Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/taut20

A novel approach to predict competency and the hidden risk factor by using various machine learning classifiers

Stalin M. & Kalyani S.

To cite this article: Stalin M. & Kalyani S. (2023) A novel approach to predict competency and the hidden risk factor by using various machine learning classifiers, Automatika, 64:3, 550-564, DOI: 10.1080/00051144.2023.2200347

To link to this article: https://doi.org/10.1080/00051144.2023.2200347

Published online: 17 Apr 2023.

AUTOMATIKA
2023, VOL. 64, NO. 3, 550–564
https://doi.org/10.1080/00051144.2023.2200347

REGULAR PAPER

OPEN ACCESS Check for updates

A novel approach to predict competency and the hidden risk factor by using various machine learning classifiers

M. Stalin and S. Kalyani

Electrical and Electronics Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, India

ABSTRACT

In a survey conducted in the year 2020, we came to know that India's around 50% of population includes young people of the age group of 25 and students. Guiding this young mass in the right way and strengthening their future is a huge responsibility put over the head of the elder citizens of India such as their parents teachers and professors. This paper aims to build a model that can predict the students' competency level and the risk factors or the fields where he needs to put their effort to improve themselves, and this model also helps the parents, professors and Educational institutes to know about their children's and students in which zone they stand, are they ready to compete with others. This analysis is done by using different ML bifurcation algorithms. Also we aim to find the best classifier which can emerge with the highest predicting accuracy among all other classifiers to the above-said problem. The accuracy of 88.5% is achieved through the proposed machine learning algorithm for particular education datasets which have been taken into consideration.

ARTICLE HISTORY
Received 6 January 2023
Accepted 1 April 2023

KEYWORDS
Decision tree; random forest; support vector; logistic regression classifier

1. Introduction

The growth and future of the nation directly depend on the youth power present in the country. Strengthening the youth power is nothing but strengthening the nation. Students are the main source of youth power they can be found in institutions/universities, their performance competency plays a major role in the socio-economic growth of the nation which can be achieved by producing graduates with problem-solving attitudes, innovators and speculators [1]. So, we focused to begin a quest under socio-economic and psychological attributes that have a major impact on the academic goal. An inept way of training system can untangle by upgrading with crucial scrutinized policies is only the ultimate aim of this study. Speculating the pre-university zone can impart a clear logic of what pupils can achieve in life. Pupils enjoy the phase of adolescence in pre-university examination, this is the intense time they are more firmly fascinated by their environment.

There is a lack of guidance during this phase, if we should properly guide them will definitely drive them from the dark phase of adolescence towards the light space where they can enjoy and build their effective career by having a fabulous academic record track. To predict and build a model that can provide us with statistical proof-based logic, we need to have profound information about the pupils, the only source to accumulate all these data is through the educational institutes' training centres and other forms of coaching centres. In this study, we mainly focus on pupils who have recently cleared their pre-university exams and are those who are filled with a dream of joining higher courses. To build an effective model there is a need to undergo certain stages, first among them is the pre-processing stage where the raw data have been preprocessed by applying data balancing, normalization and optimal equal width binning form. We have also applied three different dimensionality reduction algorithms that reduce the number of attributes that are used during training. Since we have used several preprocessing techniques, we have comprised our comparative study with the combination of different preprocessing models to derive the best data model for our system.

2. Related work

Final grades are the most important part of students' life to establish their bright careers. In 1993 experiments on Neural networks have been profoundly started by Gedeon and Turner to find out a variety of neural networks such as feed-forward networks and back propagation and also to learn how to use these networks and what kind of problems to use what kind of network, etc. [2]. They started conducting their experiments on the prediction of the final grade system. In the year 2002 Wang and Mitrovic were involved in another kind of

CONTACT M. Stalin stalinmur@gmail.com Electrical and Electronics Engineering, Kamaraj College of Engineering and Technology, Virudhunagar, Tamilnadu, India

experiment by developing a completely different system wherein they have undergone a study that can predict the numerical digit that indicates the mistakes a student would make by combining neural network, back propagation and feed-forward techniques [3]. In a slightly different way another research has been undertaken by Olado-kun, Adebanjo and Charles-Owaba in the year 2008 on predicting the acceptance rate of mistakes done by pupils at university using a multilayer perception idea [4].

In 2009 Dean, Ayers and Nugent projected a work which talks about the skill level by bifurcating the pupils based on similar skill levels which are held by the pupils, to do so they used a clustering design technique-means and other clustering algorithms available. In 2013 another work was carried out by Pal where a variety of classification techniques are used to group the pupils who need extra care, for instance, counselling from the trainer, etc. [5].

In this work, we have used multiple Machine Learning and Deep Learning techniques to predict the factors that affect the pupils' performance to a greater extent and some of the factors which are less effective in spoiling the final grade where we can give a certain permission of freedom to pupils and to which factors we should not let them any freedom will be clearly analysed by using different ML classifiers in this work different ML classifiers such as Decision Tree Classifier Random Forest Classifier, Support Vector Classifier, Logistic Regression Classifier, Ada Boost Classifier, Stochastic Gradient Descent Classifier are used on the prepared dataset to predict the final grade of the student [6].

3. Discussion on influencing fact

Building a model that can predict the competency level of the pupils requires a lot of information. Gathering this information is a hectic task [7]. After having a certain level of debate with the students and the experts we have developed around forty questions which are the most influencing and effective attributes. All these questions are spitted into a group of three such as the Pupils' educational record track, Psychological factors and econo-social status of the pupils. The following groups contain certain questions related to the corresponding group [8].

a) **Economical status in society**

Economical and social status of the family is one of the important categories of the question whose questions are having a treble influence on the final grade of the children [9,10]. Here the sex and the age of an individual have a greater impact on their educational achievement. Also, some other questions like Father and Mothers' education in society, Occupation of the Father and mother, Net worth of the family i.e. the economical status of the family in the society, Place of education and the level of private tutor, etc., all these questions have a greater impact on the pupils and children's academic track record and in the final grade of the pupils [11].

b) **Psychologically influencing attributes**

Psychology is one of the most important factors and the field which has shown a greater influence not only in our study but also in all aspects of life and each field of the world. When it comes to the pupils, the final grade this part has a major role to play in the outcome of the examination. Many more questions will come one after the other which are correlated to one another. Among them, the most important fact of attributes is considered in this discussion. The last factor that influences the Child's final grade is the size of the family and the intensity of parents' involvement in educational aspects. The number of siblings and their age difference plays a key role in the competency level and the outcome [12]. According to parents' participation in education was found to have a major influence on the education system is nothing but the parents who show active responsiveness in their children's education. In the year 1993, a study was undertaken by Hatzichristou which says parents' marital status is another psychological factor that affects or improves the final grade of the pupils. The marital status of the parents has a huge psychological effect on the pupils' minds leading to a direct effect on the final exam outcome [13]. The presence or absence of the pupils' parents is another factor which has a greater role in the outcome of the pupils' final exam. The absence of parents in one's life makes pupils feel more depressed and it lacks self-motivation towards their studies also financial matters make the pupils distracted from their studies [14]. The absence of parents is like a pupil participating in a race where there is no aim or goal. According to pupils and young people who are involved in physical activity have shown greater achievement in academics [15]. The teachers and the professors energize their pupils by supporting them to do something new and motivating them to hold a higher rank and making them feel comfortable in understanding the concept, etc., such an active teachers and professors' role has a major influence on the final grade of the students. Having motivating parents in one life has a major influence on the students' academic career. Review from most of the students says their desire to study and for higher studies and self-shown interest towards study made them achieve a major academic milestone [16]. The number of hours pupils spend their time in social media has a greater influence on their academic record. Romantic relationship and the time spent with friends by pupils has a direct influence on the academic achievements

of the students. Alcohol/drugs and smoking attributes have again a major influence on academics. Finally the health of the pupil and their family members has a major influence on the educational record of the native pupil.

c) **Pupils' educational record track**

The basic education of the students plays a major role in their further academic records. Childhood education acts as a foundational stone for students in building a strong higher education record. It has been studied that students with an excellent childhood records have shown good final grades in the higher secondary education system. There are multiple questions such as a weekly timetable for the studies where the amount of hour pupil invest in studying has a prominent influence on the final grade, extracurricular activity which have helped the students to learn the skills which are required throughout their life have a certain level of direct or indirect influence on the academic record, accessibility to the internet and 10th-grade percentiles also plays a prominent role According to Kirby and McElroy students who attended the class regularly will perform well in the final exam when compared to the students who have not attended the classes. By considering all these factors we undergo to develop a model which can give us a clear idea about which factors are most important and which are least important based on that parents and teachers and students can decide where they need to give proper attention and where they should not be predicted by using different classifiers of the machine learning.

