Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 7
June 2023

Analysis and Comparison of Student Performanceusing Machine Language Algorithms

Praveena Chakrapani
School of Computing Sciences Hindustan Institute of Technology and Science
Chennai, India
rp.18772016@student.hindustanuniv.ac.in

A Anthonisan
School of Computing Sciences Hindustan Institute of Technology and Science
Chennai, India
dyregistrar@hindustanuniv.ac.in

ABSTRACT-In recent times, forecasting students' Academic Performance has gained substantial traction and is the dynamic challenges of academic institutions. Educational Data Mining (EDM) with advanced techniques and methods plays a significant role in addressing students' academic performance. Although research has been conducted at University and College levels, only limited research has been conducted at the School level, with respect to predicting academic performance. Identifying students' academic performances at an early stage is crucial for educational institutions and parents in order to take proactive decisions concerning a student's future. The goal is to determine the factors that affect a student's scholastic performance and enable parents and educational institutions to accurately predict a student's academic performance and channelize the student's capabilities in the right direction, at the right time. This research makes use of Supervised Machine Learning approaches to analyze, filter and determine students at risk and suggest alternatives to improve their performance. In order to achieve the desired objective, this paper analyses Classification algorithms for EDM in depth, and identifies the right attributes and the most suitable Machine Learning tool which can accurately predict the scholastic performance of school students and ensure academic achievement. After conducting a comparative study of the results from various Machine Learning techniques using student data, this research shows that the Light Gradient Boost Method is the best in predicting the scholastic performance of school students.

1. INTRODUCTION

One of the biggest and growing challenges in Education is identifying students' scholastic performance and how to provide the required assistance to enable a student to succeed in it. This is widespread - starting from the foundational elementary grades to graduate levels. Although an academician's final grade, grade point average (GPA) is the determining factor, there are various other factors which influence the scholastic performance of a student. Analyzing the wealth of data available about students is very challenging due to the fact that the environment they operate upon is very dynamic and complex.

Detecting and intervening as early as possible during the students' academic period will help educational institutions and parents to implement the required interventions to improve a school students' academic performance. Through the advancement of technology, it is possible to systematically analyze their performance, identify their strengths, weaknesses and areas of improvement using which a strategy can be formulated by educational institutions and parents to guide students in focusing on their inherent strengths. For example, intelligent tutoring systems can be provided, student counselling and monitoring can be established, a policy can be made for the proactively help students.

Educational Data Mining (EDM) enables analysing large volumes of scholastic data in terms of density 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 available for Scholastic Academic Performance Prediction (SAPP).

The scope is to apply machine learning tools to infer the performance of school students using data that could influence a student's scholastic performance using which educational institutions can formulate a strategy to provide the required support to students. This research makes use of data from 557 students of a Portuguese school for prediction. The dataset contains 33 features. Supervised Machine Learning models including explicit base models, ensemble models and hybrid models have been used.This project paper is categorised into various sections, namely - Section II presents a Literature Review of Related Work in the same space; Section III discusses the Materials and Methods used in predicting students' performance; Section IV discusses the Results of the performance analysis; Section V presents the Discussions of this research's outcome and finally Section VI outlines the Conclusions of this research and scope for future research.

2. RELATED WORK

As a consequence of advancements in Science and Technology, there is immense research in the field of EDM, particularly, with respect to the Scholastic Academic Performance of students. These researches explore suitable student attributes (features) that can be used in combination with associated ML algorithms in order to devise models for accurate prediction of Academic performance.

2.1. Online Classes

During pandemics, online or virtual classes had become popular. However, online classes also have challenges that may impact student performance, such as, the lack of direct interaction with instructors and classmates, discipline, technical difficulties, and so on. Analysing student's performance under this study model was done by various researcher using different criteria. A research conducted by Mohammad Noor Injadat et al [17] to predict students who may need help in an e-learning environment, uses two different datasets with slight change in the order of the attributes. ML models - SVM, RF, NB, MLP, KNN, and LR were used for prediction. In the model proposed by Ahmed Abdul Rahman et al [3], the considerations were focused on predictions of good performance, bad performance and final grade. Using the events in the OpenDSA infrastructure, this study infers that the RF Classifier model is best suited with a resulting accuracy of 91.7%. Ghassen Ben Brahim [1], predicted students' performance during online classes through the DEEDS dataset's interaction logs using the RF Classifier whose resulting accuracy and F1-score were 97% respectively.

2.2. Hybrid Study Model

Hybrid study models combine both traditional classroom-based learning and online learning

and therefore outweighs the challenges in both methods. Factors such as the content quality, teaching effectiveness, student individuality and so on need to be considered. A research by Ansar Siddique et al [6] focused on the factors influencing student performance at the school secondary level of education. The study selects 16 key attributes and predicts performance through a 7-category grading system. ML techniques MLP, PART and J48 (DT) were used for the prediction.

2.3. Learning Management System (LMS)

With improved Access to Learning Resources, an LMS can provide students a diverse variety of learning resources, promoting self-learning and thereby improved understanding, clarity and retention as an essential outcome of a personalized learning experience. Aaditya Bhusal's [9] research infers that LMS is best suited for prediction of final grades. Kiran Fahd et al [7], in their study of LMS infer that the RF model with an accuracy of 85% is best suited for determining the attrition rate of students. Tuti Purwoningsih et al [10] used the models – RF, DT and GBT with the LMS demonstrating an accuracy of 85.03%. Sellappan Palaniappan et al [28] inferred that the DT algorithm is best suited for predicting performance using the WEKA data mining tool.

2.4. Co-curricular Activities Related Research

Research on the type of influence of Co-curricular activities on a student's academic performance is researched. Most are positive. The impact on a student's academic performance due to participation in co-curricular activities was analyzed by Shaikh Rezwan Rahman et al [14] using ML models - RF, VT, MLP, LR and highlighting the highest accuracy of 99.52% using LR.

