTS

No competing interest

AUTHORS' CONTRIBUTIONS

Reyhan Zeynep Pek: Developed the methodology, wrote the programs, conducted the experiments, wrote the article, analyzed the results, reviewed, and approved the article; and Sibel Tariyan Özyer, Tarek Elhage, Tarek Elhage, Tansel Özyer, and Reda Alhajj: Developed the methodology, analyzed the results, reviewed, and approved the article.

REFERENCES

[1] P. Cortez and A. Silva, "Using data mining to predict secondary school student performance," in *Proc. 15th Eur. Concurrency Eng. Conf. (ECEC), 5th Future Bus. Technol. Conf. (FUBUTEC)*, 2008, pp. 5–12.

[2] E. Er, "Identifying at-risk students using machine learning techniques: A case study with IS 100," *Int. J. Mach. Learn. Comput.*, vol. 2, no. 4, pp. 476–480, 2012, doi: 10.7763/ijmlc.2012.v2.171.

[3] S. Isljamovıc and M. Suknovıc, "Predicting students' academic performance using artificial neural network: A case study from faculty of organizational sciences," *Eurasia Proc. Educ. Social Sci.*, vol. 1, pp. 68–72, May 2014.

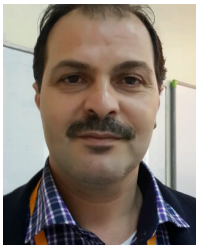
[4] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Comput. Educ.*, vol. 53, no. 3, pp. 950–965, 2009, doi: 10.1016/j.compedu.2009.05.010.

[5] 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.
[6] 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.
[7] 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.
[8] 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.
[9] 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.
[10] 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.
[11] 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
[12] 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.
[13] 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.
[14] 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.
[15] M. Adnan, A. Habib, J. Ashraf, S. Mussadiq, A. A. Raza, M. Abid, M. Bashir, and S. U. Khan, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *IEEE Access*, vol. 9, pp. 7519–7539, 2021, doi: 10.1109/ACCESS.2021.3049446.
[16] G. Sujatha, S. Sindhu, and P. Savaridassan, "Predicting students performance using personalized analytics," *Int. J. Pure Appl. Math.*, vol. 119, no. 12, pp. 229–238, 2018.
[17] 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.
[18] 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.
[19] R. Al-Shabandar, A. J. Hussain, P. Liatsis, and R. Keight, "Detecting at-risk students with early interventions using machine learning techniques," *IEEE Access*, vol. 7, pp. 149464–149478, 2019.
[20] L. Cen, D. Ruta, L. Powell, B. Hirsch, and J. Ng, "Quantitative approach to collaborative learning: Performance prediction, individual assessment, and group composition," *Int. J. Comput.-Supported Collaborative Learn.*, vol. 11, no. 2, pp. 187–225, Jun. 2016.
[21] Y. Cui, F. Chen, and A. Shiri, "Scale up predictive models for early detection of at-risk students: A feasibility study," *Inf. Learn. Sci.*, vol. 121, nos. 3–4, pp. 97–116, Feb. 2020.
[22] M. Hlosta, D. Herrmannova, L. Vachova, J. Kuzilek, Z. Zdrahal, and A. Wolff, "Modelling student online behaviour in a virtual learning environment," 2018, *arXiv:1811.06369*.
[23] S. Palmer, "Modelling engineering student academic performance using academic analytics," *Int. J. Eng. Educ.*, vol. 29, no. 1, pp. 132–138, 2013.
[24] 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.
[26] 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.
[27] G. Akçapınar, M. N. Hasnine, R. Majumdar, B. Flanagan, and H. Ogata, "Developing an early-warning system for spotting at-risk students by using eBook interaction logs," *Smart Learn. Environ.*, vol. 6, no. 1, pp. 1–15, Dec. 2019.
[28] 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.
[29] J. Berens, K. Schneider, S. Görtz, S. Oster, and J. Burghoff, "Early detection of students at risk—Predicting student dropouts using administrative student data and machine learning methods," CESifo Group Munich, CESifo Work. Paper 7259, 2018, pp. 1–41. Accessed: Dec. 30, 2022. [Online]. Available: https://ssrn.com/abstract=3275433
[30] S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Appl. Sci.*, vol. 9, no. 15, p. 3093, Jul. 2019.
[31] J. Figueroa-Canas and T. Sancho-Vinuesa, "Predicting early dropout student is a matter of checking completed quizzes: The case of an online statistics module," in *Proc. LASI-SPAIN*, 2019, pp. 100–111.
[32] C. Burgos, M. L. Campanario, D. D. L. Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Comput. Electr. Eng.*, vol. 66, pp. 541–556, Feb. 2018.
[33] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Hum. Behav.*, vol. 73, pp. 247–256, Aug. 2017.
[34] A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," *Int. J. Mod. Educ. Comput. Sci.*, vol. 8, no. 11, p. 36, 2016.
[35] S. Gupta and A. S. Sabitha, "Deciphering the attributes of student retention in massive open online courses using data mining techniques," *Educ. Inf. Technol.*, vol. 24, no. 3, pp. 1973–1994, May 2019.
[36] A. Wolff, Z. Zdrahal, A. Nikolov, and M. Pantucek, "Improving retention: Predicting at-risk students by analysing clicking behaviour in a virtual learning environment," in *Proc. 3rd Int. Conf. Learn. Anal. Knowl.*, 2013, pp. 145–149.
[37] J. Park, K. Denaro, F. Rodriguez, P. Smyth, and M. Warschauer, "Detecting changes in student behavior from clickstream data," in *Proc. 7th Int. Learn. Anal. Knowl. Conf.*, Mar. 2017, pp. 21–30.
[38] 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.
[39] C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun, and S. Ventura, "Early dropout prediction using data mining: A case study with high school students," *Expert Syst.*, vol. 33, no. 1, pp. 107–124, Feb. 2016.
[40] L. C. B. Martins, R. N. Carvalho, R. S. Carvalho, M. C. Victorino, and M. Holanda, "Early prediction of college attrition using data mining," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 1075–1078.
[41] S. Huang and N. Fang, "Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models," *Comput. Educ.*, vol. 61, pp. 133–145, Feb. 2013.
[42] J.-L. Hung, M. C. Wang, S. Wang, M. Abdelrasoul, Y. Li, and W. He, "Identifying at-risk students for early interventions—A time-series clustering approach," *IEEE Trans. Emerg. Topics Comput.*, vol. 5, no. 1, pp. 45–55, Jan. 2017.
[43] 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.
[44] 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.
[45] A. S. Imran, F. Dalipi, and Z. Kastrati, "Predicting student dropout in a MOOC: An evaluation of a deep neural network model," in *Proc. 5th Int. Conf. Comput. Artif. Intell. (ICCAI)*, 2019, pp. 190–195.
[46] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauría, J. R. Regan, and J. D. Baron, "Early alert of academically at-risk students: An open source analytics initiative," *J. Learn. Anal.*, vol. 1, no. 1, pp. 6–47, May 2014.
[47] G. Körösi and R. Farkas, "MOOC performance prediction by deep learning from raw clickstream data," in *Proc. Int. Conf. Adv. Comput. Data Sci.* Valletta, Malta: Springer, 2020, pp. 474–485.

