Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 8
August 2023

Prediction of Senior Secondary School Students' Academic Performance using Hybrid Machine Learning Algorithm

Praveena Chakrapani , A Anthonisan
School of Computing Sciences Hindustan Institute of Technology and Science
Chennai, India
School of Computing Sciences Hindustan Institute of Technology and Science
Chennai, India

Abstract-The immense growth of data in terms of volume, often presents an exciting challenge for the development of data analysis tools capable of detectingpatterns in this data. Data mining has established itself as a discipline that contributes tools for data analysis, knowledge discovery, and autonomous decision-making across a wide variety of application domains. One of these application domains isthe field of higher education. One of the key priorities of every higher education system is the review and improvement of educational systems in order to improve their services and satisfy the need of their students. Scholastic performance in the academics of school students is a growing concern in institutes for higher education, more than ever as predictive models constructed using classic quantitative methods do not generate reliable findings due to enormous volumes of data, attribute correlation, missing values, and variable non-linearity. However, data mining approaches function admirably in these circumstances. This paper proposes a novel hybrid algorithm, CNN/LVQ3, for predicting student academic performance.

1. Introduction

Data mining is a technology that has the potential to assist in the closure of the knowledge gap in the Higher Education System (HES) by predicting the scholastic academic performance of senior secondary school students using Machine learning algorithms. These enhancements may result in a variety of benefits for the higher educational system (HES), including decreased student drop-out rate, increased educational system efficiency, increased student promotion rate, transition rate, retention rate, educational improvement ratio, student learning outcome, and student success as well as cost savings associated with system processes. To accomplish the aforementioned quality enhancement, we require a data mining system capable of providing the necessary facts and insights to HES decision-makers.

Application of techniques such as, classification, prediction, clustering, and association when applied to HES, can help improve students' performance, their retention rate, their life cycle management, major selection, and course, an institution's grant/fund management.

Educational Data Mining (EDM) masters the challenges of sifting through substantial quanities of data in terms of the density of data and the rich variety of attributes through Machine Learning (ML) algorithms. Scholastic, Co-scholastic, Personal, Social and Demographic characteristics are some of many segments of attributes availablefor Academic Performance Prediction.

Numerous studies on data mining's varied applications in education have been done. Multiple authors have done literature reviews to guage the value of data mining in higher education, conducted from a domain-specific perspective on data mining applications.

This paper brings out a unique, specialized machine learning hybrid technique for the analysis and prediction of the scholastic performance of senior secondary school students using data attributes that could possibly influence a student's performance. Thereby, educational institutions can formulate a strategy to provide any required help to the students that are at the risk having low performance. This research project uses the data of 8000 students for prediction. The dataset contains a total of37 features of the students such as, their gender, nationality, parental status, access to subject information etc. Supervised Machine Learning models like DT, NB, SVM, MLP SGD and Ensemble models like LGBM, RF, AdaBoost, CatBoost, XBoost along with CNN and

LVQ3 algorithm have been used to predict the scholastic performance.

This project paper is divided into various sections, namely – Related Works, where we discuss about similar research done in predicting student's performance; Methods and Materials, where we discuss about the methods adopted for this research along with insight on the models used; Results section discuss about the result observed in this research based on various machine learning models used for prediction accuracy; Discussion section is where we will discuss about the result and finally in the Conclusion section we will summarize the inference and propose further possibilitiesfor research in this domain. References section covers the list of research papers referred for this research

Related Work

With latest development in the Information Technology, there is a lot of research that is on-going in the field of education. These researches explore suitable student attributes (features) that can be used in combination with associated ML/AI/Hybrid algorithms in order to devise models for accurate prediction of academic performance.

2.1. Online Classes

There is an increased demand for Online classes due to the increased convenience-needs in a fast-growing world. However, online classes also have some potential challengesthat may impact student performance. Analysing the student's scholastic performance under this study model has been done by various researchers using different criteria.

In researches like Ghassen Ben Brahim [15] and Tuti Purwoningsih et al [22] proposed a machine learning model to predict early academic achievement of fully online learning students. The author selected which are categorized into three main categories based on different criteria: (1) activity-type, (2) timing statistics, and (3) peripheral activity count. [15]

2.2. Hybrid Study Model

Hybrid study models combine both traditional classroom-based learning and online learning. However, hybrid study models also have some potential challenges that may impact student performance. Few studies [19] states that it can be extended to evaluate the performance of students in both physical and virtual educational settings. The authors have also noted limitations in their study.

2.3. Learning Management System

With improved Access to Learning Resources, an LMS can a variety of learning aids such as lectures, videos, e-books, and other materials. Slater et al. [3, 21] addressed why knowledge of diverse tools used in educational research is vital and required. In some researches, the data shows that the internal grades and GPA of students are the most important factors for prediction. The study found that classification approaches, particularly decision trees and Naive Bayes, are extensively used, but more research is required in the same sector.

Few researches [4, 16, 26] were based on assignments and quizzes for predicting students' ultimate grades based on early assignments and quizzes. The proposed approach helped forecast a student's course performance. Unlike the previous research, this investigation used data that the teacher considered to be available. For this reason, Moodle

[10] and other analytics [8, 9, 17] is a better option for instructors because they don't have to access the data themselves. Rather, the teacher can use Moodle based on data gathered from other course tools. Teachers, on the other hand, may struggle to grasp and apply this work. Because a specialist in engineering or computer science is necessary, work that academic professors may do is limited.

2.4. Co-curricular Activities Related Research

Since early times, combining co-curricular activities with academics have been known to enhance academic performance. Here are a few ways in which participating in co-curricular activities can benefit students: developing new skills, improving confidence, improving academic performance, enhancing social skill and boosting college applications. Shaikh Rezwan Rahman et al [25] proposed an approach on effects of co-curricular activities on student's academic performance. ML models

- RF, VT, MLP, LR were used and LR providing 99.52% accuracy.

2.5. Single Subject Research

Although, the influence of a subject on performance is mainly influenced by the teaching quality, individual student's interests and inherent ability also play a key role. Tsai et al. reported on a

case study at Taiwan's National University where they employed clustering to predict which undergraduate students will fail the university's computer competency test. The approach presented in this research aids the university in identifying these groupings [1, 2, 6, 5, 7]. Unsupervised neural networks were employed using K-means. To predict student achievement on additional university evaluations, such as English language tests, the "C5.0 decision-tree method" was used. The K-means algorithm out-performed neural networks and the "C5.0 decision-tree method" in terms of results.[7]

2.6. Behavioural Research

It is important to recognize that student performance is impacted by a complex set of factors, and that behavioural features are just one piece of the puzzle. However, by focusing on developing positive behavioural features in students, educators can help to create an environment that is conducive to academic success. Some ways in which different behavioural features can impact student performance [11, 24, 27, 30]: Motivation, Attitude, Study Habits, Learning Style, and Social Skills. In another research, focus is on identifying slow learners, done by Parneet Kaur et al [34] using Multilayer Perceptron (MLP) with 75% accuracy.

2.7. Grades and Mark Research

Positive feedback in the form of good grades can motivate students to continue working hard and performing well academically [12, 20, 31,32]. Few researches [18, 23, 28, 29, 33, 34] focus on student's who might fail in semester examination. In this study academic grade in different courses was used for prediction.

2. Methods And Materials

2.1. Proposed Workflow Diagram

The complete sequence of this research is exhibited in Fig. 1

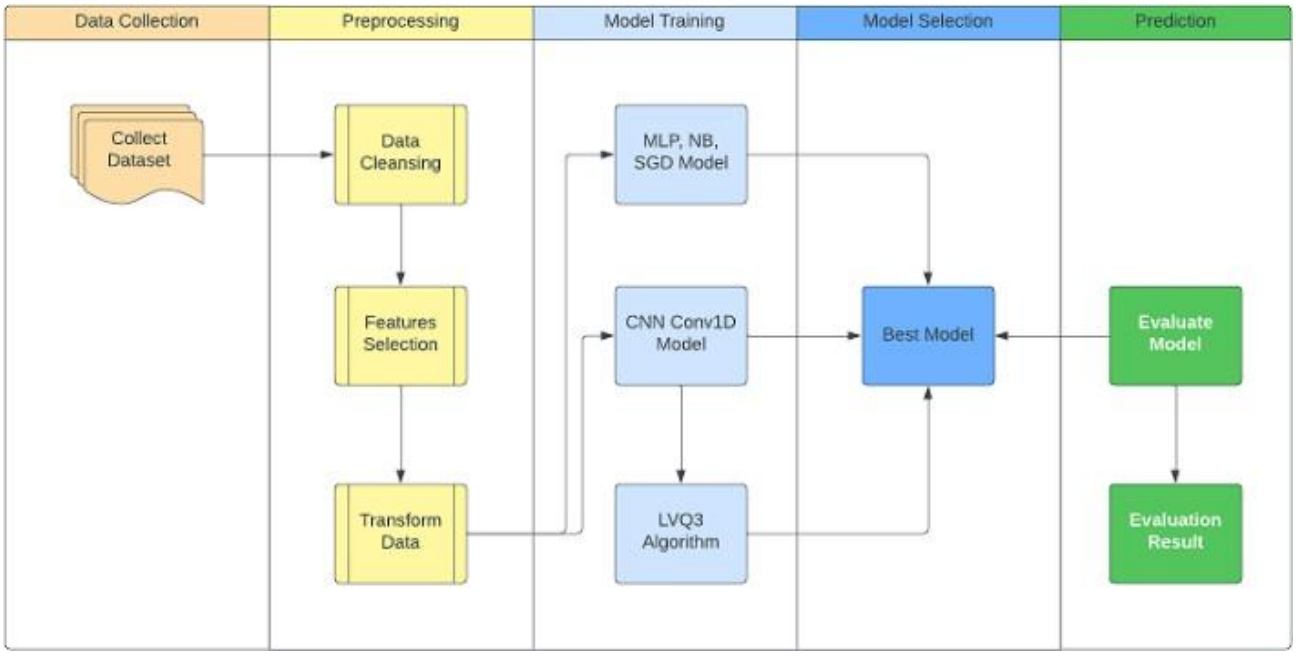


Exhibit 1 - Process Flow Diagram

The process encompasses six steps which are described as follows:

Step 1 – Data Collection: The dataset format used UCI Machine Learning Repository format. Identify the features with meaningful name.

Evaluation Criteria: In accordance with previous researches, the final grade is used to evaluate PASS or FAIL. For this simulation, it is considered that any grade above or equal to 9 is evaluated as PASS. Otherwise, the resultis evaluated as FAIL.

Step 2 – Dataset Exploration: Explore the data to find out each feature impact on the final prediction.

Step 3 – Dataset Cleansing: Make sure data integrity is maintained for better prediction.

Step 4 – Feature Selection: Find the list of features that has impact on the final prediction. Eliminate

the features that are not having any or significant impact.

Step 5 – Data Transformation: After feature selection, non-numerical data has to be transformed to numericaldata for prediction.

Step 6 – Prediction: Use the emerging machine learning models to train and test data to fabricate the final prediction. One of the models will become apparent as suggested model for student's performance prediction. The final phase of data pre-processing is to convert all values to quantitative values.

2.2. Dataset Collection and Identification

The data attributes include academic achievements, student population study, social and school related features and it was collected by using school reports and questionnaires. A total of 8000 student's records were used for this research.