4. Methodology

Machine learning is showing tremendous influential growth in the prediction domain in recent years because of its simplicity and the effective classifiers it owns. In our studies, we make use of most of the latest classifiers present in the ML to make a final grade judgment on the pupils. We have classified these students into three categories, "good", "fair" and "poor", according to their final exam performance. Then we analysed a few features that have a significant influence on students' final performance, including Romantic Status, Alcohol Consumption, Parents' Education Level, Frequency Of Going Out, Desire Of Higher Education and Living Area. Finally, leveraging available features, we have created various machine learning models to predict students' final performance classification and have compared models' performance based on one-out sample accuracy score [17]. There are multiple steps needed to be followed the first and foremost is data gathering second one is data cleaning and the final step is to build a model. All these steps will be discussed in detail.

a) **Data collection and preparation**

Gathering data is the most important step in creating any model. These data are collected by the pupils in surveys and questionaries'. This questionnaire includes all the attributes and the question which we have discussed in section three. All these data have to be cleaned and verified because some missing value data attributes have been found in the survey from some students and we have not considered such data it may result in some variation in the final results or in getting the accuracy. Most of the data have been properly given by the pupils and all these data have to be cleaned and maintained in a proper Excel format. Once the data set is prepared the next step is to find the factors which have a greater influence and those factors which don't have a greater influence all this relationship between the attributes are drawn in the form of the Correlation Heat map. This correlation heat map shows how romantic status, alcohol consumption, parents' education level frequency of going out with friends, the desire for higher education and the place are interrelated to each other and in a mathematical way.

b) **Decision tree classifier**

A decision tree is one of the main streams of the machine-learning approach. In this approach decisions are made based on the graphical representation of the attributes in the form of a tree. Whenever there are noisy data and the data which are represented in the form of the attribute value pair and it's a kind of analysis problem, it is highly recommended to use the decision tree classifier algorithm. Describing the graphical representation it contains a root node and each routing node either contains a child or a sub-root node attribute with the final target attribute values of yes or no. Based on a certain combination of the condition it becomes possible to predict.

In Figure 1 the tree representation of the pupils' attributes is shown through which it is possible to decide which kind of students will score good marks. The pupils whose SSC marks are less there is a greater chance of failing the examination, and the pupils whose marks are more and not engaged in any kind of romantic relationship and have no habit of consuming alcohol and attend examination with proper preparation will score more in the examination. In a similar manner the pupils who score high in the SSC exam and are in a romantic relationship and have a habit of alcohol consumption and attend the examination have a more chance of not scoring high marks. However, the structure shown is just a sample where A lot of other attributes are also need to be considered for the prediction which had a greater influence in the outcome. A decision tree can be built by finding the entropy of each attribute and the total information gain of all

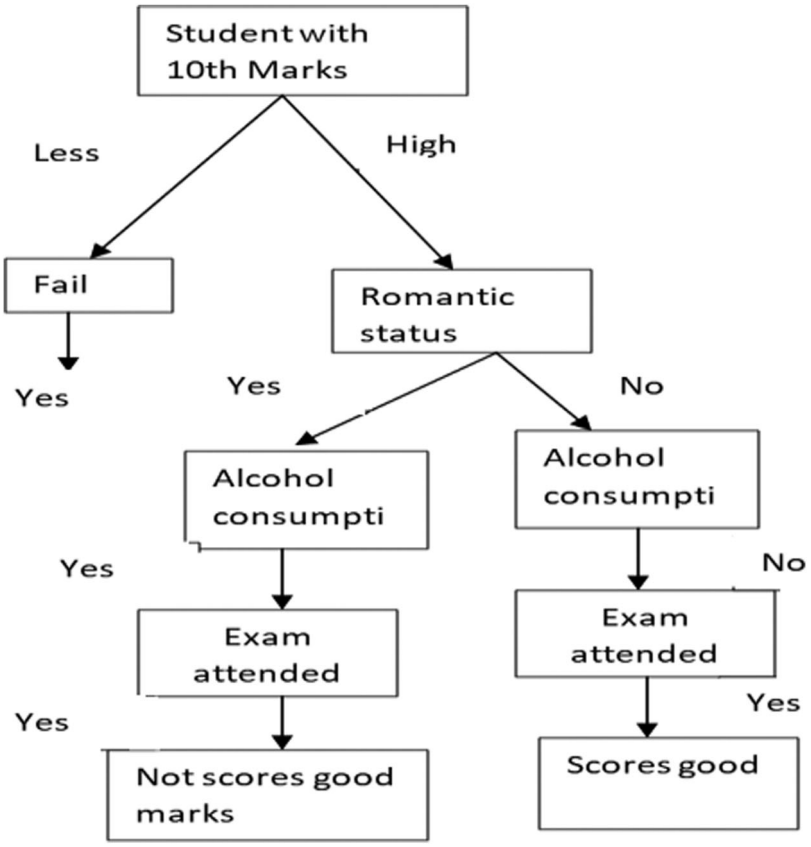

Figure 1. Decision-tree-based competency prediction structure.

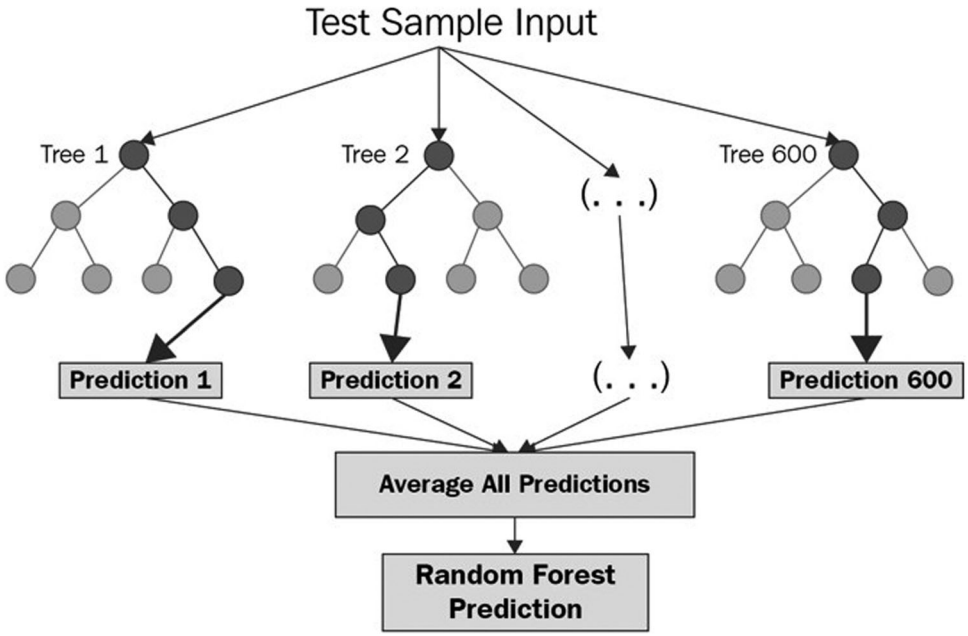

Figure 2. Random forest-based competency prediction structure.

the attributes participating in the model. The attribute with the highest information gain is considered the root node. Similarly the sub-root nodes are selected and the tree is represented to make a proper prediction.

c) **Random forest classifier**

Random forest method is extensively used for decision-making regression and classification problems. Figure 2 shows the graphical representation of the random forest. In this method, the entire dataset is taken into consideration. This dataset is further clustered into multiple sub-trees and a final evaluation score is taken by votes [18]. When a new data instance is passed into a model of random forest based on the majority of similarity found in the cluster is collected by votes and the prediction for the corresponding given instance is done.