2.5. Single Subject Research

The quality of the educational content for a single subject and the effectiveness of the teaching and learning strategies used can significantly impact student performance. In particular, many researches focus on performance in computer science [22, 26, 29, 34] using various machine learning techniques such as, NB, DT, LR, ANN and so on. with prediction rates ranging from 66.9% to 77.04%.In some of these researches [21, 26, 34], data was collected using several methods such as, questionnaires that contain the students' prior academic performance, preferences, cultural background, co-scholastic strengths, family background, among others, to name a few.

2.6. Behavioural Research

Behavioural research enables us to recognize the impact of positive behavioural habits and values on student performance. A student's behaviour and attitude towards education is influenced by both external and internal factors. External factors such as, the neighbourhood he lives in, social relationships, friends, influence a student's behaviour towards academics. Internal factors such as, the domestic environment, the attitude of the family members towards education, the attitude of the family members towards the student, also have a significant influence in determining the habits, values, level of motivation and attitude towards education in general [23]. In another research by Parneet Kaur et al, Multilayer Perceptron (MLP) is used to identify slow learners [35] with 75% accuracy.

2.7. Teaching Method Research

Students are characterized based on their capabilities and academic capabilities as a factor of their strengths and weaknesses. This study groups students into two categories and models them based on their performance as a combination of their intellectual quotient and emotional quotient based on a balance between their skills, emotions, knowledge is achieved [19].

2.8. Demographic Research

Certain Demographics [11, 18], factors have a direct impact on students' academic performance. This is in addition to other co-scholastic factors discussed above. Several researchers have analyzed how demographic factors influence a student's performance, the extent of impact and whether the impact is positive or negative towards academic success. Such research enables educators and parents to provide a more holistic education to the student, that goes beyond disparities.

2.9. Grades and Mark Research

The results of an academic assessment can influence students' academic performance –

either positive or negative. Good grades and positive feedback encourage a student to perform academically better and improve their scholastic performance [2, 20, 31. Consistency in positive feedback indicate clarity and mastery of study material and academic progress [4, 5, 25], whereas, consistency in negative feedback is indicative of a student's weakening grades and the need for intervention and an additional support system [8, 32, 33]. Senior school students are even more impacted as it is a prime determinant for their graduate education. High weightage is given by Universities for a student's grade (GPA) for decisions with respect to admission. Therefore, a high GPA is indicative of the success rate for a student to get into a preferred university and in addition, the course of choice.

2.10. Research on Predicting Pass or Fail

Various internal and external factors influence the academic performance of a student, as discussed above. Success or Failure of a student at the school level is a crucial determinant of the success or failure in his higher education and career [13,24]. Failures at the school level may cause a student to even drop out of school [12].

Archived data on student performance can be an invaluable resource for educators, as it can provide timely insight into patterns of a student's academic achievement and suggest corrective measures. Therefore, predictions on whether a student would pass or fail [24, 27, 30] are effective measures to render corrective action towards student success.

3. METHODS AND MATERIALS

3.1. Proposed Workflow Diagram

The modelled workflow is sequenced in Fig. 1

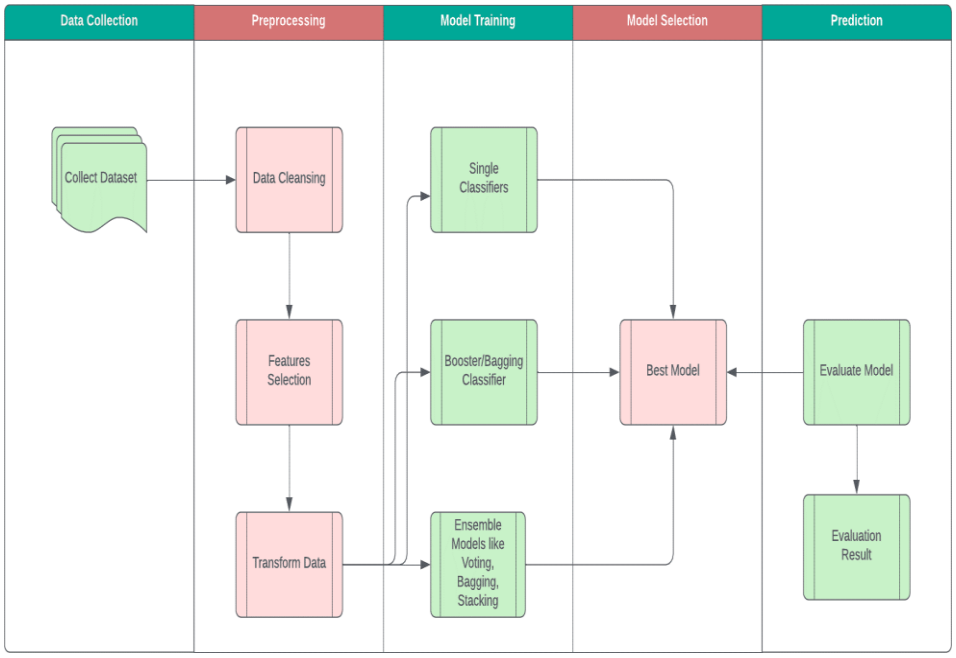


Fig. 1. Sequence Flow Diagram

The process comprises of 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 result is 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 featuresthat 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 resultant machine learning techniques to evaluate the data to make the finalprediction. One of the model will render itself as a suggested model for student's performance prediction.

3.2. Dataset Collection and Identification

The data set is sourced from an anonymous school. A total of 8000 student's records were used for this research.

TABLE I. DATASET SIZE

Data Source	**Attributes**	**Records (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	(<15min., 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 thecourse 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 - veryhigh)
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)

3.3. Data Exploration

To understand the data, the dataset will be imported for an Exploratory Analysis and to have an overview of the distribution of the numerical data. Below is the histogram plotting on numerical data. The Histogram shows how each feature maps against the target value.