TABLE I. DATASET SIZE

Data Source	Attributes	ecords (Total Students)
http://archive.ics.uci.edu/ml/machine-learning-databases/00320/	33 *	8000
* Refer below for list of attributes considered.		

There are totally 33 features available for selection. Table II provides brief description of each features considered for performance prediction.

TABLE II. DATASET DESCRIPTION

S. No.	Features	Expansion	Explanation
1	School	Student's school	GP, MS
2	Sex	Gender	Male, Female
3	Age	Student's Age	15 to 22
4	Address	Address type	Urban, Rural
5	Famsize	Family size	Less than 3, Greaterequal to 3
6	Pstatus	Parent's cohabitation status	Together, Apart
7	Medu	Mother's education	None, Primary, Elementary, Secondary, Higher
8	Fedu	Father's Education	None, Primary, Elementary, Secondary, Higher
9	Mjob	Mother's job	(at home, teacher, health, services, other)
10	Fjob	Father' s job	(at home, teacher, health, services, other)
11	Reason	Reason to choose the school	(home, reputation, course, other)
12	Guardian	Student's guardian	(mother,father, other)

13	Traveltime	Home to school travel time	(<15 min., 15 to 30 min., 30 min. to 1 hour, >1 hour)
14	Studytime	Weekly study time	(<2 hours, 2 to 5 hours, 5 to 10 hours, >10hours)
15	failuers	Number of past class failures	(nif 1<=n<3, else 4)
16	Schoolsup	Extra educational support	(yes,no)
17	Famsup	Family educational support	(yes,no)
18	Paid	Extra paid classes within the course subject	(yes,no)
19	Activities	Extra-curricular activities	(yes,no)
20	Nursery	Attended nursery school	(yes,no)
21	Higher	Wants to take higher education	(yes,no)
22	Internet	Internet access at home	(yes,no)
23	Romantic	Relationship	(yes,no)
24	Famrel	Family relationships	(from 1 - very bad to 5 - excellent)
25	Freetime	Free time after school	(from 1 - very low to 5 - very high)
26	Goout	Going out with friends	(from 1 - very low to 5 - very high)
27	Dalc	workday alcohol consumption	(from 1 - very low to 5 - very high)
28	Walc	Weekly alcohol consumption	(from 1 - very low to 5 - very high)
29	Health	Current health status	(from 1 - very bad to 5 - very good)
30	Absentism	Number of school leaves	(from 0 to 93)
31	G1	Score 1	(from 0 to 20)
32	G2	Score 2	(from 0 to 20)
33	G3	Final	(from 0 to 20)

2.3. Data Exploration

Exploratory analysis on the data is done in order to have a quick recapitulation of the istribution of the numerical data. Let's plot histogram on numerical data. Histogram shows how each feature maps against the target value.

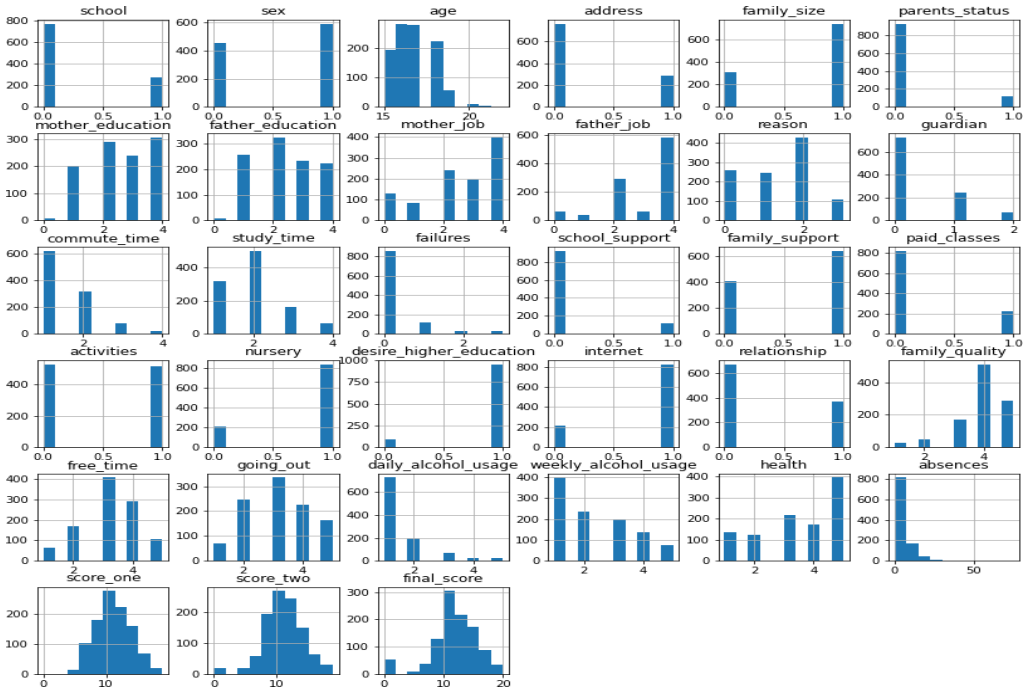


Fig. 1. Histogram of each feature vs final_score (target value)

2.4. Data Cleansing

After data exploration, data integrity is ensured. Integrity in this context here is to make sure that there is no missing values or duplicates.

This data cleansing step makes sure that any row(s) from the original raw dataset having missing value(s) or duplicate entry is removed and only the relevant data with all the values is kept for further processing. The missing values row were removed to avoid computational complexity.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10

Fig. 2. First 5 rows of the dataset

2.5. Machine Learning Algorithms

In this research project, we will be focusing on the below models to select the final recommended model forprediction.

- Base Classifiers
- Ensemble Methods
- Deep Learning (CNN)
- CNN with LVQ3 Algorithm

3.5.1. Base Classifiers

A total of 7 classifier models, namely DT, Logistic Regression (LR), SVM, MLP, SGD, KNN and NB were usedto build models using input dataset.

3. RESULTS

3.1. Environment

Jupyter Lab, Python Libraries for data mining were used to evaluate the proposed classification models to arrive at a conclusion. A structured sequential approach was conducted to determine the student performance.

Goal 1: Determine the relevant features that has impact on the students' performance.

Goal 2: Use cross validation with 10 K-Fold and compute the accuracy score of the models for final fitting.Eliminate less fitting model.

Goal 3: Perform comparative analysis of all the models - Base Classifiers, Ensemble, and Hybrid.

3.2. Feature Selection

For feature selection, we didn't use the Wrapper, Forward Elimination, Backward Elimination, Low Variance Filter, SelectKBest methods as each gave different set of features that could possibly give better prediction. The variance is much compared to common features among them.

In this research, final_score feature is used as target value. We are going to plot boxplot or bar chart or pie chart to find the impact of the independent characteristics.

After feature selection process, 18 aspects overall were used for training the models for prediction.

Features – **school, guardian, daily alcohol, weekly alcohol and romantic relationship** was not considered forfeature selection. We feel these are not applicable for predicting school students' performance. Fig. 4 represents dataset after feature selection.

4.2.1. Included Features

The below features are included in the final prediction features list. Their mapping against the final_score feature is shown below to know why they are included in the final feature selection list.

4.2.1.1. Age

As there is significant variance, it is included as part of included feature.

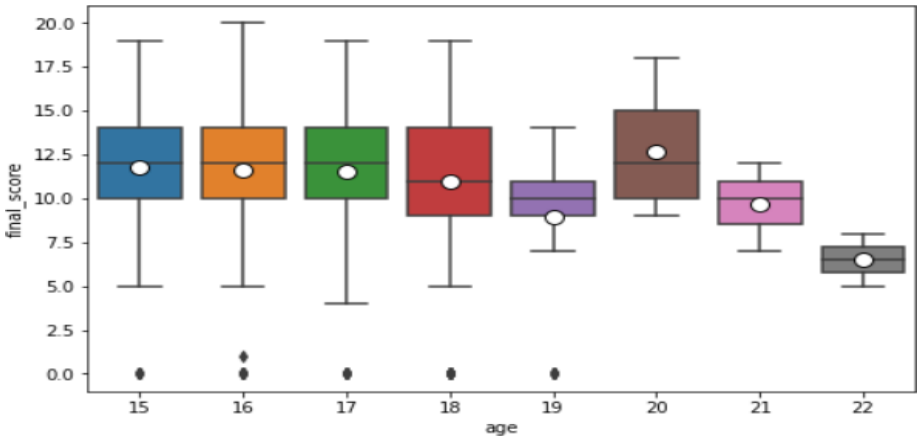


Fig. 5. Age impact on the students' performance

4.2.1.2. Address

As the residence of the student location has significant variance, it is included as part of included feature.

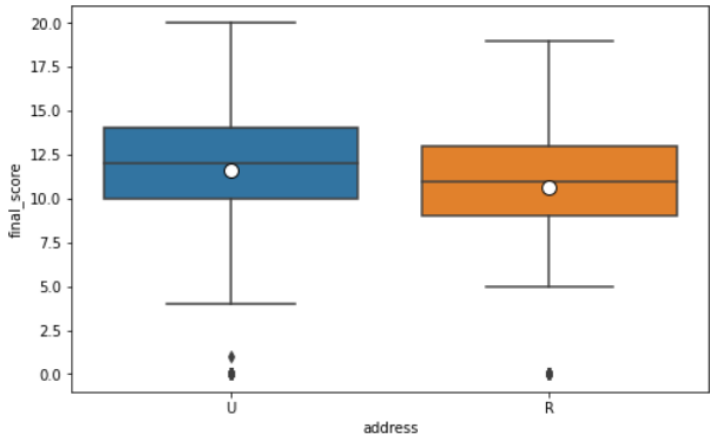


Fig. 6. Urban vs Rural place of stay impact

4.2.1.3. Mother's Education

This graph shows that Mother's education influences scholastic performance.

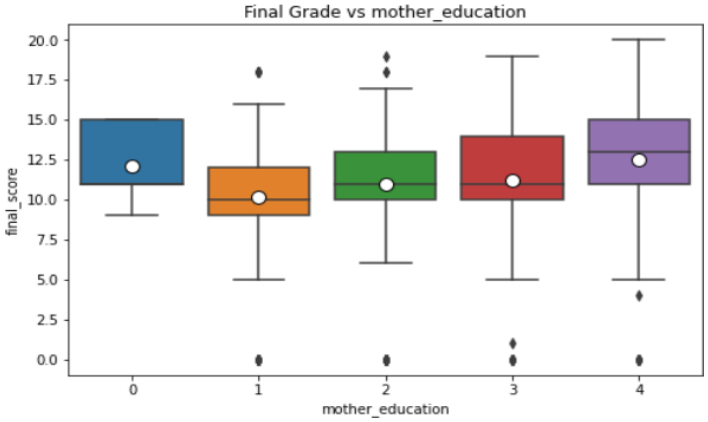


Fig. 7. Mother's Education impact on the scholastic performance

4.2.1.4. Father's Education

The graph indicates that Father's education influences scholastic performance.