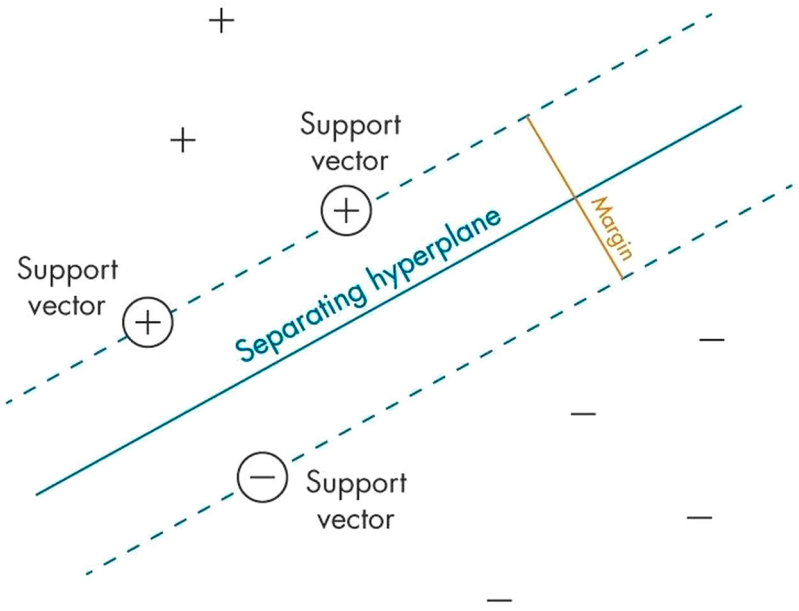

Figure 3. Representation of hyper plain and the data points in the *n*-dimensional space.

d) **Support vector classifier**

Support Vector Machine, SVM, is one of the terminologies associated with machine learning. This model learns from the past input data and makes a future prediction as an output. By continuously training we teach the model which is an apple and which is not. Once the model is completely trained and it is going to predict the apple. This way of training the system is known as supervised learning. So this supervised learning is a subset of machine learning which contains the SVM algorithm [19,20]. This algorithm can be used for both classification and regression kind of problems. This SVM works on the concept of the hyper-plane where all the attributes are taken into consideration and each data point is plotted or represented in an n-dimension plane space with the value of each attribute being the value of the corresponding coordinate. Plotting the hyperplane can exactly make the attribute separate from one another.

Figure 3 shows how the hyperplane has been created by considering the data points and how the plane is successfully separated the two different points and the margin distance is being measured.

e) **Logistic regression classifier**

The supervised machine contains a list of machine learning classifiers which are most effective in solving the problems; one such algorithm is known as the logistic regression classifier which works by considering the number of dependent and independent variables. The prediction of dependent attributes can be done based on the given set of independent instances of attributes. Mainly this classifier is used for classification and regression kind of problems where it brings the probabilistic values kind of output i.e. the output values lie between 0 and 1. This output is either categorical or discrete-valued outputs. The output of logistical regression is in the form of the S-shaped curve we call a logistic function which is responsible to predict two maximum values 0 or 1. The S-shaped curve always shows the likehood between the attributes [21]. This method is capable of identifying the most influencing attribute from the data given on which the comments can be concluded. Figure 4 shows the shaped curve for the logistic regression.

f) **Ada boost classifier**

Ada Boost or Adaptive Boosting classifier is one of the proposed classifiers of Yoav Freund and Robert Schapire in the year 1996. This is novel technique which is an integration of multiple low-performing classifiers to get the maximum accuracy out of those classifiers. The working of this classifier is done by initially setting the weight to classifiers and training the data model in every iteration in such a way that it has to predict the unobserved data accurately.

The working of Ada Boost is depicted in Figure 5. At the beginning level, Ada Boost takes some samples of training data randomly. Iteration is done on the training dataset by obtaining the proper prediction values of the previous data values. This classifier is going to give a greater value weight to those data points which are falsely classified by assuming that in the next iteration, these weighted points get greater importance in the classification process [22]. Not only the weights are assigned to those points which have been wrongly classified but also have been given greater importance to those classifiers which have classified the points with the highest accuracy. The iteration is

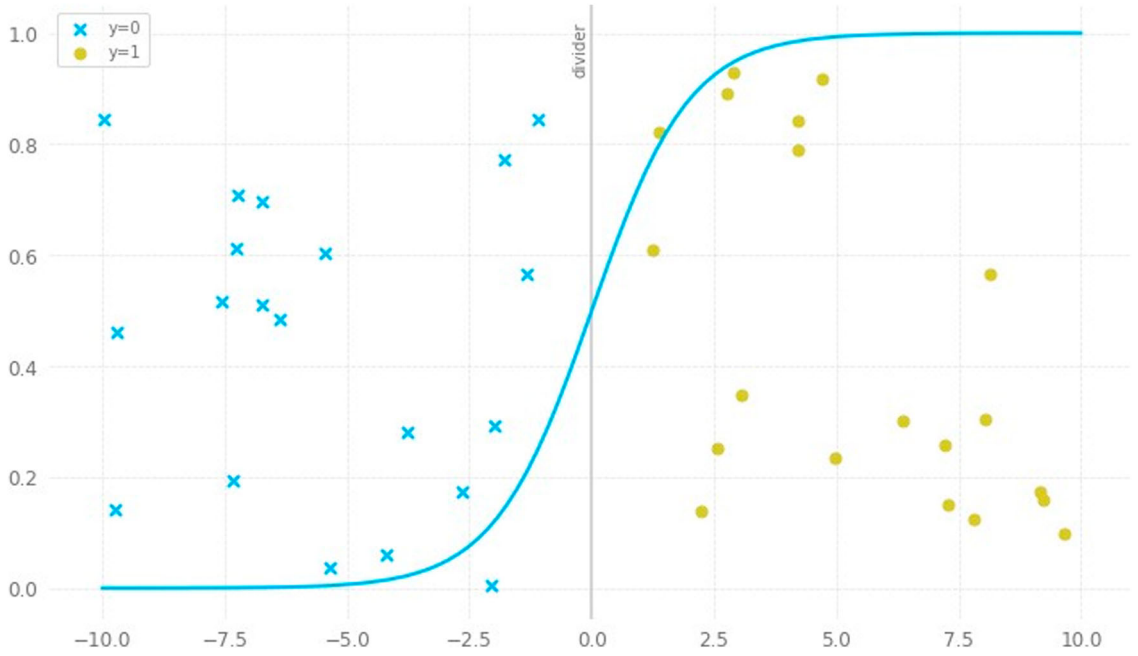

Figure 4. S-shaped curve for the logistic regression.

processed till reaching the level where no more error has been found.

g) **Sochastic gradient descent classifier**

There is an interconnection between random probability and the Stochastic Gradient Descent because the working way of the gradient descent is by selecting the probable value of a sample from a large number of datasets for each iteration. In this classifier, we can encounter a term called batch which is the total number of samples from a dataset which are used for calculating gradients. So the entire batch is taken as a whole dataset which is maximally useful for getting minima in a less noisy and less random manner but the only obstacle is when our dataset becomes big. This sample can be solved by introducing a selection process to select a single sample i.e. a batch size of one to perform each iteration [23].

5. Result and analysis

All the above-discussed classifiers are practically used on the obtained data set to analyse the risk factor and the most important attribute and the least important attributes in finding the final grade of the pupils. By these, we can prioritize the schedule of the pupil which has to be done and which should not be done or to be done infrequently.

a) **Sample of data preparation**

A huge amount of data on pupils has been collected from the schools and the colleges and the training institutions. The data obtained are not in the proper structure and not properly arranged and some missing value data attribute fields or been ignored and structured in a form which is supportive for training the model. Figure 6 shows the sample of the data that have been structured in the row and Coolum way which contains all the attributes which have been discussed in the section of questioners.

b) **Final grade distribution**

In the entire work, the aim is to predict the attributes which are most important and which are least important along with that we also need to decide to which category of students its most important and the least important so here the pupils are categorized into three classes based on their previous academic record into poor fair and the good. This can be statically shown in Figure 7 which shows the number of students who fall under which category.