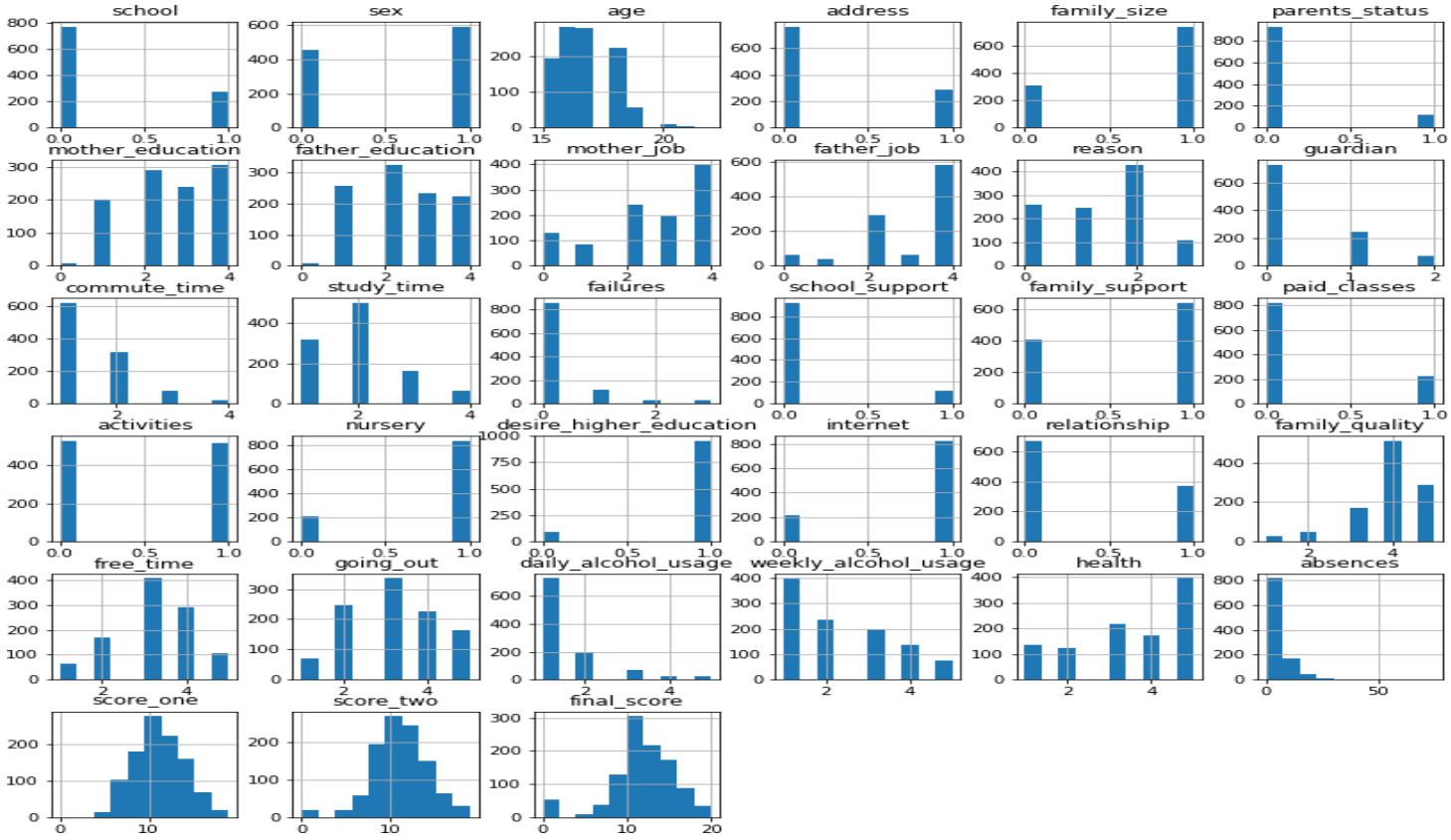


Fig. 2. Histogram of each feature vs final_score (target value)

3.4. Data Cleansing

Data exploration leads to the second stage of Data Cleansing in order to ensure that data integrity is ensured. In this context, data integrity ensures that there are no missing values or duplicates.

This data cleansing step (Fig. 3) makes sure that any row(s) from the original raw dataset having missing value(s) or duplicate entry is removed and only relevant data with all the values is retained for further processing. Rows with missing values have been removed to avoid computational complexity. Using Python code in Jupyter Lab environment, all rows that had empty values in any column have been removed.

	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. 3. First 5 rows of the dataset

3.5. Machine Learning (ML) Algorithms

In this research, the following ML models will be used to select the final, recommended model for prediction.

- Base Classifiers
- Ensemble Methods
- Hybrid Methods (Ensemble + Base)

3.5.1. Base Classifiers

A total of 7 Classifier models, namely:

- K-Nearest Neighbour (KNN)
- Support Vector Machine (SVM),
- Stochastic Gradient Descent (SGD)
- Decision Tree (DT),
- Multi-layer Perceptron (MLP),
- Logistic Regression (LR) and,
- Naïve Bayes (NB)

have been used to build models using the input dataset.

3.5.2. Ensemble Classifiers

Ensemble Learning techniques can be classified as:

- Bagging
- Boosting
- Adaboost
- Gradient Boost
- XGBoost
- Light Gradient Boost
- CatBoost

3.5.3. Hybrid (Heterogeneous) Classifiers

Ensemble Multiple classifier models can be classified into:

- Voting
- Stacking (Stacked Generalization)

4. RESULTS

4.1. Environment

Jupyter Lab and Python Libraries for data mining were the tools used to evaluate the proposed classification models to arrive at a conclusion. A structured sequential approach was implemented to predict 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 models.

4.2. Feature Selection

For feature selection, the Wrappers, viz., Forward Elimination, Backward Elimination, Low Variance Filter, SelectKBest methods were not used 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. Boxplot or bar chart or pie chart is used to understand the influence of features.