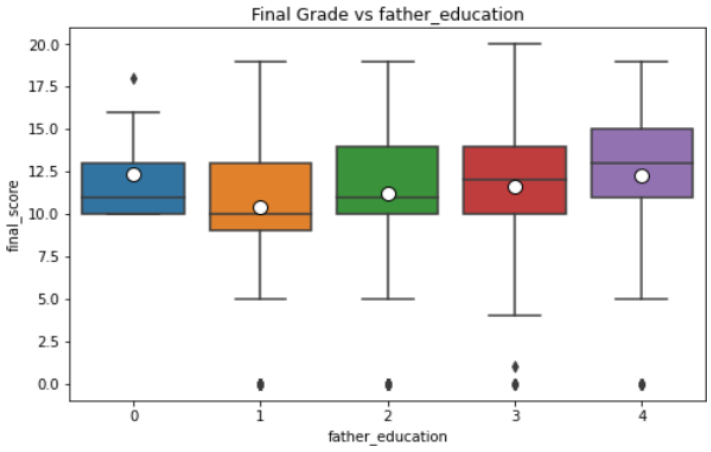
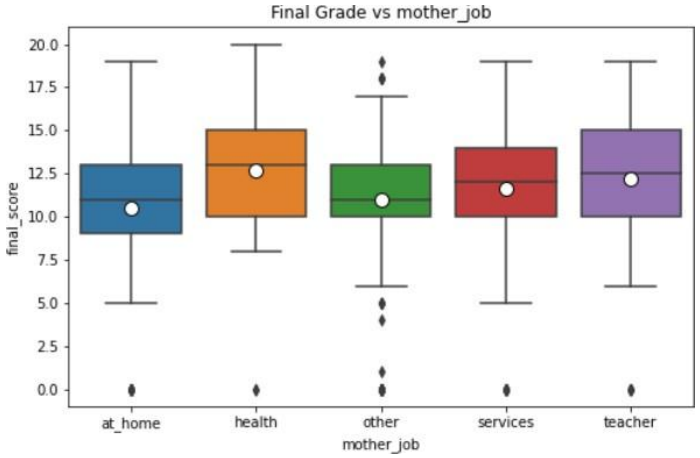


Fig. 8. Father's education impact on scholastic performance

4.2.1.5. Mother's Job

As there is much variance in mother's job, impacting the student performance, it is part of included feature.

Fig. 9. Mother's job impact on the student performance



4.2.1.6. Father's Job

As there is much variance in father's job, impacting the student performance, it is part of included feature.

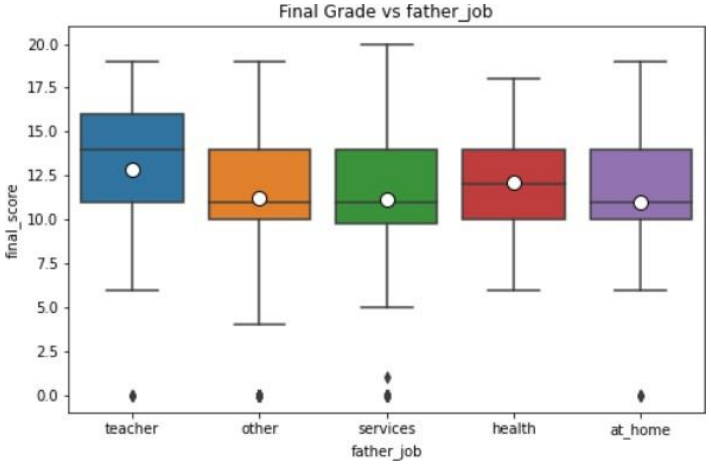


Fig. 10. Father's job impact on the student performance

4.2.1.7. Travel Time

As the students commuting less tends to score more, it is included as part of the final features list.

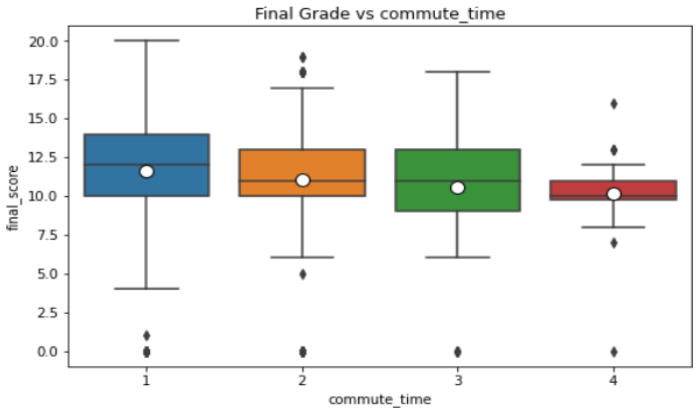


Fig. 11. Commute time impact on the student performance

4.2.1.8. Study Time

Student who spend more time to study, performs better.

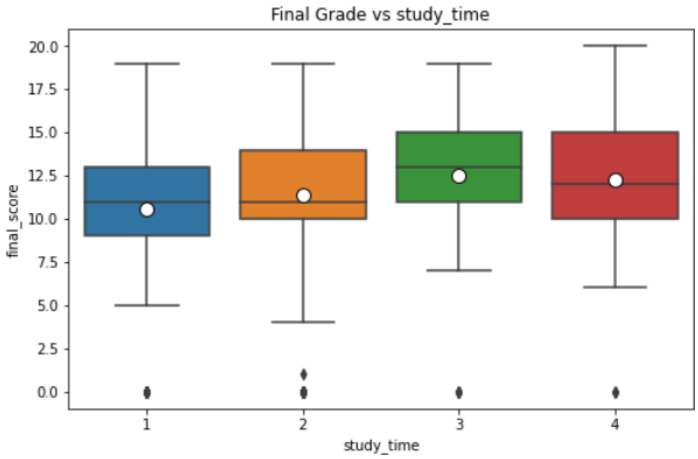


Fig. 12. Study time impact on the student performance

4.2.1.9. Failures

Student who do not have past failures record seems to score more.

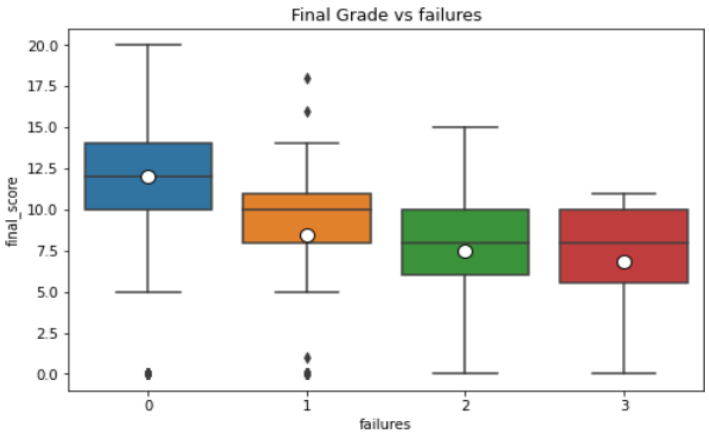


Fig. 13. Past failure impact on the student performance

4.2.1.10. School Support

Students with school support have much higher change of passing the exam.

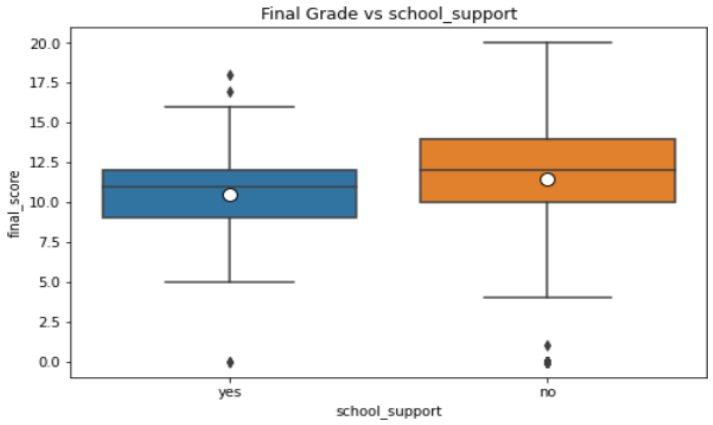


Fig. 14. School Support impact on the student performance

4.2.1.11. Higher Education

Students who have interest in higher studies seems to perform better.

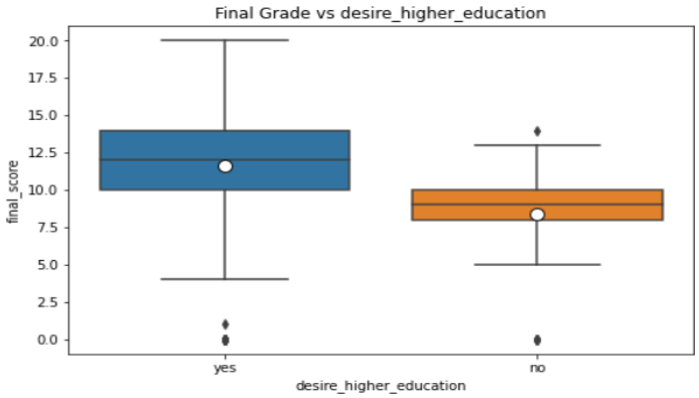


Fig. 15. Higher Education Interest impact on the student performance

4.2.1.12. Internet Impact

Internet access improves education.

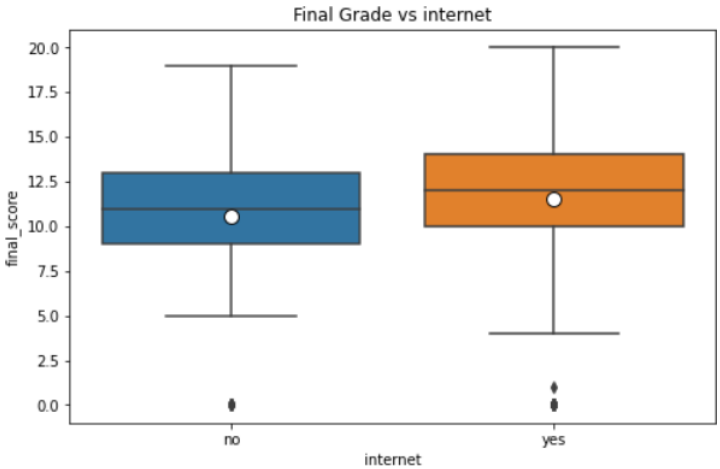


Fig. 16. Internet access impact on the student performance

4.2.1.13. Going Out

Students who go out more tends to score less.

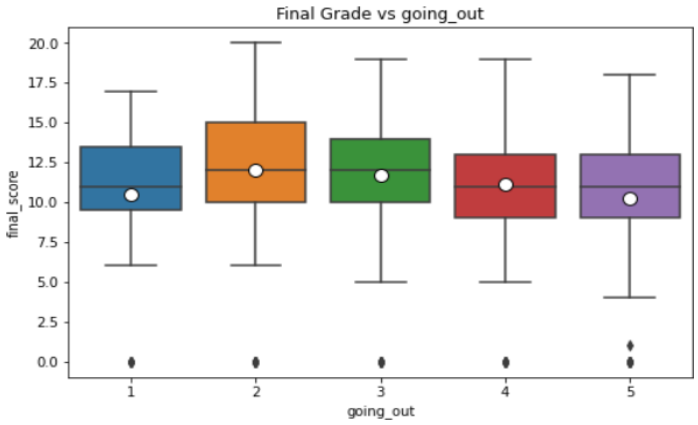


Fig. 17. Going out habit impact on the student performance

4.2.1.14. Study Time

Student whose health is better seems to perform better.