The bar graph with the red colour indicated the number of students belonging to the poor class which means the students whose marks range from 0 to 40 is considered poor or danger zone. The bar graph in blue colour which indicates the pupils belongs to the fair class means the marks obtained by the pupils are in the range of 31–40. The bar with green colour indicates the student belongs to the class well with marks 41–100.

c) **Correlation heat map**

To predict the final grade of the pupils we need to have a clear idea of the variables which we have considered for the prediction. By considering these variables it is also important to understand the relationship between one

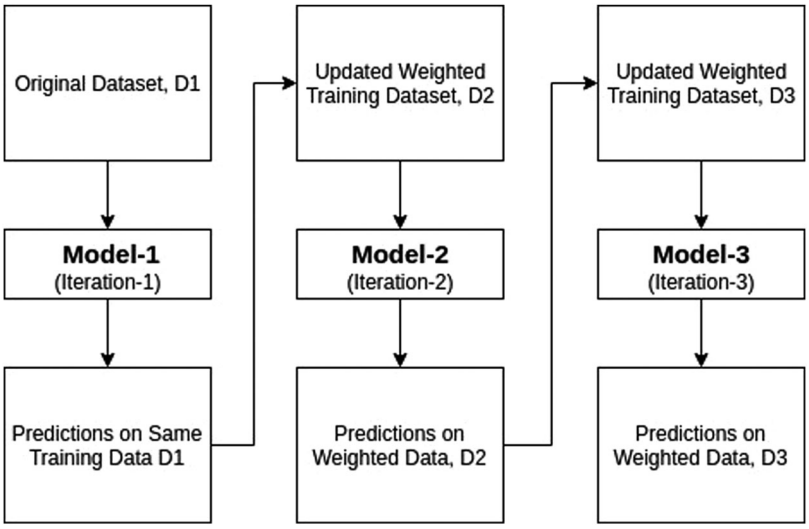

Figure 5. Ada boost classifier mode.

	school	sex	age	address	family_size	parents_status	mother_education	father_education	mother_job	father_job	...	free_time	go_out	weekday_alcohol_
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	3	4	
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	3	3	
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	3	2	
3	GP	F	15	U	GT3	T	4	2	health	services	...	2	2	
4	GP	F	16	U	GT3	T	3	3	other	other	...	3	2	

5 rows × 34 columns

Figure 6. Structured arrangement of the pupils' data.

variable and the other and how one attribute can affect the presence of another variable. For example, fathers and mothers' education has a direct impact on the children's final academic grade and the number of hours the pupil study has a direct impact on the final grade of the pupil. Also, the IQ level and the number of hours the student study per week vary from one student to another.

Finding all these relationships between the variable can be done by plotting a correlation graph using a heat map. The correlation heat map depicted in Figure 8 shows the mathematical relationship between the numbers of variables.

d) Final grade by romantic status

A romantic relationship is the one major attribute which should be mainly considered because pupils who have attained the phase of adolescent have a natural tendency of being attracted to the opposite gender. Whether the parents and society should allow them to have a romantic relationship or not is the major question that has to be answered here. If we allow them to enjoy the romantic relationship what is the effect on the final grade. Whether it is going to be boosted or it will be drained. Answers to this question can be done using the graph that is depicted below.

Figure 9 shows the final grade of the pupils belonging to all three classes. With the above graph, it is clear that the romantic status of the pupils has a high correlation with the final grade of the pupils. And this hypothesis with all three class such as poor fair and good have shown that romantic status will definitely affect the final grade of the student. By having the result of the chi-square test we obtained a score of around 0.038810128743959726, which clearly says that parents and teachers should definitely not allow their children and pupils to have a romantic relationship? Saying NO to the romantic relationship will boost the final grade at a good speed.

e) Final grade by alcohol consumption

Figure 10 The legal age to consume alcohol in India is between 18 and –25. This is the age when a lot of pupils need to be taken care of by their parents and professors. Again proper understanding of how effective the consumption of alcohol during this level of age and what is the impact on the final grade need to be analysed by the parents and the pupils to make their life better and

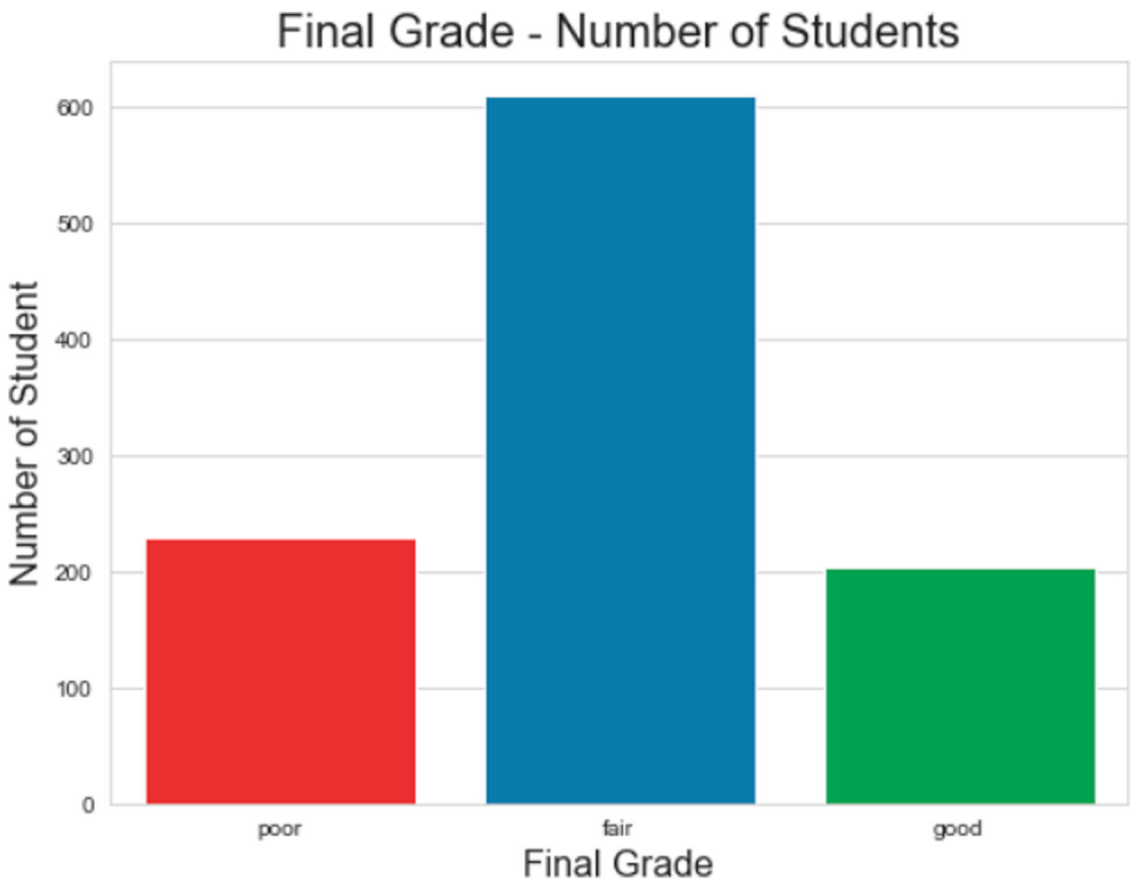

Figure 7. Bar graph showing the count of students belonging to their respective classes.

better both in academic grade and also in one's health point of view. The following depection of the graph will answer whether allowing the pupils to consume alcohol will improve their grades or not the bar graph of Figure 11 which shows the weekly consumption of alcohol by the three different classes of the student and their corresponding final grade in logarithmic percentage. After performing the hypothesis testing we came to know that there is a significant influence on the final grade of the variable called alcohol. With the help of chi-square test score of 0.007592294056368298 it is suggested to stay away from alcohol which can influence or affect the final grade of the pupils.

f) **Final grade by parents' education level**

Parents play an important role in the development of children in all aspects. The presence or absence of parents makes a major difference in one's life. In this work we need to find an answer to the question that does the education level of the parents will influence the final grade of the students or not. If it is influencing or related in some aspect again we need to answer the question that among the parents whose education is influencing more fathers or mothers. This question can be answered statically with the help of the graph that is shown in Figures 12 and 13 where Figure 12 shows the fathers' educational level of good and poor students and Figure 13 indicates the mothers' educational level with students class poor and good.