[48] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, vol. 103, pp. 1–15, Dec. 2016.
[49] N. Wu, L. Zhang, Y. Gao, M. Zhang, X. Sun, and J. Feng, "CLMS-Net: Dropout prediction in MOOCs with deep learning," in *Proc. ACM Turing Celebration Conf.*, May 2019, pp. 1–6.
[50] S. N. Liao, D. Zingaro, K. Thai, C. Alvarado, W. G. Griswold, and L. Porter, "A robust machine learning technique to predict low-performing students," *ACM Trans. Comput. Educ.*, vol. 19, no. 3, pp. 1–19, Sep. 2019.
[51] 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.
[52] K.-W. Hsu, "A theoretical analysis of why hybrid ensembles work," *Comput. Intell. Neurosci.*, vol. 2017, pp. 1–12, Jan. 2017.
[53] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.

REYHAN ZEYNEP PEK received the B.Sc. degree from the Computer Engineering Department, Istanbul Medipol University, Turkey, where she is currently pursuing the graduate degree in computer engineering. Her research interests include data science, machine learning, assessment in education systems, and data analysis.

SIBEL TARIYAN ÖZYER received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Çankaya University, Ankara, Turkey. She is an Assistant Professor of computer engineering with Ankara Medipol University, Turkey. Her research interests include data science, machine learning, network systems, health informatics, image processing, and computer vision.



TAREK ELHAGE received the B.Sc. degree from the Department of Chemistry, Middle East Technical University, Ankara, Turkey. He is a Teacher with the ABC Private School, Abu Dhabi, United Arab Emirates. His research interests include course development, assessment in education systems, and data analysis.

TANSEL ÖZYER received the B.Sc. degree from the Computer Engineering Department, METU, the M.Sc. degree from the Computer Engineering Department, Bilkent University, and the Ph.D. degree in computer science from the University of Calgary. He is a Professor of computer engineering with Ankara Medipol University, Turkey. His research interests include data science, machine learning, bioinformatics, health informatics, XML, mobile databases, image processing, and computer vision.

REDA ALHAJJ (Senior Member, IEEE) is a Tenured Professor with the Department of Computer Science, University of Calgary, Alberta, Canada. He is also with Medipol University Istanbul, Turkey, and the University of Southern Denmark, Odense, Denmark. He has published over 500 papers in refereed international journals, conferences, and edited books. He currently leads a large research group of Ph.D. and M.Sc. candidates. His primary work and research interests include various aspects of data science and big data with emphasis on areas, such as scalable techniques and structures for data management and mining; social network analysis with applications in computational biology and bioinformatics, homeland security, and disaster management; sequence analysis with emphasis on domains like financial, weather, traffic, and energy; and XML, schema integration, and reengineering. He served on the program committee of several international conferences. He received the Best Graduate Supervision Award and the Community Service Award from the University of Calgary. He recently mentored a number of successful teams, including SANO who ranked first in the Microsoft Imagine Cup Competition in Canada and received the KFC Innovation Award in the World Finals, Russia, TRAK who ranked in the top 15 teams in the Open Data Analysis Competition in Canada, Go2There who ranked first in the Imagine Camp Competition organized by Microsoft Canada, and Funiverse who ranked first in Microsoft Imagine Cup Competition in Canada. He is the Founding Steering Chair of the flagship conference "IEEE/ACM International Conference on Advances in Social Network Analysis and Mining" and three accompanying symposiums FAB (for big data analysis), FOSINT-SI (for homeland security and intelligence services), and HI-BI-BI (for health informatics and bioinformatics). He is a member of the Editorial Board of the *Journal of Information Assurance and Security*, *International Journal of Data Mining and Bioinformatics*, and *International Journal of Data Mining, Modeling and Management*. He is a guest editor of a number of special issues and edited a number of conference proceedings. He is the Founding Editor-in-Chief of the *Social Networks Analysis and Mining* (Springer), *Lecture Notes on Social Networks* (Springer), *Network Modeling Analysis in Health Informatics and Bioinformatics* (Springer), and *Encyclopedia on Social Networks Analysis and Mining* (Springer) (ranked top third in most downloaded sources in computer science in 2018).