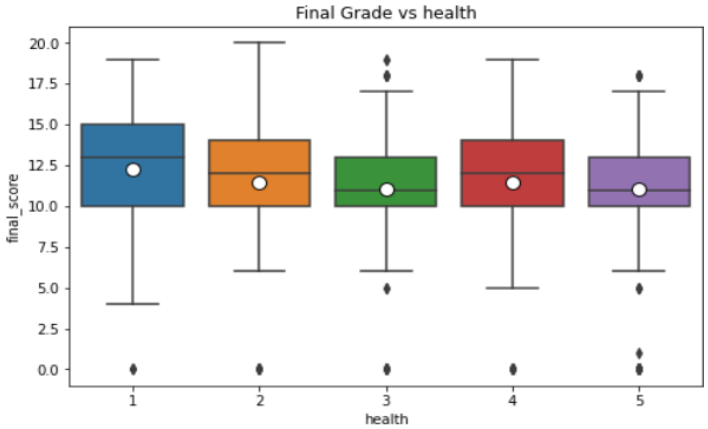


Fig. 18. Student's health impact on the student performance

4.2.1.15. FatherAbsences

Student who has less absences seems to perform better.

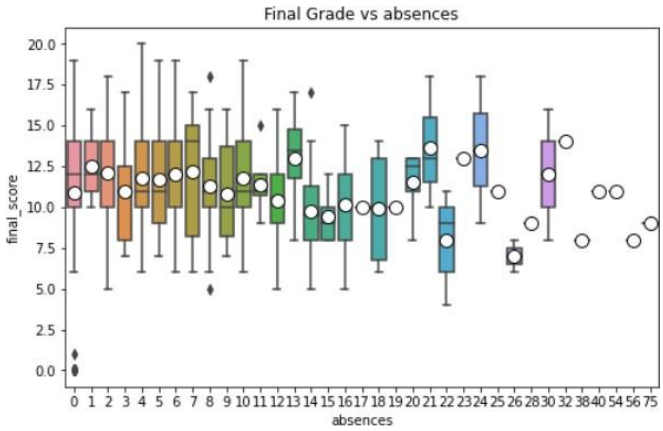


Fig. 19. Absences impact on the student performance

4.2.2. Excluded Features

The below features are excluded in the final prediction features list. Their mapping against the final_score feature is shown below to know why they are included in the final feature selection list.

4.2.2.1. Gender (Sex)

Student's gender is comparable and from the below graph it is clear that it has no impact on their performance.

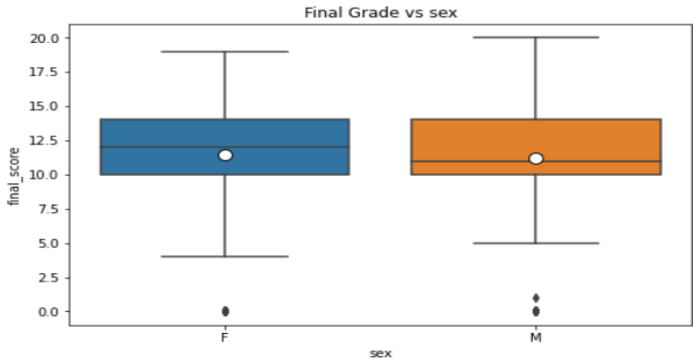
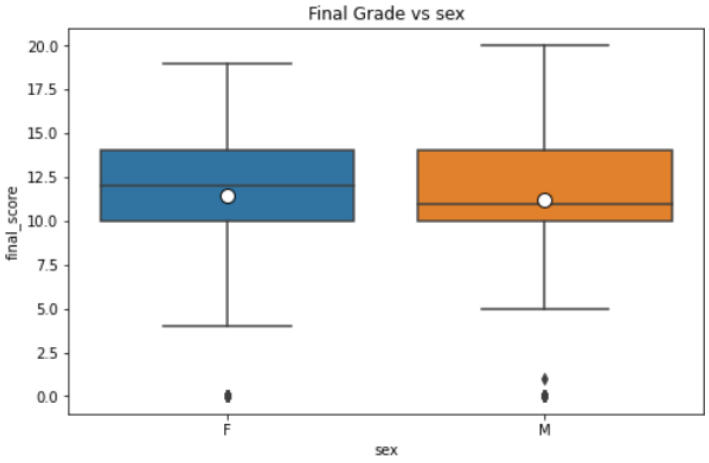


Fig. 20. Gender impact on the performance

4.2.2.2. Family Size

Almost negligible, from the below graph it is clear that family size of the student has not major impact on theirperformance.

Fig. 21. Family size impact on the performance



4.2.2.3. Parent Cohabitation)

Parent living status seems to have not impact based on the below graph.

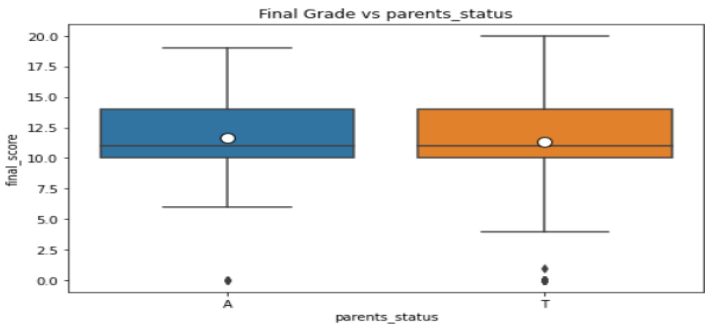


Fig. 22. Parent Cohabitation impact on the performance

4.2.2.4. Reason to join

Student reason to join the school shows no major impact on the performance from the graph below. This can beexcluded.

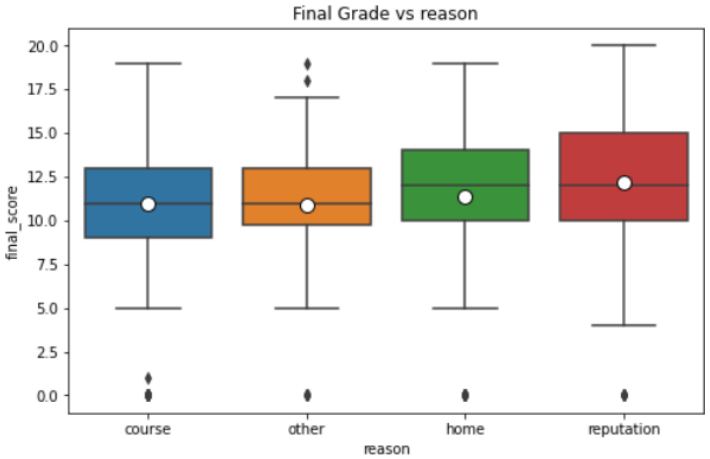


Fig. 23. Reason to join the school impact on the student performance

4.2.2.5. Family Support

Student's family support doesn't seems to have an impact on their performance.

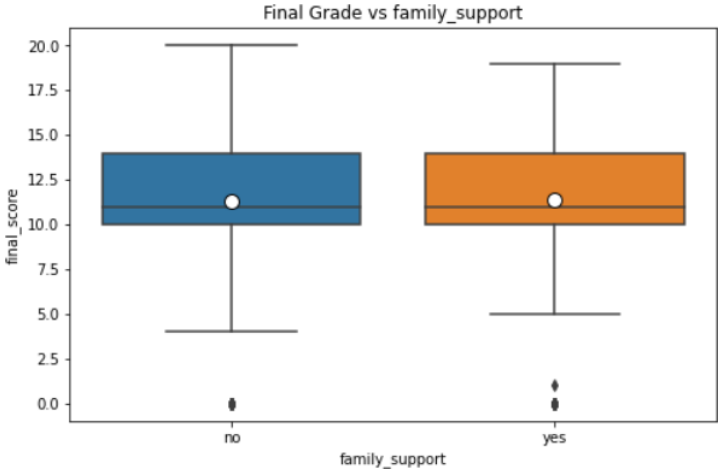


Fig. 24. Family support impact on the student performance

4.2.2.6. Paid Study

Student who pay for extra class seems to show not much improvement. This can be excluded from the feature list.

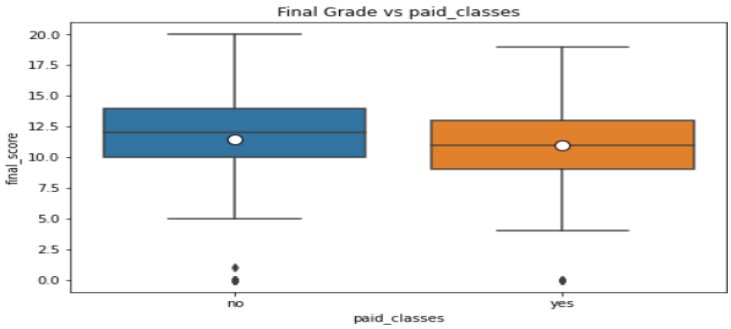


Fig. 25. Extra classes impact on the student performance

4.2.2.7. Other Activities

Students' who are involved in other activities seems to show no major improvement in their performance.

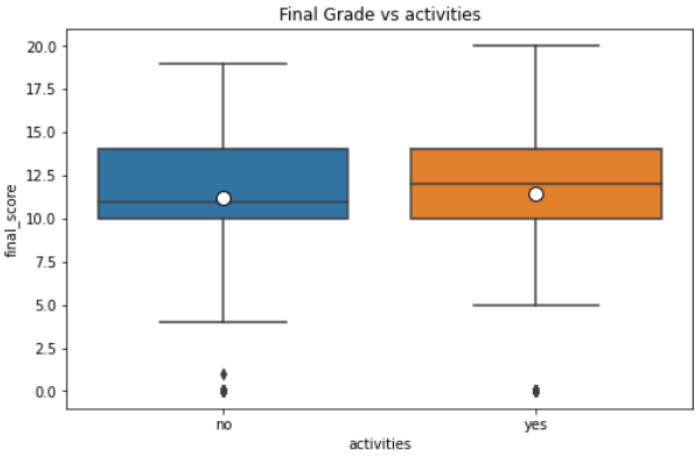


Fig. 26. Other activities impact on the student performance

4.2.2.8. Nursery Study

Students' who went to school from nursery seems to show no major impact on their performance.

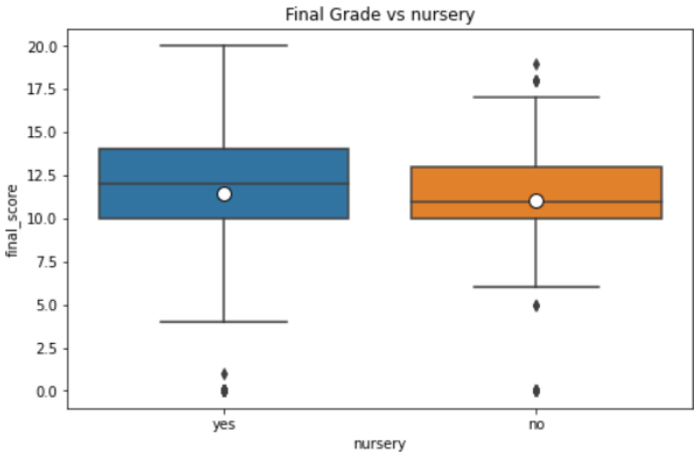


Fig. 27. Nursery study impact on the student performance

4.2.2.9. Family Relationship

Family relationship (quality) can be ignored as Family Size and Support is already excluded from studentperformance impact.