In Figures 12 and 13 where the red colour indicates the educational level of the student's father and mother belonging to good class and the graph with a blue peak indicates the education level of the student's father and mother belonging to the poor class. Number one indicates the very low level of education obtained by the parents and number five indicates the high level of education obtained by the parents in the graph. From Figure 14 OLS tells that parents' education level has a positive correlation with students' final scores. The mother's education level has a bigger influence than the father's education level.

g) **Final grade by the frequency of going out**

Students often used to go out with friends to enjoy each and every moment of their life at a young age and try to grab and learn the experience which is out of their academic world. While learning the things which are out of academic will help them score good marks in the final exam or not. Parents need to take the decision whether allowing their children to go out with their friends is good or bad with respect to their final exam point of view. The graph shown below will give a clear idea about

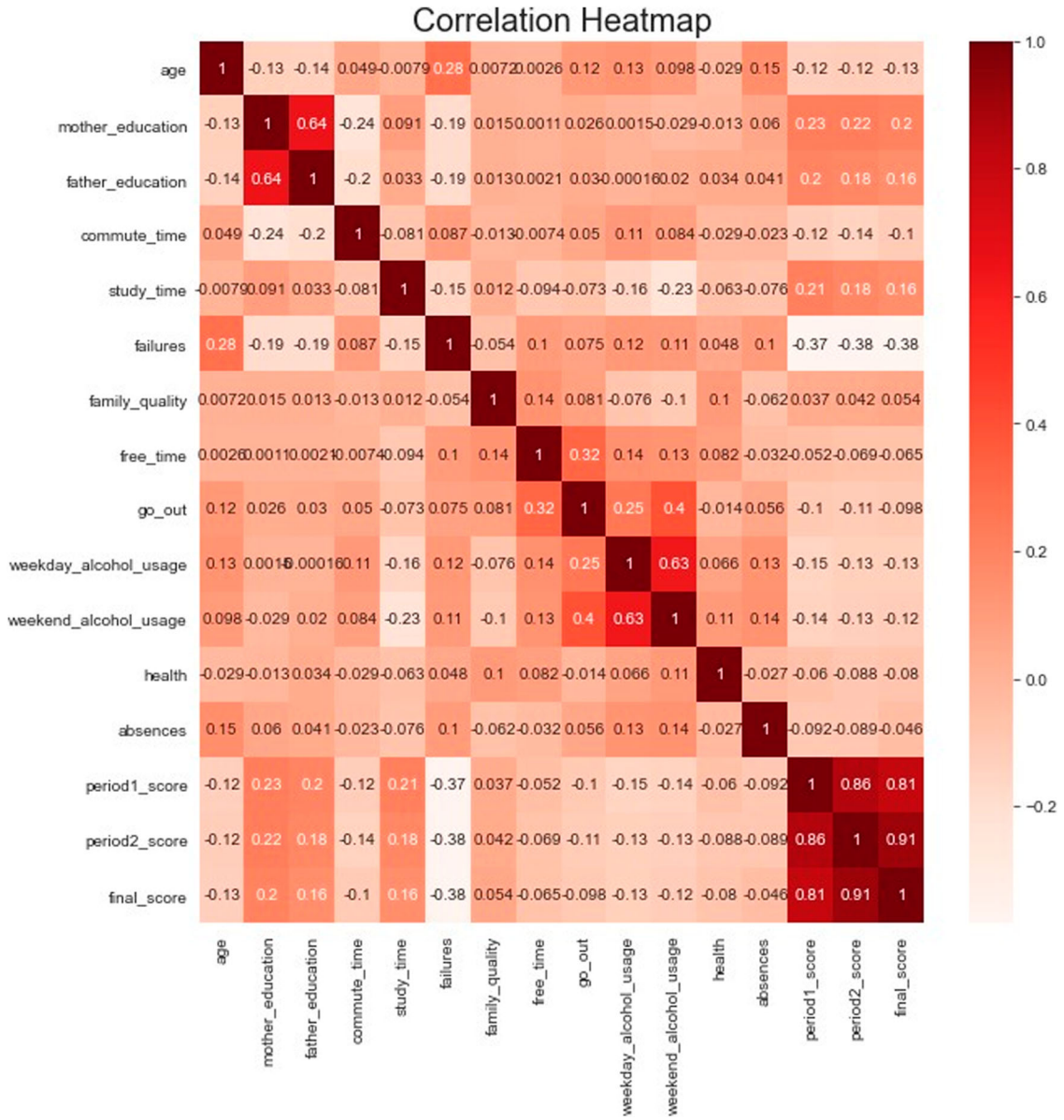

Figure 8. Depiction of dependency of variables.

whether frequently going out will help their kids to score more or less on exams is there any relationship between going out with friends and the final exam will be analysed using the below graph.

Figures 15 and 16 show the statistics of the pupils going out and their corresponding marks. The graph with a score of one indicates the students who go out very less and the graph with a score of five on the *x*-axis shows the students who go out more. And the y-axis shows the per cent of marks scored by the students who go out according to a scale of five.

Based on the chi-square score 4.365198328282194e-05 obtained i.e. Hypothesis Testing confirmed, the frequency of going out with friends has a significant correlation with students' final performance. So it is suggested to limit the frequency of going out to have a better final grade.

h) **Final grade by desire to go to college**

The desire to have something is the main force which drives the world. This desire is found to be more strong in the age group between 15 and 25. Parents and teachers who motivate their children to go for higher study and the students who have the natural tendency to go for higher education will study well to get good grades. This influence on the final grade is measured by the study time and the age group the pupil belongs to. Figure 17 shows the graphical representation where the pink colour lines show the desire to have higher education measured with respect to age and the study time of the pupils, and the yellow colour lines indicate the pupils with no desire of going to higher study. Based on the chi-square score 2.2470914018413168e13 obtained i.e. Hypothesis Testing confirmed that the desire to

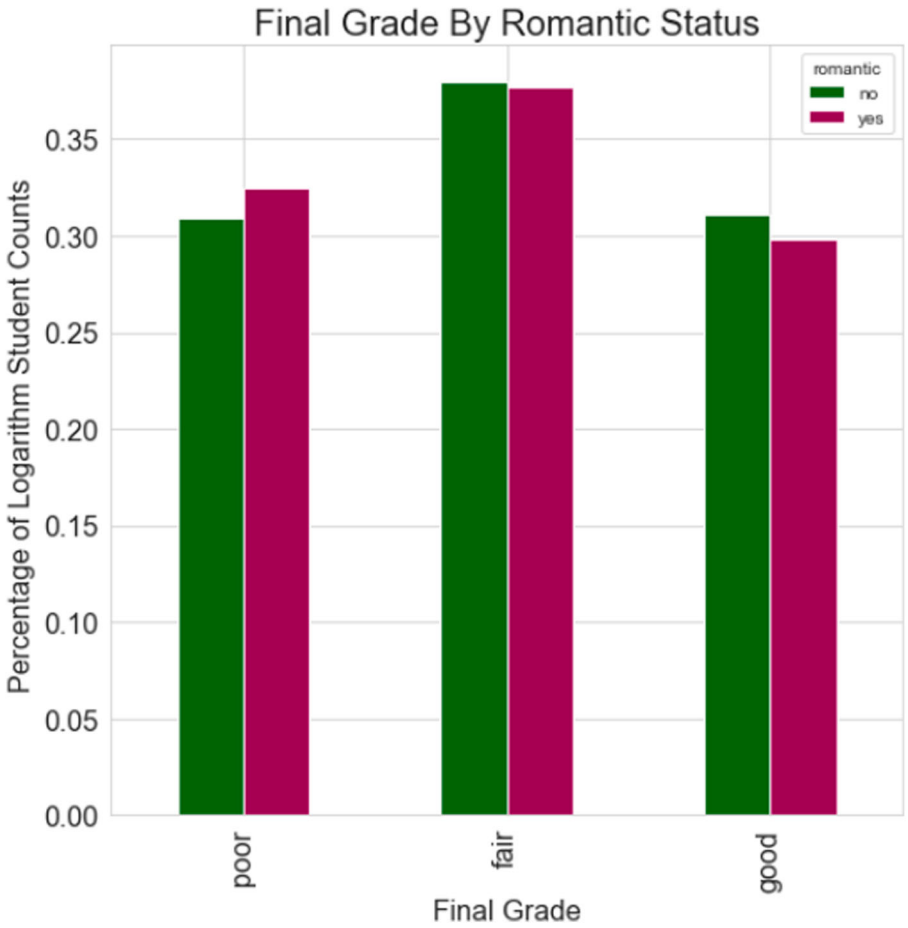

Figure 9. Final grade based on romantic status.