After feature selection, eighteen features have been used for training the models for prediction.

Features – **school, guardian, daily alcohol, weekly alcohol and romantic relationship** were not considered forfeature selection. It was inferred that these features 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 featureis shown below to know why they are included in the final feature selection list.

4.2.1.1. Age

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

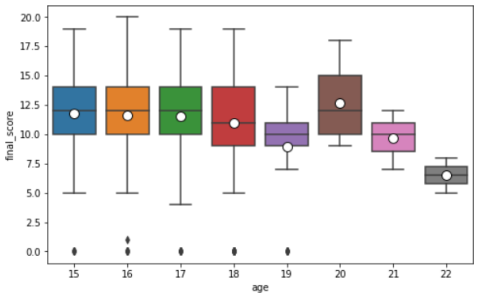


Fig. 6. Age impact on students' performance

4.2.1.2. Address

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

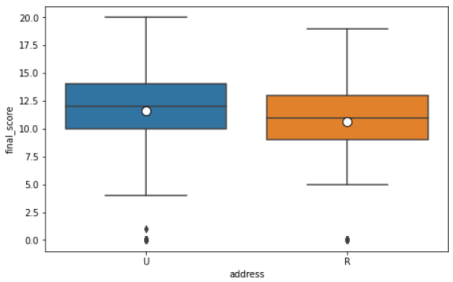


Fig. 7. Urban vs Rural place of stay impact on the student's performance

4.2.1.3. Mother's Education

From the below figure it is clear that Mother's education has an impact on the students' performance.

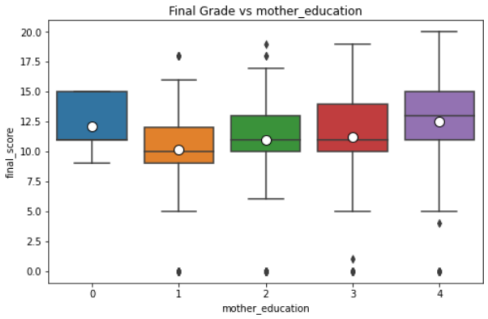


Fig. 8. Mother's Education impact on the students performance

4.2.1.4. Father's Education

From the below figure it is clear that Father's education influences student's academic performance.

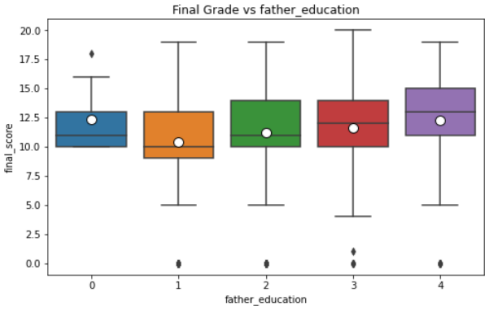


Fig. 9. Impact of father's education

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.

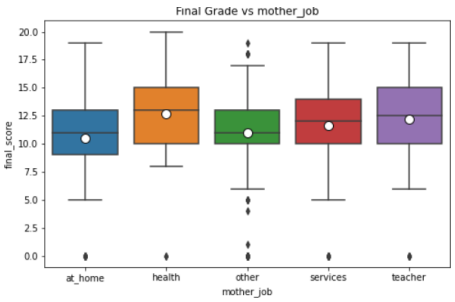


Fig. 10. Mother's job impact on the student performance

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.

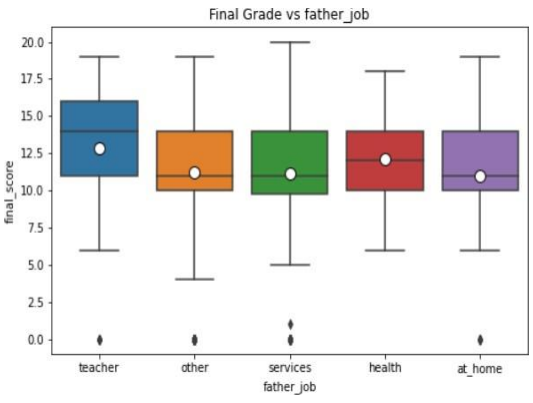


Fig. 11. Father's job impact on the student performance

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.

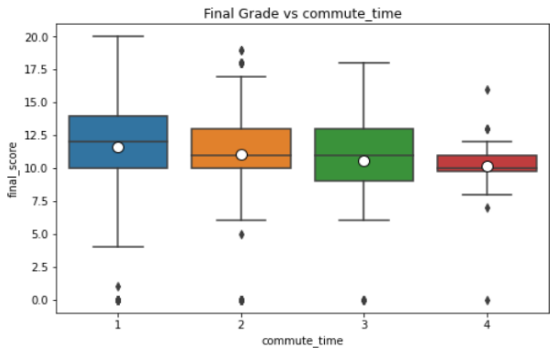


Fig. 12. Commute time impact on the student performance

4.2.1.8. Study Time

Students who spend more time to study, performs better.