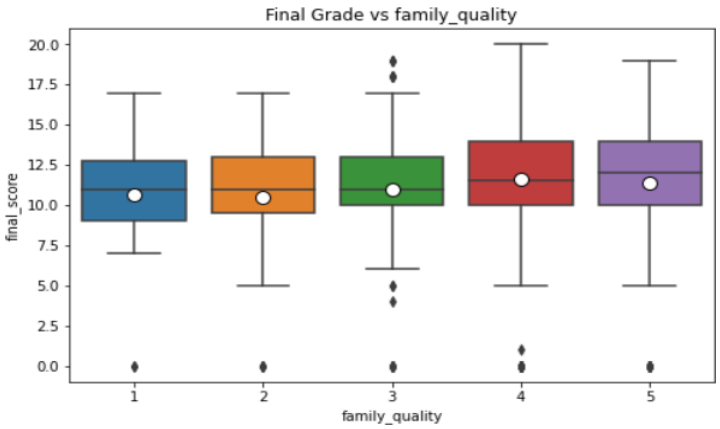


Fig. 28. Family relationship (quality) impact on the student performance

4.2.2.10. Free Time

Student free time is not having major impact and some students with less free time tends to score more.

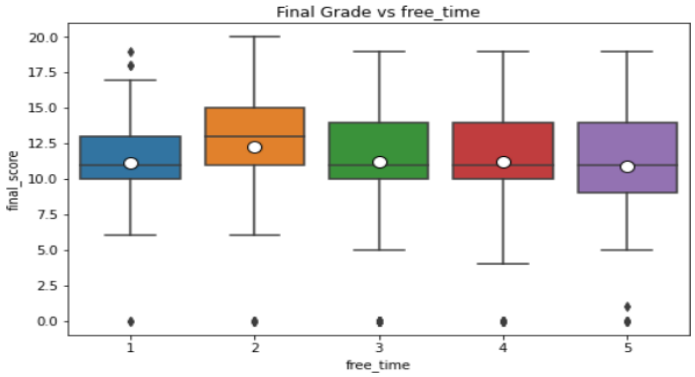


Fig. 29. Free time impact on the student performance

4.2.3. Features Correlation Map

The below figure shows how each features correlate to the final score. From this figure, we can see the positiveand negative impact features contributing to the student's performance.

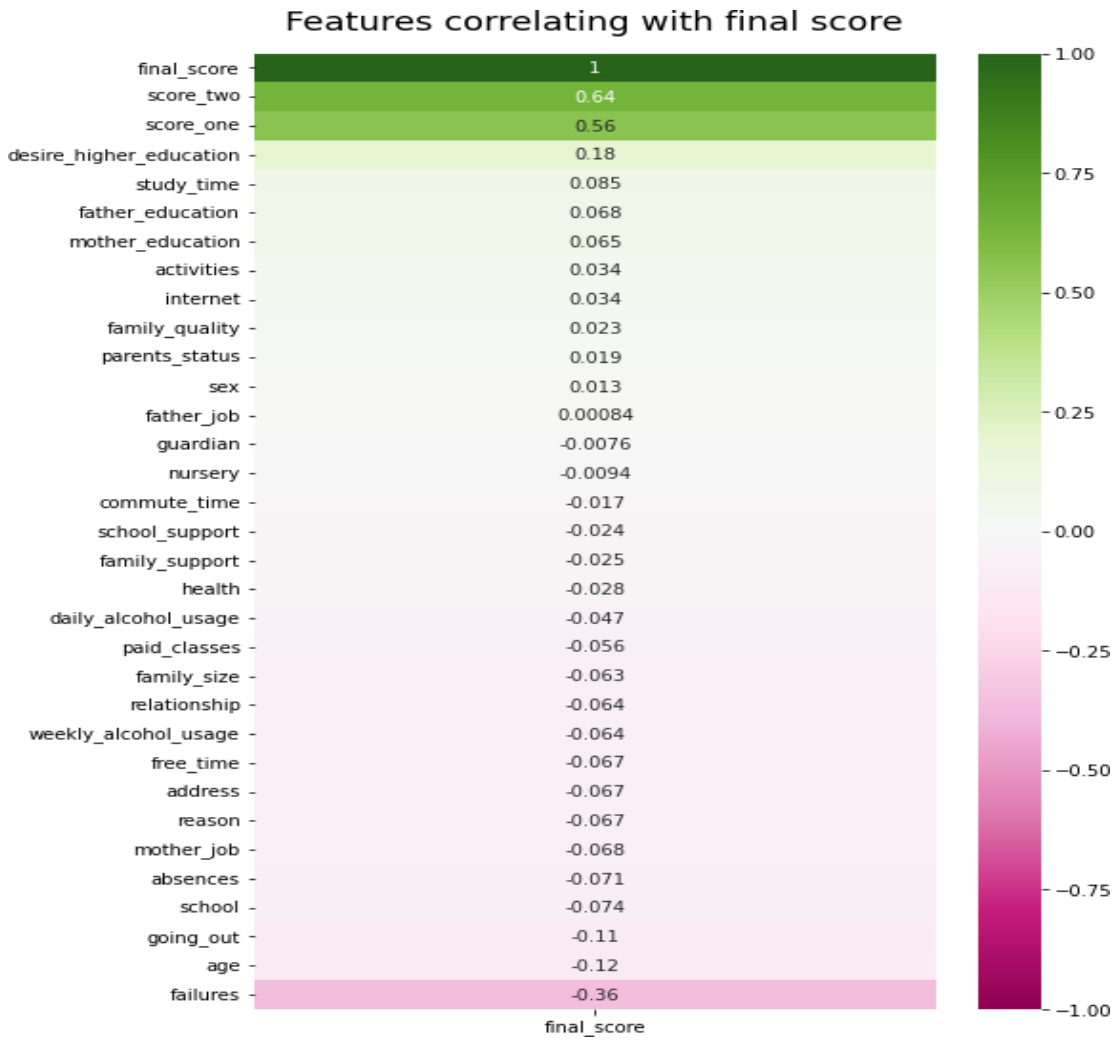


Fig. 30. Feature correlation map

3.3. Data Transformation

The final stage of data pre-processing converts all values to numerical values. For this purpose, we convert all the non-numeric features to numeric value by mapping them to corresponding number. For example, 'yes' and 'no' values will be mapped to1 and 0 respectively.

Select features has values that are numerical but the value is spread very wide. Explicit data transformation is done for those features so that they are in a reasonable range.

sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel
F	18	U	GT3	A	4	4	at_home	teacher	...	4
F	17	U	GT3	T	1	1	at_home	other	...	5
F	15	U	LE3	T	1	1	at_home	other	...	4
F	15	U	GT3	T	4	2	health	services	...	3
F	16	U	GT3	T	3	3	other	other	...	4

Fig. 31. Dataset before transformation

age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc
18	0	1	1	4	4	3	0	...	4	3	4	1
17	0	1	0	1	1	3	4	...	5	3	3	1
15	0	0	0	1	1	3	4	...	4	3	2	2
15	0	1	0	4	2	1	2	...	3	2	2	1
16	0	1	0	3	3	4	4	...	4	3	2	1

Fig. 32. Dataset after transformation

mother_education	father_education	mother_job	father_job	study_time	failures	school_support	desire_higher_education
4	4	3	0	2	0	1	1
1	1	3	4	2	0	0	1
1	1	3	4	2	3	1	1
4	2	1	2	3	0	0	1
3	3	4	4	2	0	0	1

Fig. 33. Dataset after feature selection

4.3.1. Feature – Score_One and Score_Two

TABLE III. SCORE ONE AND TWO MAPPING

Grade Range	New Value	Grading	Description
18 – 20	6	A+	Excellent
16 - 17	5	A	Very Good
14 – 15	4	B	Good
12 – 13	3	C	Satisfactory
9 – 11	2	D	Sufficient
0 – 8	1	E	Fail

4.3.2. Final Score Mapping

TABLE IV. FINAL SCORE MAPPING

Grade Range	New Value	Description
9 – 20	1	Pass
0 – 8	0	Fail

4.3.3. Age Mapping

TABLE V. AGE GROUP MAPPING

Age	New Value	Description
< 17	0	Age group less than 17
< 19 and 18	1	Age group between 17
>= 19	2	All other ages.

4.3.4. Absences Mapping

As absences value is spread out very widely, it is better to group them in ranges. The below table depicts how they are grouped.

TABLE VI. ABSENCES MAPPING

Absences	New Value
0 – 3	2
4 – 7	6
8 – 10	9
10 – 15	13
16 – 20	18
21- 25	23
>= 26	30

3.4. Model Checking

K-fold cross-validation is used to guage the model's performance.

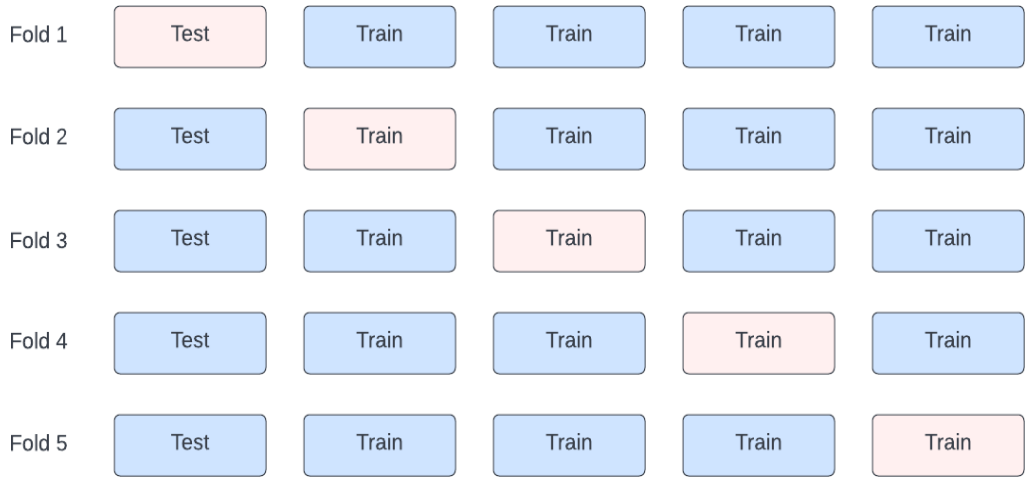


Fig. 34. K-Fold Approach

All data is used to check the model score. It has to be noted that cross validation approach will not retain the model as predicting the score. For actual model building we will be using 80:20 train-test split of data. Fig. 37 represents K-fold result.