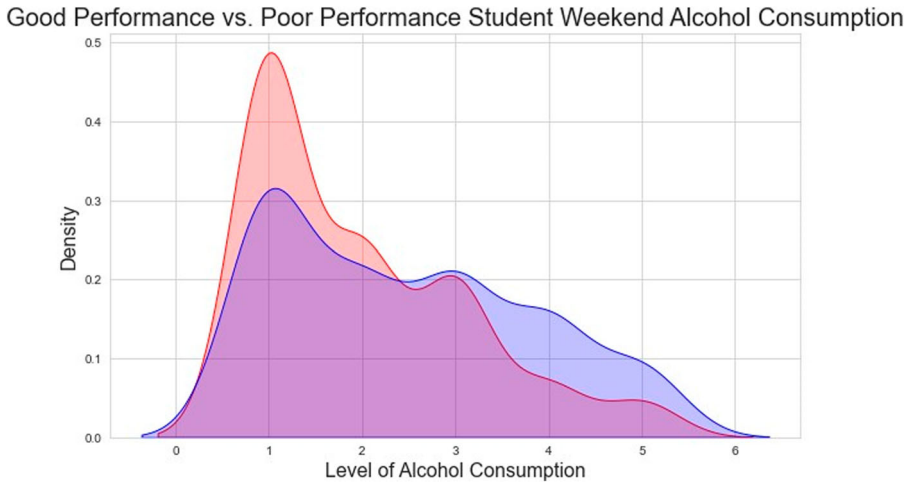

Figure 10. Density versus the level of alcohol consumption by the pupils.

go for higher study has a significant correlation with students' final performance. So it is suggested for the parents and the teachers to motivate their kids and fill them with the dream of going and achieving and having higher level goals and better final grades.

i) **Final grade by living area**

Most of the literatures has shown that the place where we leave has a greater influence on the way we leave the way we speak and behave and the facility which is available near the place or surrounding creates the resource for our leaving hood and the majority of the occupations can be created locally by using the resource that is available in our surrounding. Now in this study, we want to find if is there any kind of influence on the final grade of the pupils based on the place where they leave. Whether the family needs to be shifted to the nearest urban place from the village just to make their kids boost their grades or not. Answers to all these queries can be found by considering the graph which is depicted in Figure 18 where the bar graph shows that

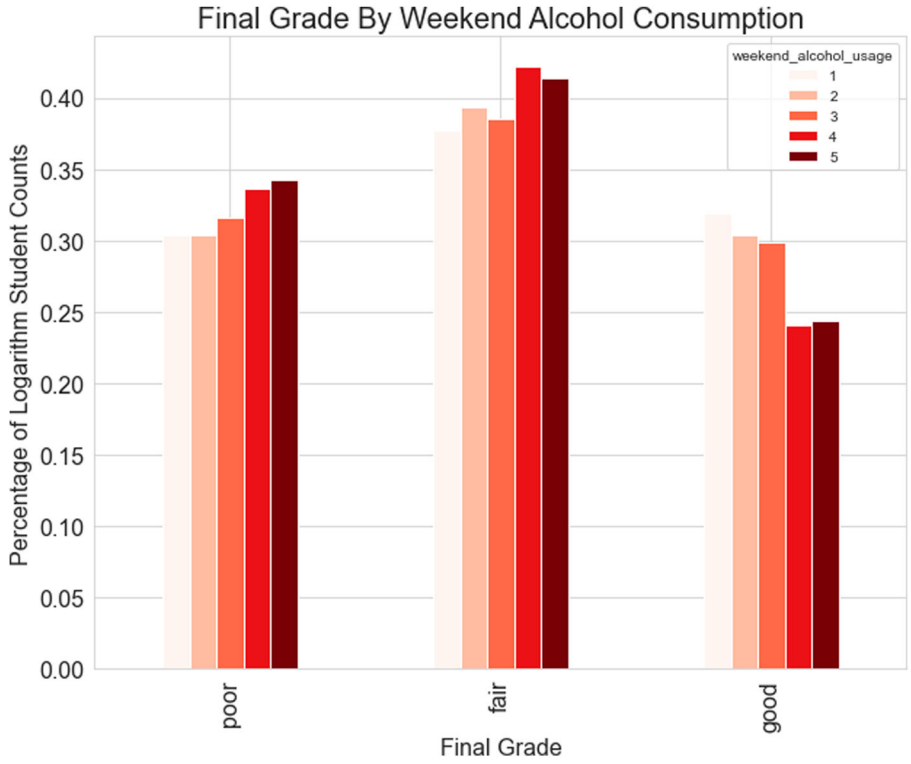

Figure 11. Final grade with respect to weekend alcohol consumption.

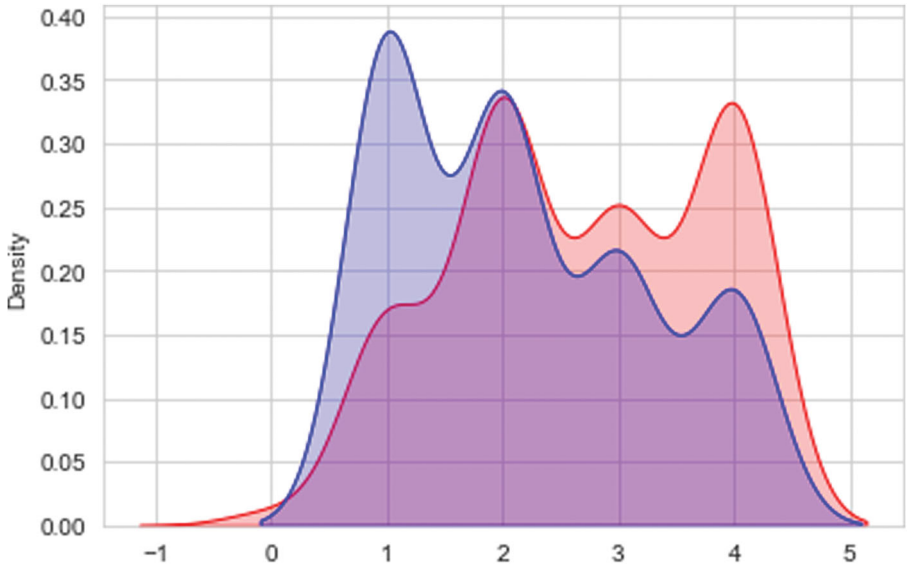

Figure 12. Fathers' education level.

the green colour indicates the rural and the red colour indicates the urban place this influence of place on the final grade is measured by using the final percentage of the students. Hypothesis Testing confirmed that the place where we leave has a significant correlation with students' final performance. The graph says that the students who score more or the toppers of the class are those students who stay in the city and who can have access to all kinds of resources and materials are the ones who score good marks. So it is suggested to stay in a place where all the resources are easily accessible to the students so that they can have very good final grades in their final examination.

j) **Use Students' information to predict their final grades by different classifiers**

In the process of constructing a model to predict the final grade of the students, it is very important to choose the classification algorithm which can classify the given data points more effectively. The first step involved in building a model for the final grade prediction is the

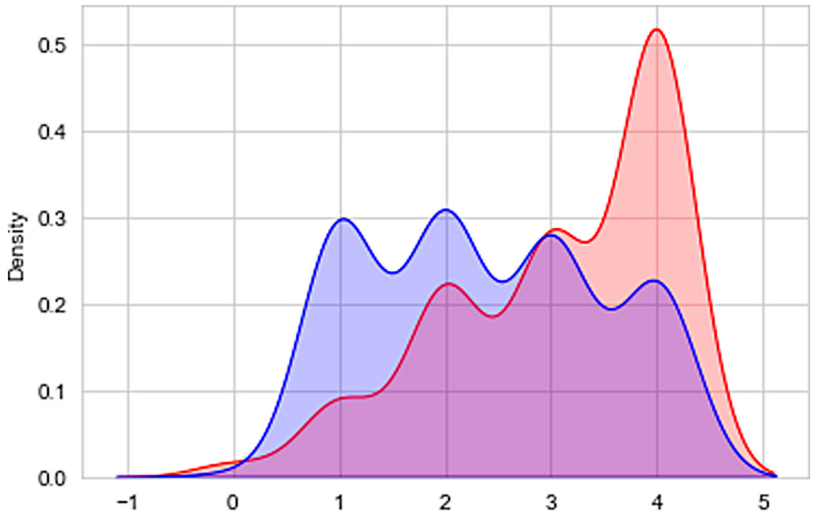

Figure 13. Mothers' education level.