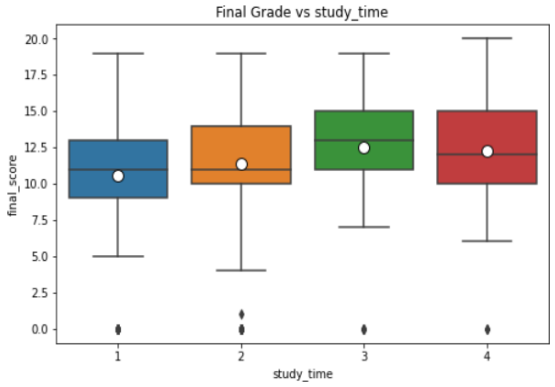


Fig. 13. Study time impact on the student performance

4.2.1.9. Failures

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

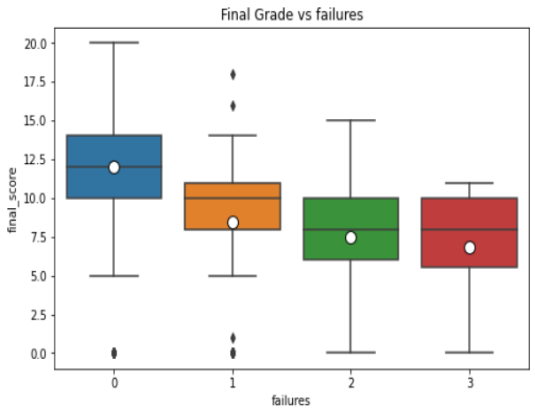


Fig. 14. Past failure impact on the student performance

4.2.1.10. School Support

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

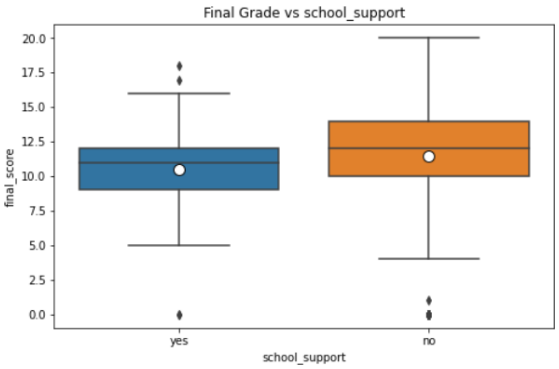


Fig. 15. School Support impact on the student performance

4.2.1.11. Higher Education

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

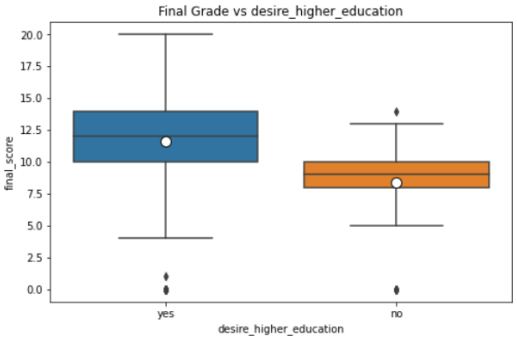


Fig. 16. Higher Education Interest impact on the student performance

4.2.1.12. Internet Impact

Internet access helps students perform better.

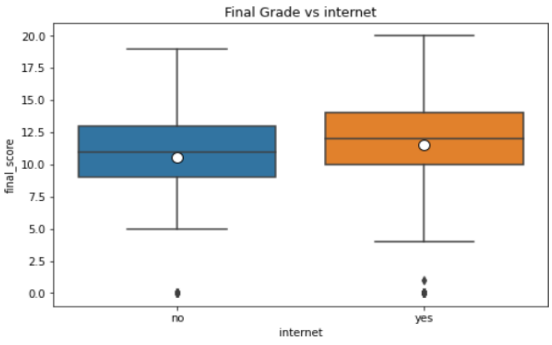


Fig. 17. Internet access impact on the student performance

4.2.1.13. Going Out

Students who go out more tends to score less.

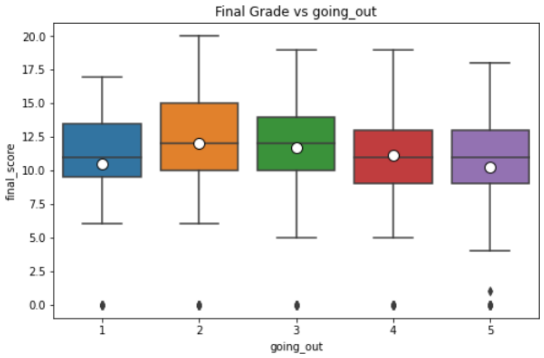


Fig. 18. Going out habit impact on the student performance

4.2.1.14. Student Health

Student whose health is better seems to perform better.

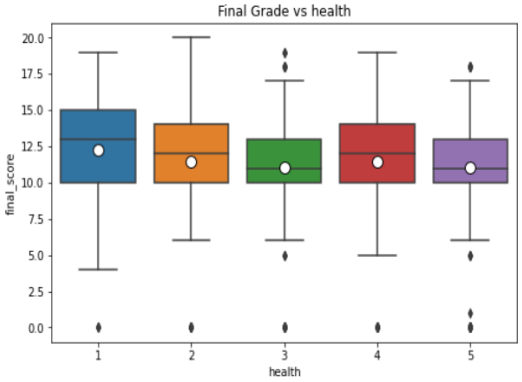


Fig. 19. Student's health impact on the student performance

4.2.1.15. Absentism

Student who has less absences seems to perform better.

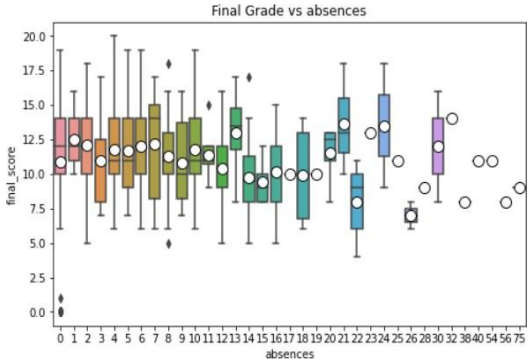


Fig. 20. Absences impact on the student performance

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

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

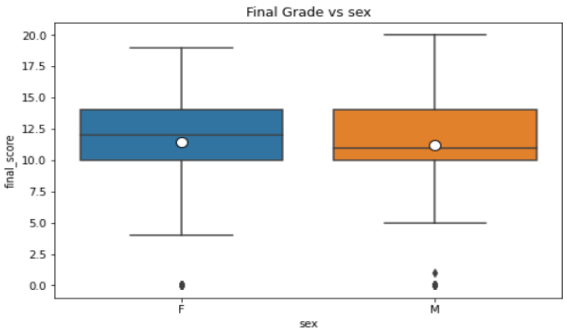


Fig. 21. Gender impact on the performance

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.