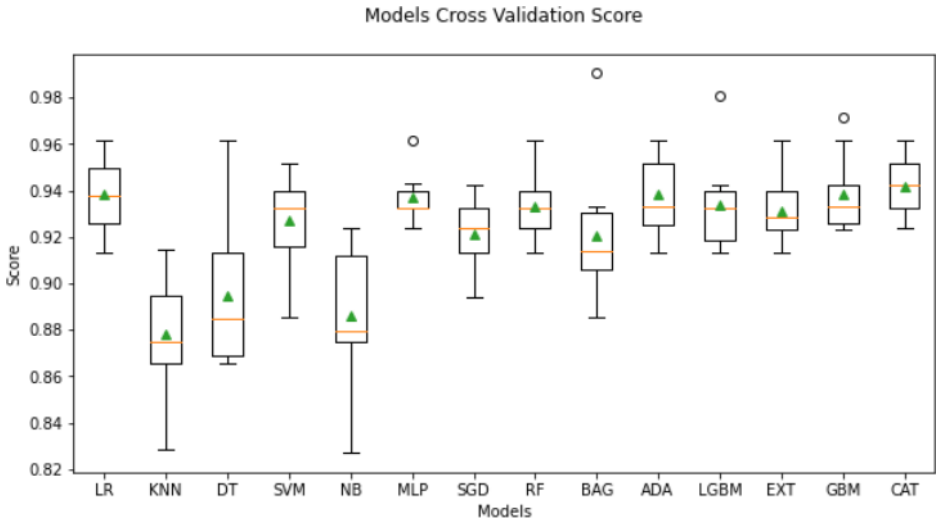
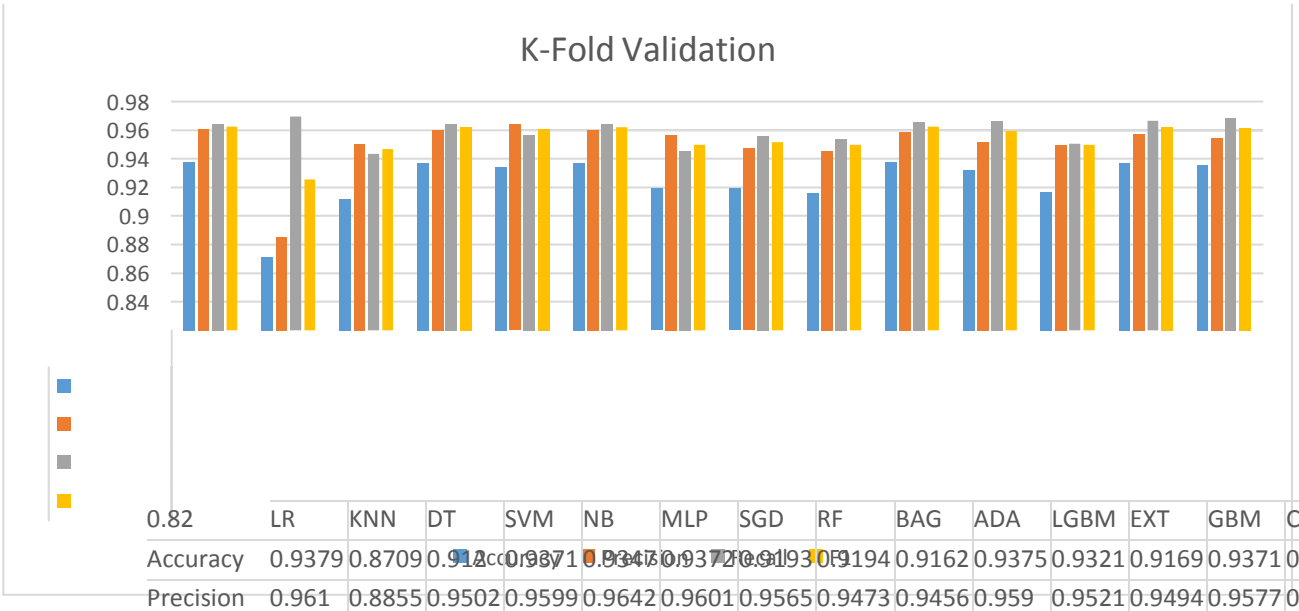


Fig. 35. Model Fitting using K-fold Approach

3.5. *Model Fitting*

In this research experiment, first we want to determine whether the dataset fits into the models selected for prediction. Instead of using train and test split of data from the source dataset, we want to use all the data. The 10 K-Fold cross-validation approach has been used to determine how the fitting of data into the selected models. By this way every data is part of train and test subset.

This is done to achieve the goal of finding whether the model fits the data before model building. The evaluation result showed all the base and ensemble models (hybrid is excluded) fitting the data and no over fitting happened. The cross-validation score is 90% or above in all the models. The score could appear less because it is mean of 10 folds of the entire dataset.



	LR	KNN	DT	SVM	NB	MLP	SGD	RF	BAG	ADA	LGBM	EXT	GBM	CAT
Accuracy	0.9379	0.8709	0.912	0.9371	0.9347	0.9372	0.9193	0.9194	0.9162	0.9375	0.9321	0.9169	0.9371	0.9355
Precision	0.961	0.8855	0.9502	0.9599	0.9642	0.9601	0.9565	0.9473	0.9456	0.959	0.9521	0.9494	0.9577	0.9545
Recall	0.9642	0.9697	0.9434	0.9644	0.9569	0.9644	0.9459	0.9561	0.9538	0.966	0.9668	0.9504	0.9668	0.9684
F1	0.9626	0.9256	0.9467	0.9621	0.9605	0.9622	0.95	0.9516	0.9496	0.9624	0.9593	0.9499	0.9622	0.9613

Fig. 36. Model Fitting using K-fold Approach

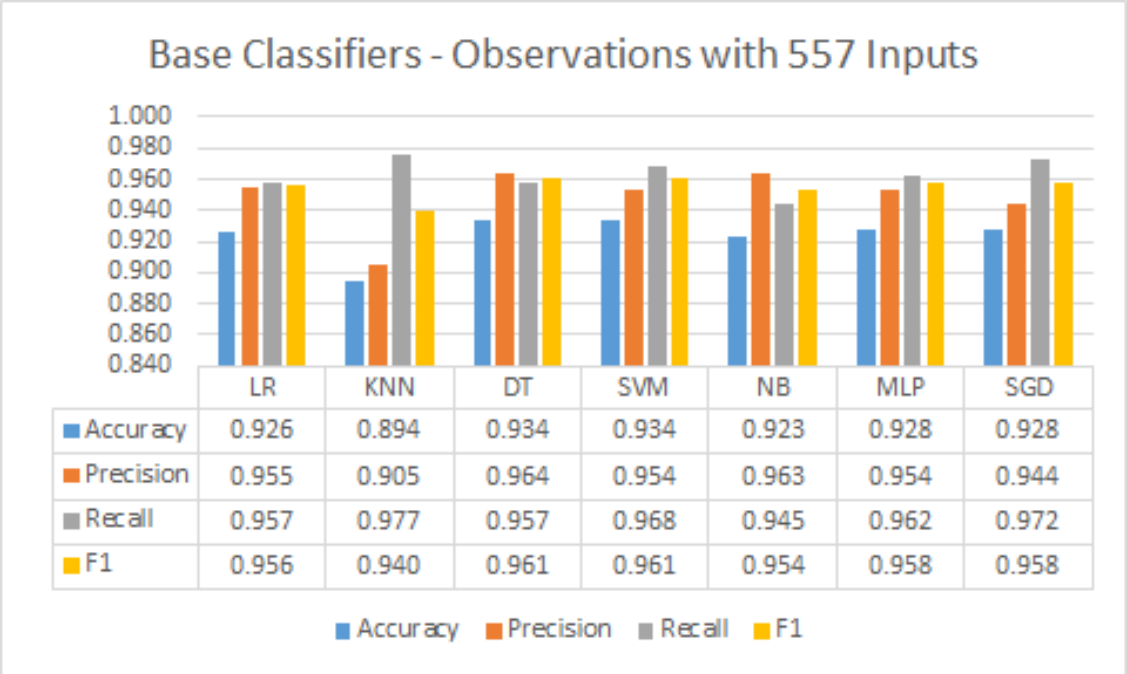
3.6. *Evaluate Base Classifiers*

A total of 7 classifier models, namely DT, Logistic Regression (LR), SVM, MLP, SGD, KNN and NB were used to build models using input dataset. Unseen external data is used to determine the prediction accuracy. Accuracy score is used to determine the bests among base classifiers with Precision, Recall and F1 score supplementing whythe model should be used.

For evaluation of the model, we ran four external dataset to predict the students' performance, each with 557, 349 and 26 observations respectively.

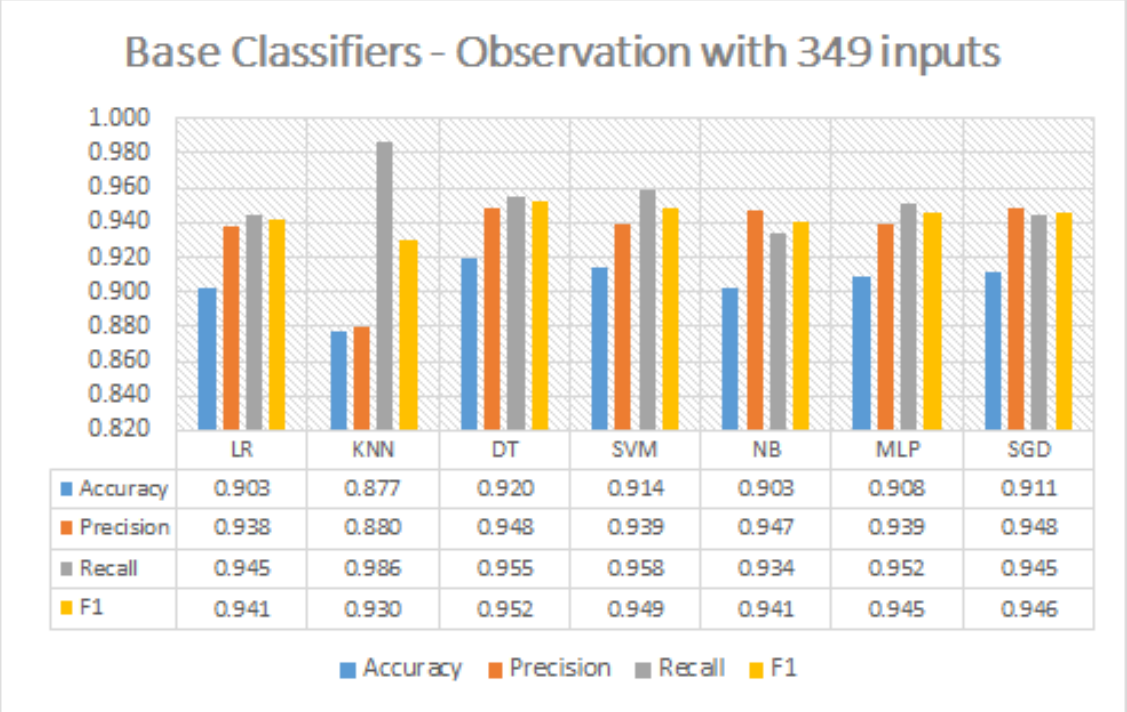
From the 3 different input observation, it appears that Decision Tree Classifier predicts better at an accuracy of 93.4%. Though SVM also predicts 93.4%, Decision Tree is better due to precision, recall and F1 score.

Fig 37, Fig 38 and Fig 39 outlines the observations for each input set.