OLS Regression Results

Dep. Variable:	final_score	**R-squared (uncentered):**	0.814
Model:	OLS	**Adj. R-squared (uncentered):**	0.814
Method:	Least Squares	**F-statistic:**	2287.
Date:	Sat, 08 Jan 2022	**Prob (F-statistic):**	0.00
Time:	16:15:24	**Log-Likelihood:**	-3194.8
No. Observations:	1044	**AIC:**	6394.
Df Residuals:	1042	**BIC:**	6403.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
mother_education	2.4078	0.166	14.527	0.000	2.083	2.733
father_education	1.5746	0.179	8.806	0.000	1.224	1.926

Omnibus:	35.858	**Durbin-Watson:**	1.631
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	39.773
Skew:	-0.427	**Prob(JB):**	2.31e-09
Kurtosis:	3.430	**Cond. No.**	5.72

Figure 14. OLS regression result analysis.

preparation of the dataset by the available dataset after which preprocessing stage is done. This preprocessing stage is common irrespective of selecting any classifier in building a model. Once the data set is preprocessed the next stage is to load the dataset and split the dataset into training and testing purposes and some part of the data set is considered for cross-validation where to increase the accuracy level of the prediction during the training phase. The spitted data will be fed as input into the model by calling the classifier it can be a decision tree, random forest, SVM, Ada boost, etc. After training the model with the dataset accuracy and classification measurement of each classifier is tabulated in Table 1 where we notice that logical regression has emerged as

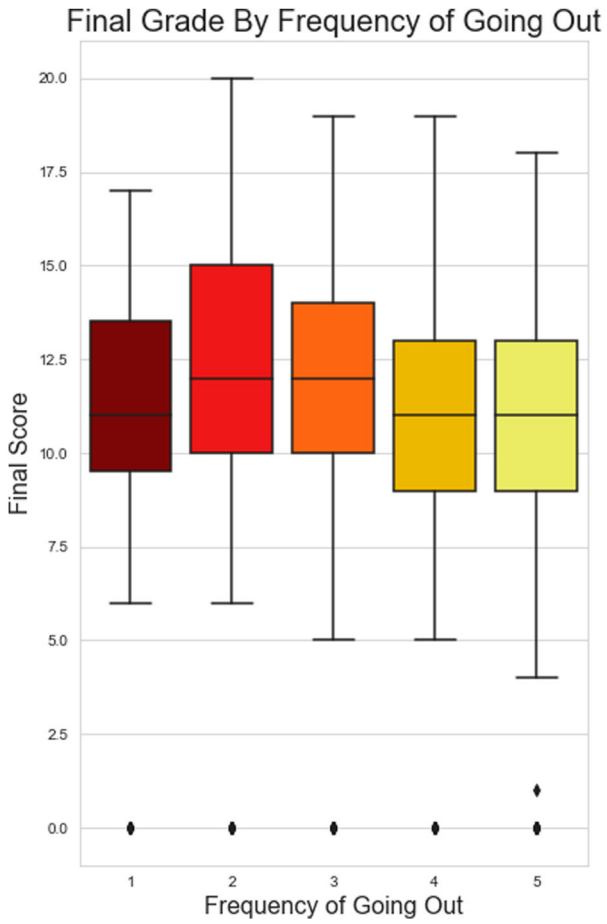

Figure 15. Pupils' frequency of going out.

the best classifier among all the other classifiers with a training model accuracy of about 0.887671 and a cross-validation accuracy of about 0.904459. The accuracy score of the other classifiers is shown in the table below.

6. Conclusion and future scope

The entire study is conducted to predict the valedictorian of the class and to find the risk factor associated with the pupils who belong to the poor class category. After analyzing all the necessary attributes that influence the final grade of the students we know that each valedictorian of the class has a decent profile which says that a pupil who is not in a romantic relationship, does not consume alcohol and lives in an urban area, does not go out with friends frequently and have a strong desire of receiving higher education and both parents received higher education and who studies more than 10 h weekly, healthy, no absences to classes are the ones who likely to emerge as a valedictorian of the class. Any pupil who follows and builds such kind of profile and any pupil who belongs to the poor class follows the above statements can shift himself from the poor class category to the good valedictorian category. Also we notice that among all the classifiers present in machine learning logistic regression emerged as the winner of the race by producing the model score of about 88.5%

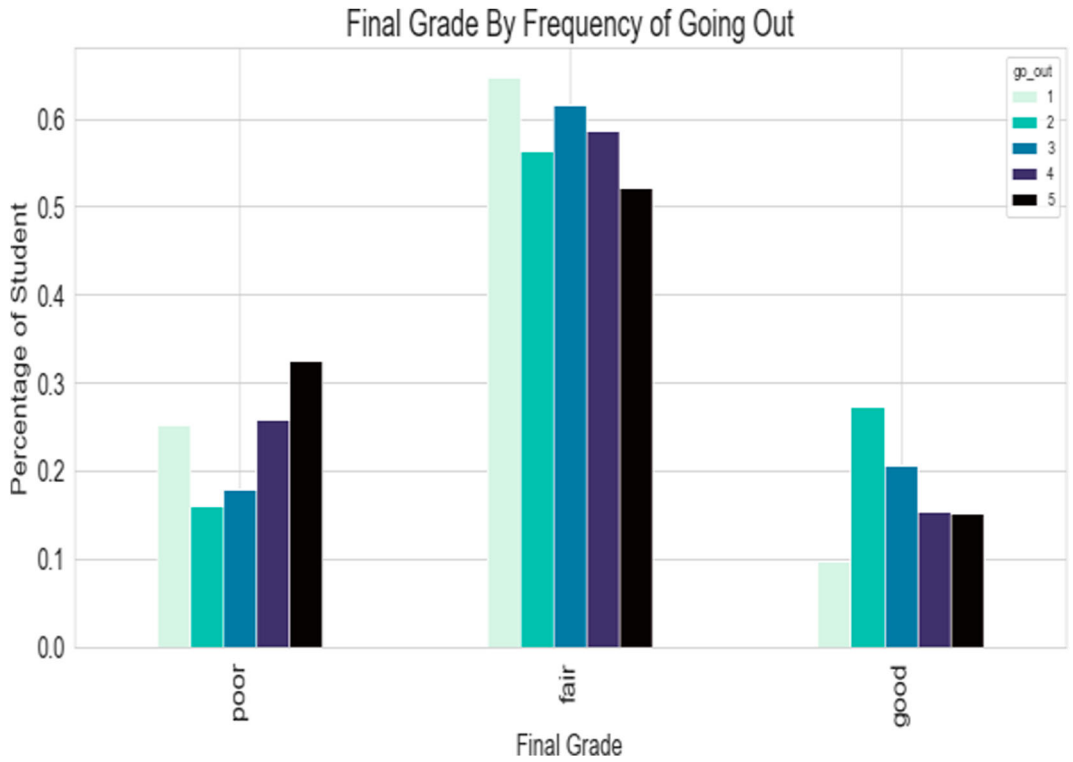

Figure 16. Pupils' frequency of going out and their respective percentage.