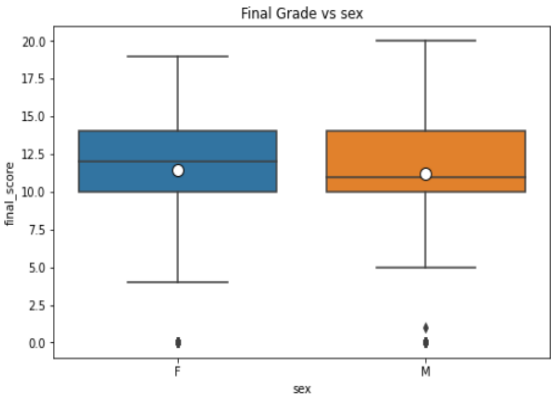


Fig. 22. Family size impact on the performance

4.2.2.3. Parent Cohabitation

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

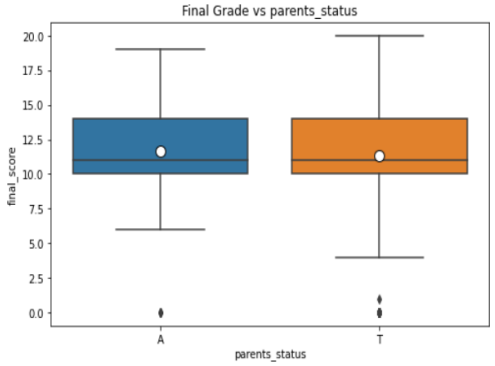


Fig. 23. Parent Cohabitation impact on the performance

4.2.2.4. Reason to join

Student reason to join the school shows no major impact on the performance. This can be excluded.

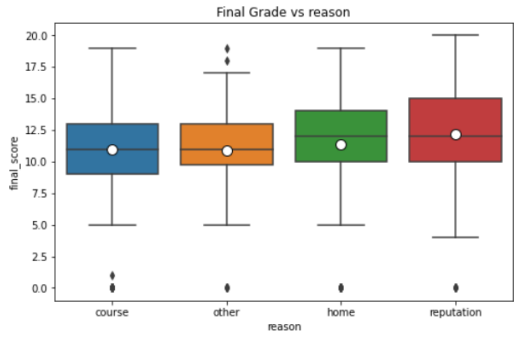


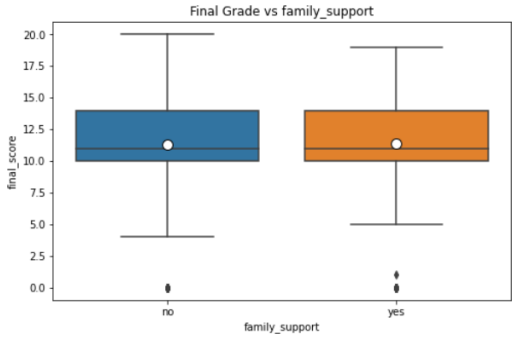
Fig. 24. Reason to join the school impact on the student performance

4.2.2.5. Family Support

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

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.



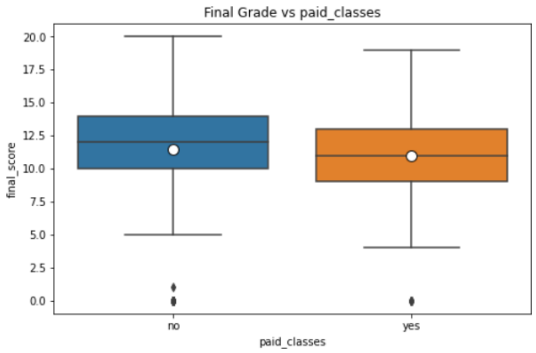


Fig. 26. Extra classes impact on the student performance

4.2.2.7. Other Activities

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

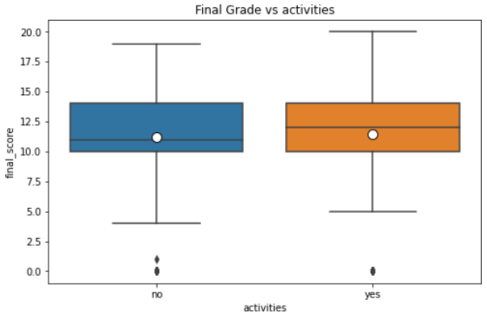
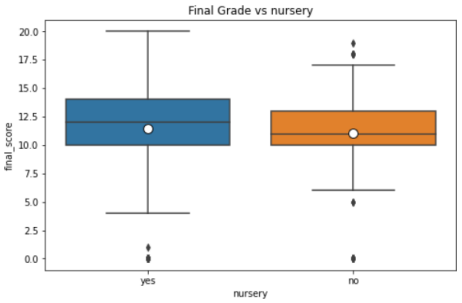


Fig. 27. Other activities impact on the student performance

4.2.2.8. Nursery Study



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

Fig. 28. 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.

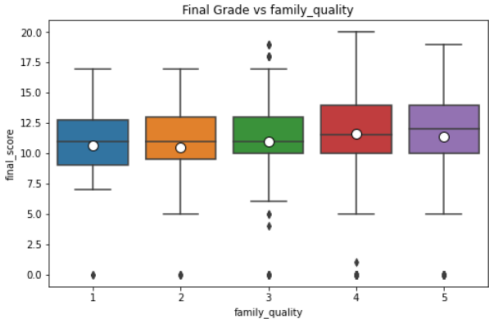


Fig. 29. Family relationship (quality) impact on the student performance

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.