	LR	KNN	DT	SVM	NB	MLP	SGD
Accuracy	0.926	0.894	0.934	0.934	0.923	0.928	0.928
Precision	0.955	0.905	0.964	0.954	0.963	0.954	0.944
Recall	0.957	0.977	0.957	0.968	0.945	0.962	0.972
F1	0.956	0.940	0.961	0.961	0.954	0.958	0.958

Fig. 37. Observation with 557 inputs



	LR	KNN	DT	SVM	NB	MLP	SGD
Accuracy	0.903	0.877	0.920	0.914	0.903	0.908	0.911
Precision	0.938	0.880	0.948	0.939	0.947	0.939	0.948
Recall	0.945	0.986	0.955	0.958	0.934	0.952	0.945
F1	0.941	0.930	0.952	0.949	0.941	0.945	0.946

Fig. 38. Observation with 349 inputs

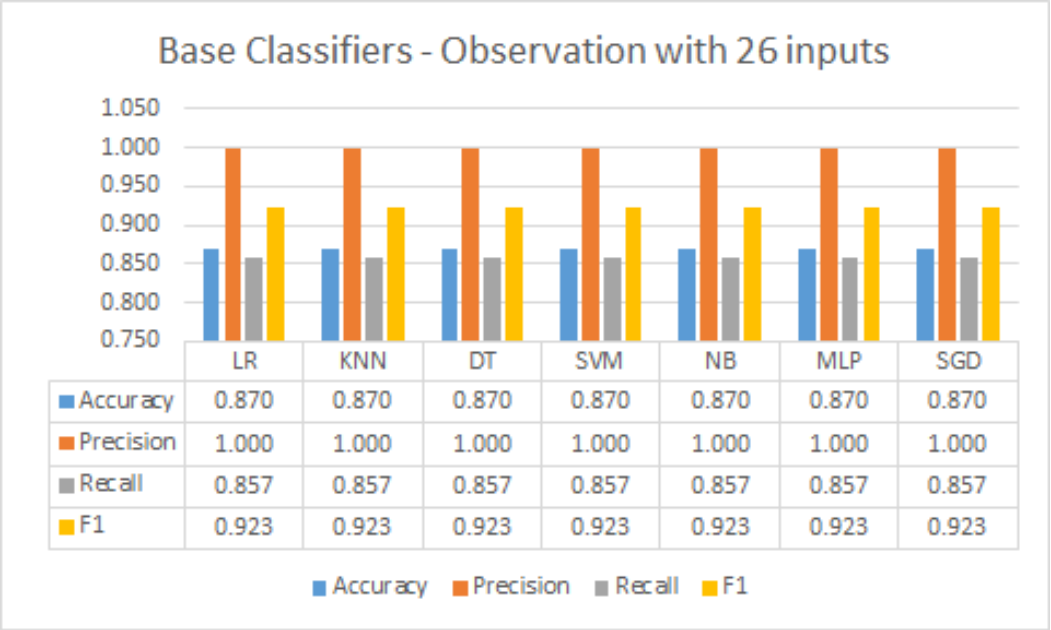


Fig. 39. Observation with 26 inputs

3.7. *Ensemble Model Result*

A total of 7 ensemble models, namely RF, Extra Tree Classifier (EXT), Gradient Boost (GBM), LightGBM (LGBM), AdaBoost (AB), CatBoost (CB), and XGBoost) were used to build models using input dataset. Unseen external data is used to determine the prediction accuracy. Accuracy score is used to determine the bests among base classifiers with Precision, Recall and F1 score supplementing why the model should be used.

For evaluation of the model, we ran four external dataset to predict the students' performance, each with 557, 349 and 26 observations respectively.

From the 3 different input observation, it appears that Extra Tree Classifier predicts better at an accuracy of 94.3% with 557 input and 93.1% with 349 inputs for prediction. It is also observed that when the input data is very low, all the models predicts the same.

Fig 40, Fig 41 and Fig 42 outlines the observations for each input set.

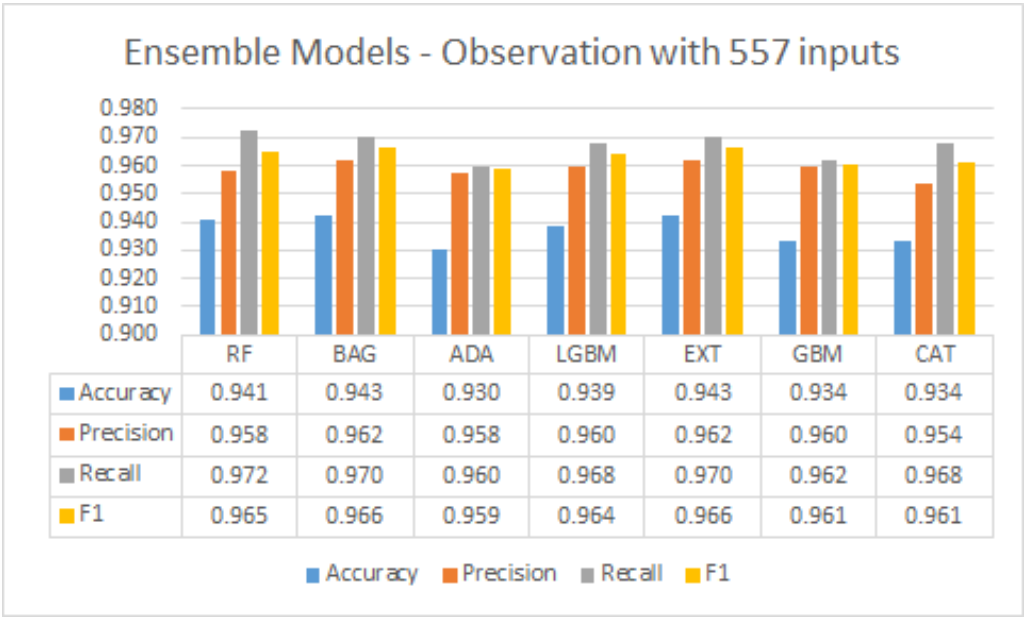


Fig. 40. Ensemble Models with 557 inputs

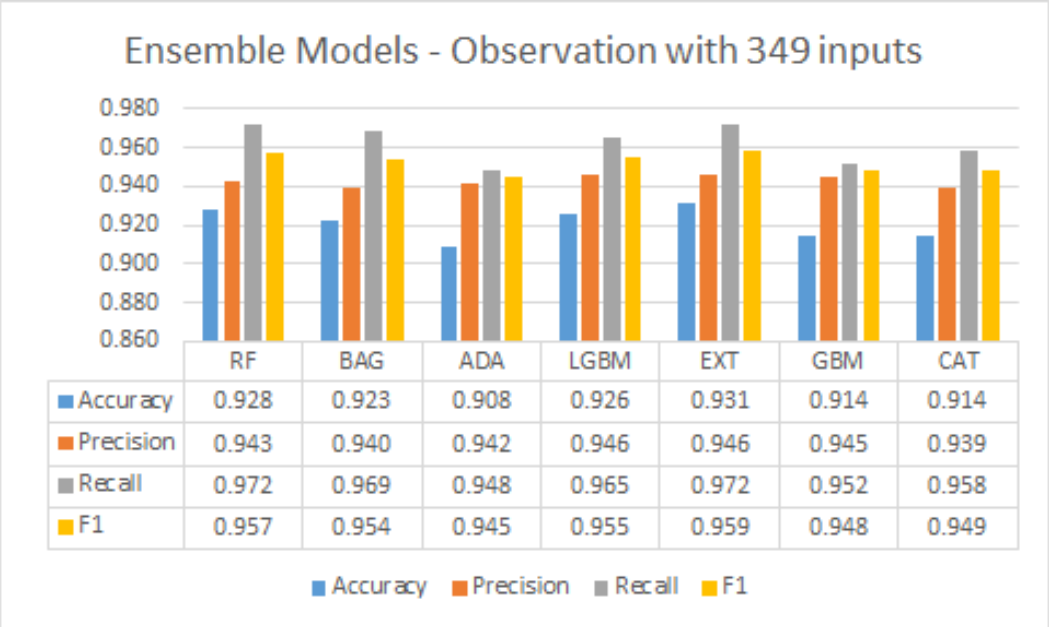


Fig. 41. Ensemble Models with 349 inputs

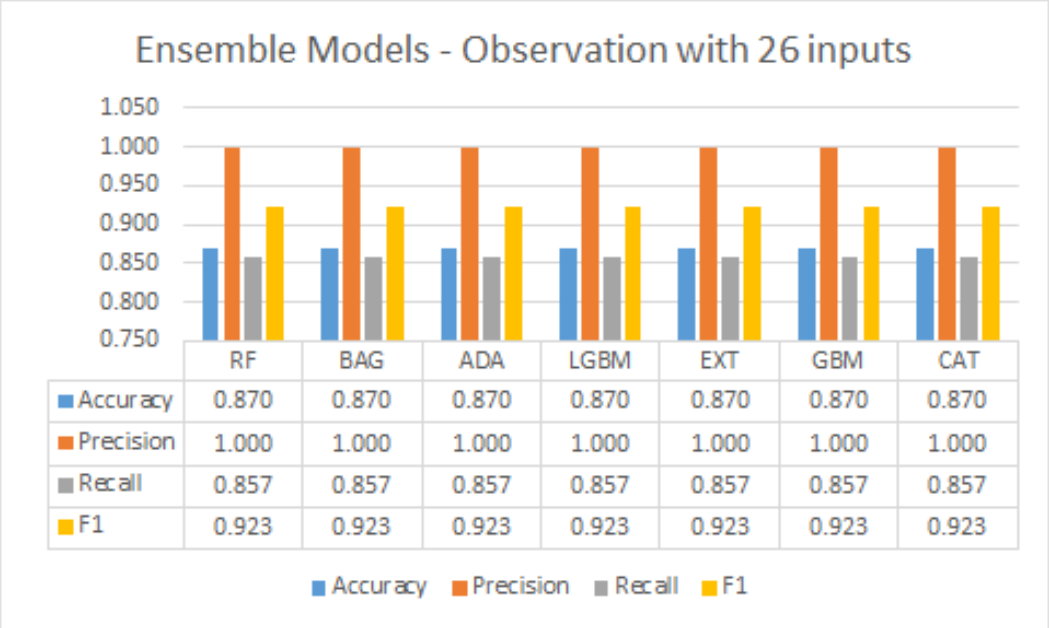


Fig. 42. Ensemble Models with 26 inputs

3.8. Convolutional Neural Network (CNN)

In this deep-learning strategy, we will be using Conv1D of Keras/TensorFlow library. We'll fit the model with train data then will check the training accuracy. The accuracy is 84.2% only, which means CNN model prediction is lower than the classic machine learning models discussed above.

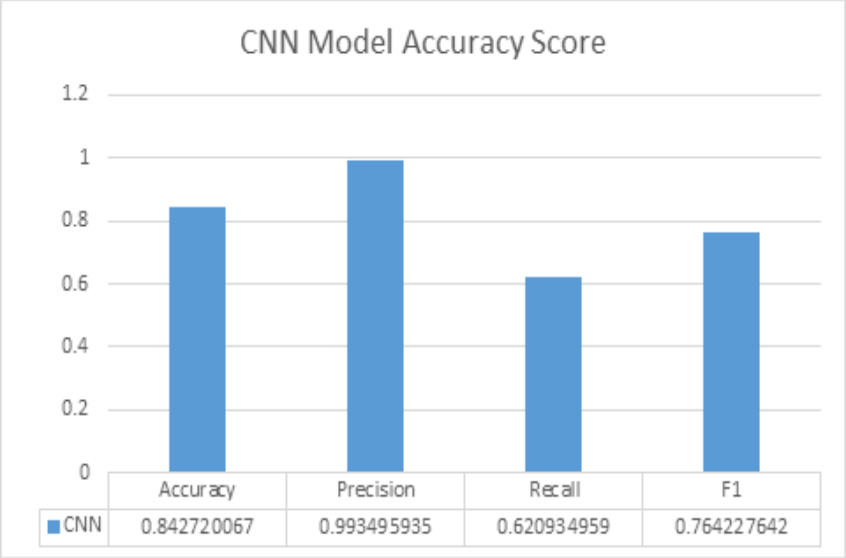


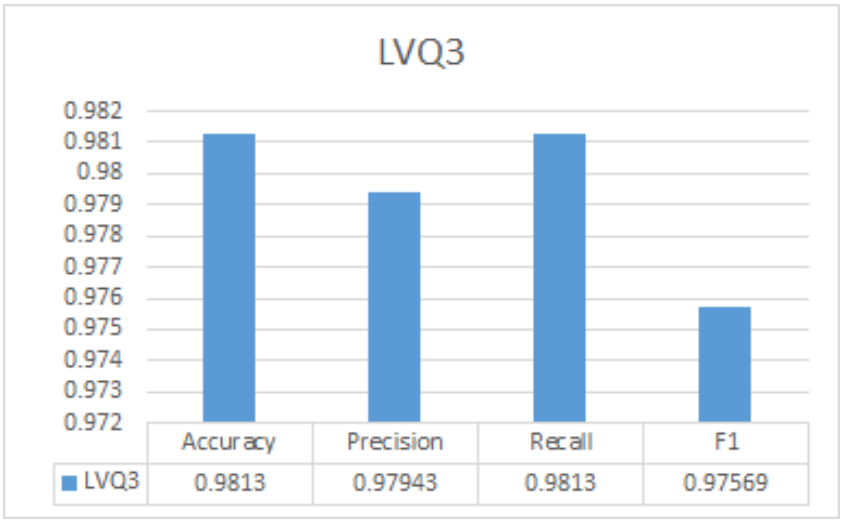
Fig. 43. CNN Model Accuracy

CNN approach produces an accuracy of 84.27%, which is far less than the conventional approach of using baseand ensemble classifiers.