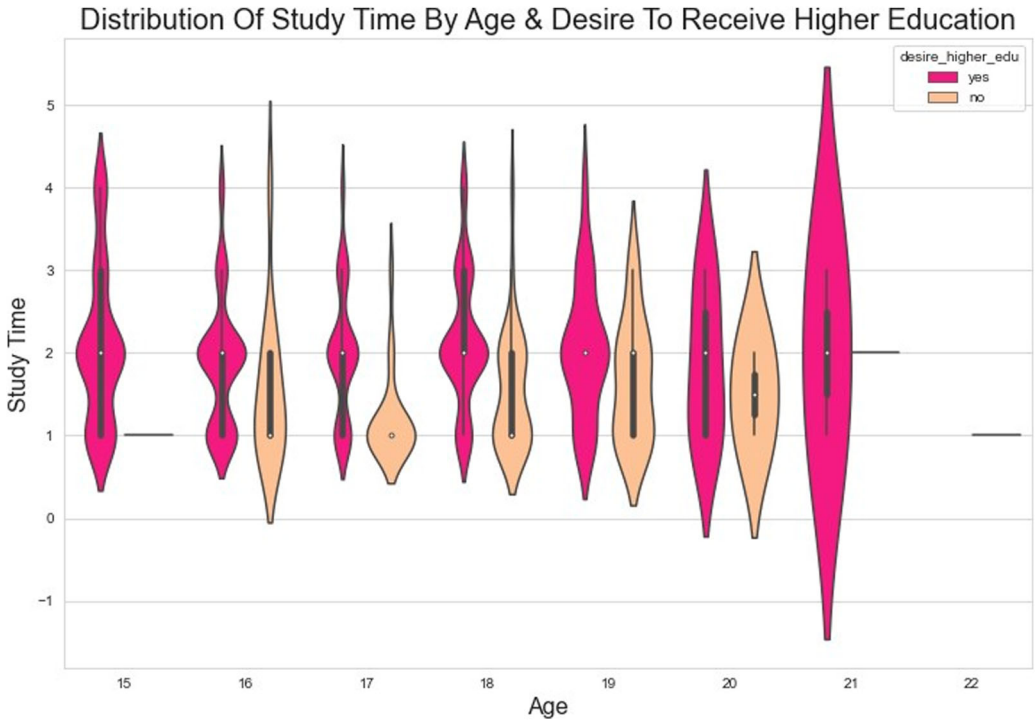

Figure 17. Distribution of study time by age and the desire to have higher education.

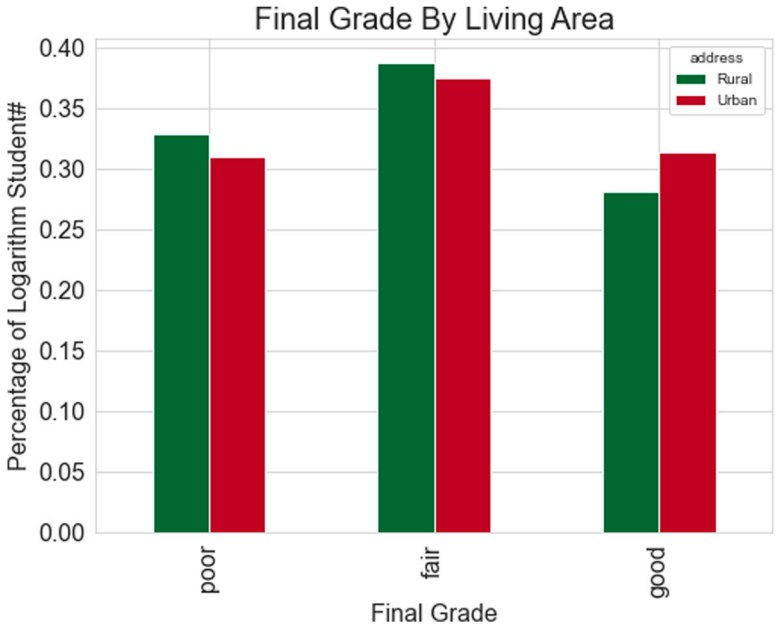

Figure 18. Final grade based on the living place.

Table 1. Accuracy score of the different classifiers.

Model	Decision tree	Random forest	Logical regression	SVM	Ada boost	SGD
Model score	0.88	0.98	0.88	0.93	0.86	0.85
Cross-validation score	0.89	0.86	0.9044	0.85	0.856	0.82

and the cross-validation score of about 90.44%. In the future, the study will be extended to find the variables and the factors which are influencing to do the post graduations and what after the completion of the PG what is the static score or the likeliness of one who is going to choose a teaching profession.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Aljohani NR, Davis HC. Learning analytics in mobile and ubiquitous learning environments. 11th World Conference on Mobile and Contextual Learning: mLearn 2012; 2012 Oct 15–17; Helsinki, Finland; 2012.

[2] Asif R, Merceron A, Pathan MK. Investigating performance of students: a longitudinal study. In: Fifth International Conference on Learning Analytics and Knowledge (LAK '15), New York, USA; 2015. p. 108–112.

[3] Chatti MA, Dyckhoff AL, Schroeder U, et al. A reference model for learning analytics. Int J Technol Enhanc Learn. 2012;4(5/6):318–331.

[4] Fournier N, Kop R, Sitlia H. The value of learning analytics to networked learning on a personal learning environment. In: 1st International Conference on Learning Analytics and Knowledge; 2011 Feb 27; 2011. p. 104–109.

[5] Lotsari E, Verykios V, Panagiotakopoulos C, et al. A learning analytics methodology for student profiling. Artif Intell Methods Appl. 2014;8445:300–312.

[6] Ma Y, Liu B, Wong CK, et al. Targeting the right students using data mining. In: 6th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD '00), New York, USA; 2000. p. 457–464.

[7] Minaei-Bidgoli B, Kashy DA, Kortemeyer G, et al. Predicting student performance: an application of data mining methods with an educational web-based system. In: 33rd Annual Frontiers in Education (FIE 2003); 2003 Nov 5–8; Westminster, CO; 2003.

[8] Mishra T, Kumar D, Gupta S. Students' employability prediction model through data mining. Int J Appl Eng Res. 2016;11(4):2275–2282.

[9] Osmanbegović E, Suljić M. Data mining approach for predicting student performance. Econ Rev. 2012;10(1): 3–12.

[10] Oyedotun OK, Tackie SN, Olaniyi EO. Data mining of students' performance: turkish students as a case study. Int J Intell Sys Appl. 2015;7(9):20–27.

[11] Pardos ZA, Heffernan NT, Anderson B, et al. Using fine-grained skill models to fit student performance with Bayesian networks. In: Antunes C, editor. Handbook educational data mining. CRC Press, Taylor & Francis group; 2010. p. 417–426.

[12] Arivudainambi D, Varun Kumar KA, Satapathy SC. Correlation based malicious traffic analysis system. Int J Knowl-Based Intell Eng Syst. Jan 1 2021;25:195–200.

[13] Ramaswami M, Bhaskaran R. A CHAID based performance prediction model in educational data mining. Int J Comput Scie. 2010;7(1):10–18.

[14] Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. Expert Syst Appl. 2007;33(1):135–146.

[15] Santos JL, Verbert K, Govaerts S, et al. Addressing learner issues with StepUp!: An evaluation. In: International Conference on Learning Analytics and Knowledge; April 2013. p. 14–22.

[16] Shalem B, Bachrach Y, Guiver J, et al. Students, teachers, exams and MOOCs: Predicting and optimizing attainment in web-based education using a probabilistic graphical model. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases; April 2014. p. 82–97.

[17] Sharabiani A, Karim F, Sharabiani A, et al. An enhanced Bayesian network model for prediction of students' academic performance in engineering programs. In: IEEE Global Engineering Education Conference (EDUCON); 2014 Apr 3–5; 2014. p. 832–837.

[18] Siemens G, Long P. Penetrating the fog: analytics in learning and education. Educause Rev. 2011;46(5):1–6.

[19] Slater S, Joksimović S, Kovanovic V, et al. Tools for educational data mining: a review. J Educ Behav Stat. 2016;42(1):1–12.

[20] Sree GS, Rupa C. Data mining: performance improvement In education sector using classification and clustering algorithm. Int J Innovative Res Dev. 2013;2(7): 101–106.

[21] Strecht P, Cruz L, Soares C, et al. A comparative study of classification and regression algorithms for modelling students' academic performance. In: International Educational Data Mining Society; 2015. p. 392–395.

[22] Yadav SK, Bharadwaj B, Pal S. Data mining applications: a comparative study for predicting student's performance. Int J Innov Technol Creat Eng. 2011;1(12):13–19.

[23] Zimmerman BJ, Kitsantas A. Comparing student's self-discipline and self- regulation measures and their prediction of academic achievement. Contemp Educ Psychol. 2014;39(2):145–155.