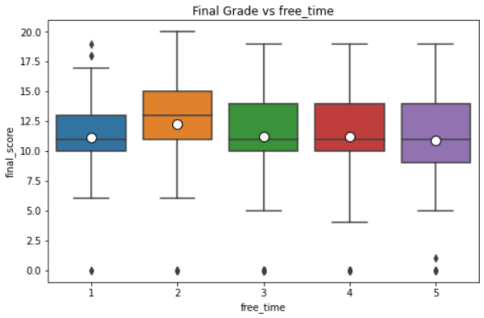


Fig. 25. Family support impact on the student performance

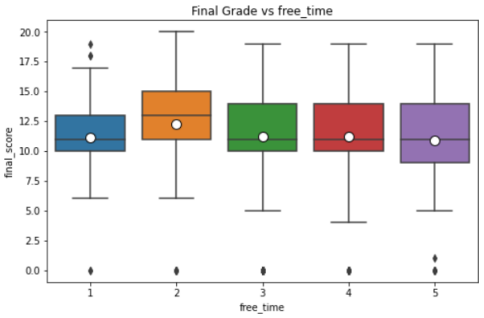
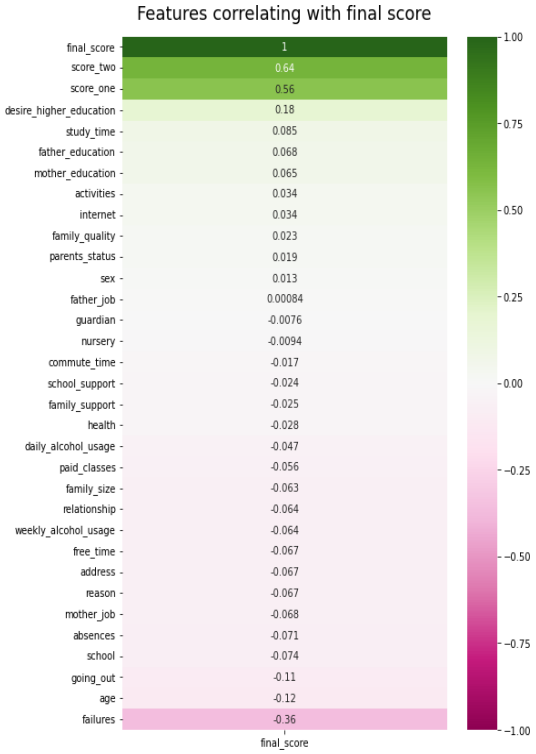


Fig. 30. Free time impact on the student performance

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

Fig. 31. Feature correlation map

4.2.3. Features Correlation Map



The below figure shows how each features

4.3. Data Transformation

This is the last stage of data pre-processing. In this stage, all the non-numerical values are converted to numerical values. For this purpose, the non-numeric data is mapped to its corresponding number. For instance, 'yes' and 'no' values will be mapped to1 and 0 respectively.

The selected features have values that are numerical but the value is spread very wide. Explicit data transformation is done for those features so that they are within 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. 32. 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. 33. 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. 34. 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+	E
16 - 17	5	A	VG

14 – 15	4	B	G
12 – 13	3	C	SA
9 – 11	2	D	SU
0 – 8	1	E	F

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	1 and 18	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 howthey 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

4.4. Model Checking

The K-fold cross-validation validates the model's performance. Data is split into folds (sections) and each is used for validation.

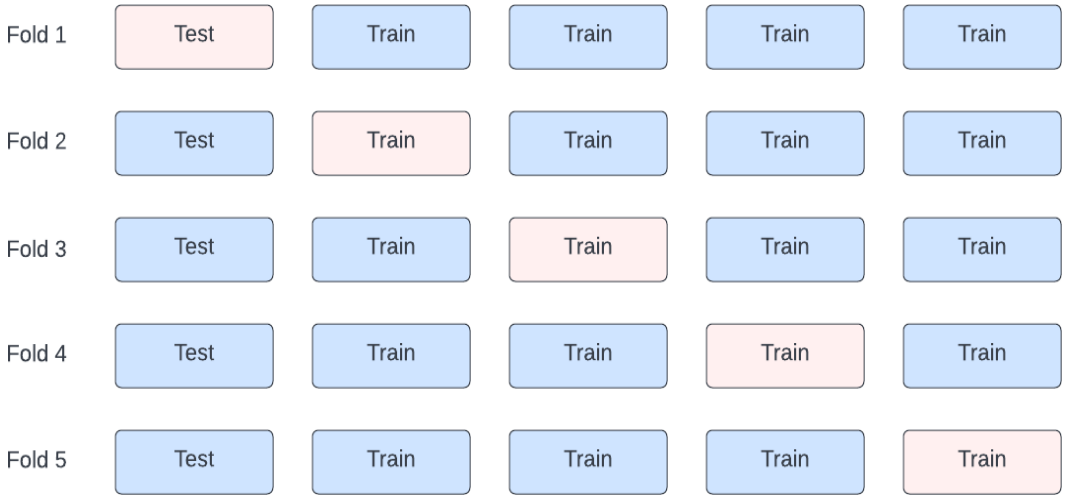


Fig. 35. K-Fold Approach

The model has been trained using a part of data and is tested for its ability to predict using the data it hasn't seen. The 'Train and Test' split approach, is used to predict and evaluate the model's ability. Generally, it is always random selection of x% and, not all data is used for train and test.

In this research, all data has been utilized for accuracy. A 10-fold cross validation has been conducted. This is done to ensure single utilization of data. The K-fold algorithm's cross valiation technique is therefore used for checking a model and not for building a model.

For actual model building we will be using 80:20 train-test split of data. Fig. 37 represents K-fold result.