3.9. CNN + LVQ3

Now that we have the result from binary classification models and CNN approach, let us see how LVQ can give much better prediction. In this approach, we use the output model of CNN as input to LVQ algorithm to predict the result using the test dataset.

Fig. 44. CNN + LVQ3 Model Accuracy



From the above graph, it can be concluded that LVQ3 algorithm can produce better prediction result when compared to other out of the box models. Here by making the output model of CNN as input to the LVQ3 algorithmand then predicting the result using the test dataset produced accuracy

score of 98.13% which is an excellent prediction.
Therefore, we can strongly recommend that a hybrid approach as the one explained above can be used to predict the scholastic performance.

4. DISCUSSION

The goal of this discussion is to evaluate the possibilities and to make educated assumptions regarding the presentation of information mining techniques that might be employed in such a situation. LVQ and CNN were coupled on the Student Academic dataset in a sequential manner before being combined. Following the proposed strategy, an approach is adopted in order to increase the prospective qualities for forecasting the introduction of understudies in academics as well as the chance at a specific place.

In light of the exploratory data, it can be inferred that when compared to the individual LVQ classifier, the combination proposed computation, Linear Vector Quantization + CNN (or Hybrid LVQ), is considered appropriate for expectation. This hybrid HLVQ's forecast delivers an 98.13% accuracy, which is significantly higher than the accuracy provided by other methods. We have obtained precision=97.9%, recall=98.1%, accuracy=98%, and F1 score is 97.5% with this hybrid algorithm. The limitation of this work is we can't predict the student placement value for higher education because it is not necessary that all the students further choose higher education they might choose some technical courses or directly choose the option.

Now let us discuss and finalize the best model for students' performance prediction using the top three models from base and ensemble models along with stacking approach from hybrid model. From the above tables, the following were considered for final outcome.

TABLE VII. FINAL COMPARISON

Model Name	**Type**	**Accuracy**
Decision Tree	Base Classifier	93.4%
Extra TreeClassifier	EnsembleClassifier	94.3%
CNN	Deep Learning	84.27%
CNN + LVQ3	**Novelty Hybrid**	**98.13%**

Limitations:

1. The sample dataset used is commonly available for any research and includes only the features more applicable for students in Portugal, though it can be used worldwide. Further study is required to see how these features can relate to schools and colleges in other countries where such research is in the emerging stage.
2. A small dataset of 8000 observations has been used for this experimental analysis. It has been inferred fromthis analysis that changes in the volume of observations leads to different ML model for the most accurate prediction of academic performance of students.

5. CONCLUSION

Over several years various researches has been done in education sector to identify student success as early as possible in their academic career. It is very essential and crucial that expeditious prediction of students' scholastic performance can aid educational institutions to develop a strategy and plan to implement necessary policies to succeed. It has been proved in this research that using Machine Learning Techniques, it is possible to handle even huge datasets to precisely and accurately predict scholastic performance of students.

Though it is widely observed that the same machine learning techinque give different accuracy levels for various researchers, it is has been proved that data features play an even more critical role in determining the level of the final accuracy. It can be concluded that samemachine learning model gives different accuracy score because of using varying dataset with different attributes.

Overall, this research project after various trial and error experiments has provided support the world of Educational Data Mining that data when used with Machine Learning Technology can help identify students' whoneed timely intervention in order to ensure scholastic achievement and success.

6. FUTURE WORK

This research paper has provided a simulation using various ML representations that are commonly used for forecasting scholastic performance of students. In this initial study, out of the various methods used to determine the student performance, the **LVQ + CNN (hybrid LVQ)** proved

to be best model for prediction with an accuracy score of **98.13%.**

In this research, 8000 observations were used for analysis. The same approach can also be utilized in order to forecast the grades of students, in addition to PASS or FAIL criteria, with slight modifications to the data set.

Using this research paper outcome, future work can include the following:

1. Provide a visual interface so that education institutions can easily enter input data to see the outcomeinstead of using traditional model of spreadsheet or statistical analysis.
2. It is possible to visualize the impact using only one feature and for a particular student

References

[1] Shukla M, Malviya AK: Modified classification and prediction model for improving the accuracy of student placement prediction. In: International conference on advanced computing and software engineering, pp 483–487 (2019).

[2] Dutt A, Ismail MA, Herawan T: A systematic review on educational data mining. IEEE Access 5:15991–16005 (2017).

[3] Slater S, Joksimović S, Kovanovic V, Baker RS, Gasevic D: Tools for educational data mining: a review. J Educ Behav Stat 42(1): 85– 106, (2017).

[4] Meier, Y., Xu, J., Atan, O. And Van Der Schaar, M.: Predicting Grades. IEEE Transactions On Signal Processing, 64(4), Pp. 959-972, (2016).

[5] Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. And Abreu, R.: A Comparative Study Of Classification And Regression Algorithms For Modelling Students' Academic Performance. International Educational Data Mining Society (2015).

[6] Veeramuthu, P., Periyasamy, R. And Sugasini, V.: Analysis of Student Result Using Clustering Techniques (2014).

[7] Huebner, R.A.: A Survey Of Educational Data-Mining Research. Research In Higher Education Journal, 19 (2013).

[8] Vahdat, M., Ghio, A., Oneto, L., Anguita, D., Funk, M. And Rauterberg, M.: Advances In Learning Analytics And Educational Data Mining. Proc. Of ESANN2015, Pp. 297- 306 (2015).

[9] Papamitsiou, Z.K. And Economides, A.A.: Learning Analytics and Educational Data Mining In Practice: A Systematic Literature Review Of Empirical Evidence. Educational Technology & Society, 17(4), Pp. 49-64 (2014).

[10] Natek, S. And Zwilling, M.: Student Data Mining Solution–Knowledge Management System Related To Higher Education Institutions. Expert Systems With Applications, 41(14), Pp. 6400-6407 (2014).

[11] Shahiri, Amirah & Husain, Wahidah & Abdul Rashid, Nur'Aini: A Review on Predicting Student's Performance Using Data Mining Techniques. Procedia Computer Science. 72. 414-422. 10.1016/J.Procs.2015.12.157 (2015).

[12] Chaudhari KP, Sharma RA, Jha SS, Bari RJ: Student performance prediction system using data mining approach. Int J Adv Res Comput Commun Eng 6(3):833–839 (2017).

[13] Karbhari N, Deshmukh A, Shinde VD: Recommendation system using content filtering: a case study for college campus placement. In: IEEE International conference on energy, communication, data analytics and soft computing (ICECDS), pp 963–965 (2017).

[14] Karnad A, Yadappanavar S, Hiremath PS: Evaluation and validation of problem solving and thinking skills based on student academic performance. In: IEEE International Conference on recent trends in electronics, information and communication technology (RTEICT), pp 642–646 (2017).

[15] Ghassen Ben Brahim: Predicting Student Performance from Online Engagement Activities Using Novel Statistical Features (2022).

[16] Ahmed AbdElrahman, Taysir Hassan A Soliman, Ahmed I. Taloba, Mohammed F. Farghally: A Predictive Model for Student Performance in Classrooms Using Student Interactions With an eTextbook (2022).

[17] V. Vijayalakshmi, K. Venkatachalapathy: Comparison of Predicting Student's Performance using Machine Learning Algorithms (2019).

[18] Muhammad Sudais, Danish Asad: Student's Academic Performance Prediction – A Review (2022).

[19] Ansar Siddique, Asiya Jan, Fiaz Majeed, Adel Ibrahim Qahmash, Noorulhasan Naveed Quadri and Mohammad Osman Abdul Wahab: Predicting Academic Performance Using an Efficient Model Based on Fusion of Classifiers Prediction Using Machine Learning Techniques (2021).

[20] Kiran Fahd, Shah Jahan Miah, Khandakar Ahmed: Predicting student performance in a blended learning environment using learning management system interaction data (2021).

[21] Aaditya Bhusal: Predicting Student's Performance Through Data Mining (2021).

[22] Tuti Purwoningsih, Harry B. Santoso, Kristanti A. Puspitasari, Zainal A. Hasibuan: Early Prediction of Students' Academic Achievement: Categorical Data from Fully Online Learning on Machine Learning Classification Algorithms (2021).

[23] Bhavesh Patel: Performance Based Machine Learning Model to Enhance Performance of Students (2021).

[24] Lonia Masangu, Ashwini Jadhav, Ritesh Ajoodha: Predicting Student Academic Performance Using Data Mining Techniques (2021)

[25] Shaikh Rezwan Rahman, Md.Asfiul Islam, PritidhritaPaul Akash, Masuma Parvin, Nazmun Nessa Moon, FernazNarin Nur: Effects of co-curricular activities on student's academic performance by machine learning (2021).

[26] Mohammad Noor Injadat. Abdallah Moubayed, Ali BouNassif, Abdallah Shami: Systematic ensemble model selection approach for educational data mining (2020).

[27] Durgesh Ugale, Jeet Pawar, Sachin Yadav, Dr. Chandrashekhar Raut: Student Performance Prediction Using Data Mining Techniques (2020).

[28] Randhir Singh, Saurabh Pal: Machine Learning Algorithms and Ensemble Technique to Improve Prediction of Students Performance (2020).

[29] Hussein Altabrawee, Osama Abdul Jaleel Ali, Samir Qaisar Ajmi: Predicting Students' Performance Using Machine Learning Techniques (2019).

[30] Samuel-Soma M Ajibade, Nor Bahiah Binti Ahmad, Siti Mariyam Shamsuddin: Educational Data Mining: Enhancement of Student Performance model using Ensemble Methods (2019).

[31] Akm Shahariar Azad Rabby, Syed Akhter Hossain: Machine Learning Algorithm for Student's Performance Prediction (2019).

[32] Fergie Joanda Kaunang, Reymon Rotikan: Students' Academic Performance Prediction using Data Mining (2018).

[33] Ahmad F, Ismail N, Aziz A: The Prediction of Students' Academic Performance Using Classification Data Mining Techniques (2015).

[34] Parneet Kaur, Manpreet Singh, Gurpreet Singh Josan: Classification and prediction based data mining algorithms to predict slow learners in education sector (2015).