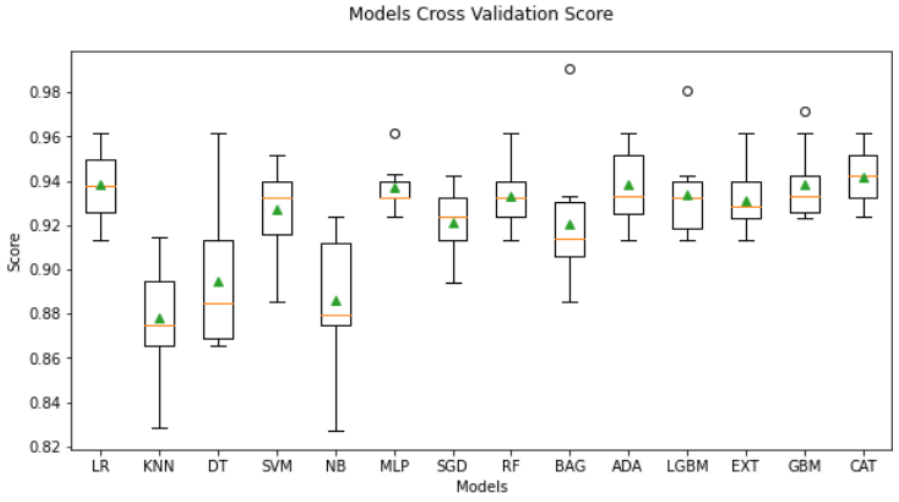


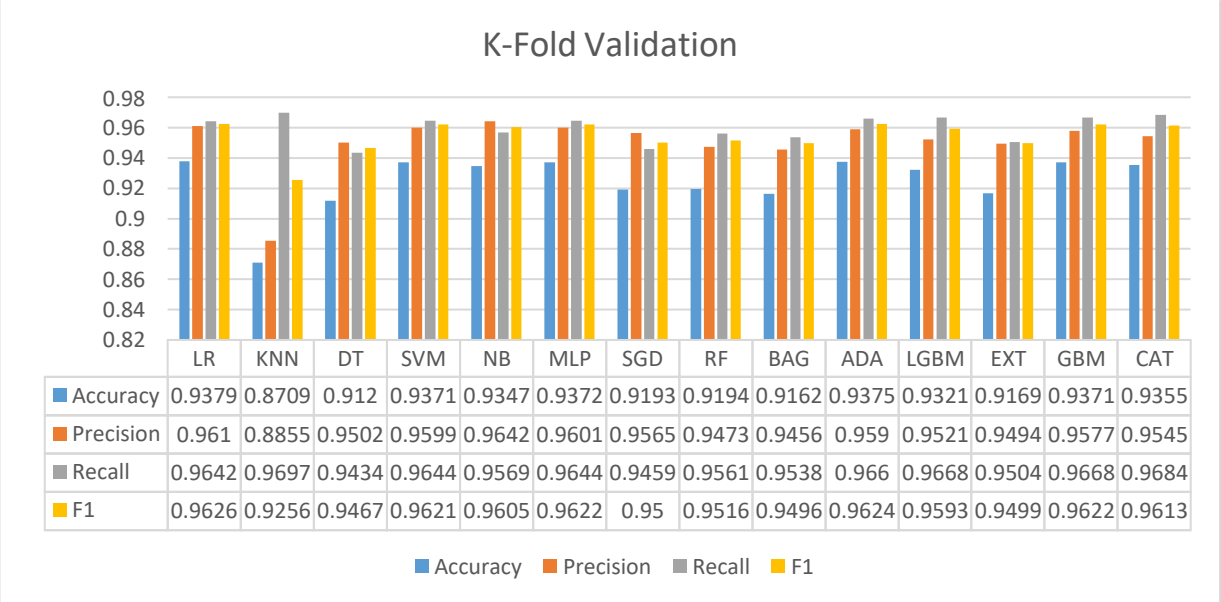
Fig. 36. Model Fitting using K-fold Approach

4.5. Model Fitting

In this research, it is initially determined whether the data set fits the model selected for prediction. This is known as model fitting. All data has been used instead of using only the train and test split of data. The ten K-Fold cross validation technique is used to determine model fitting. 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 fitted in this research. The score could appear less because it is the 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. 37. Model Fitting using K-fold Approach

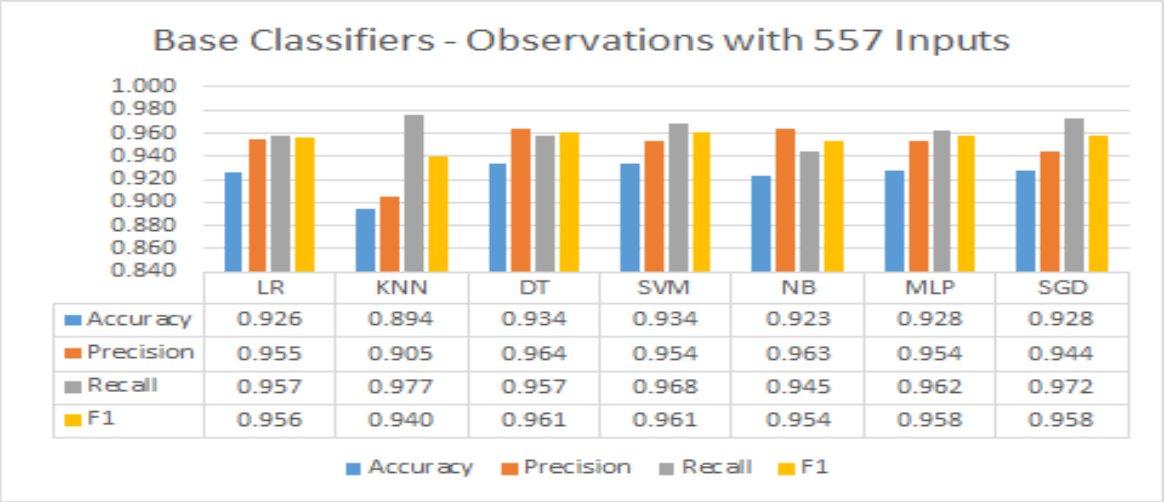
4.6. Evaluate Base Classifiers

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

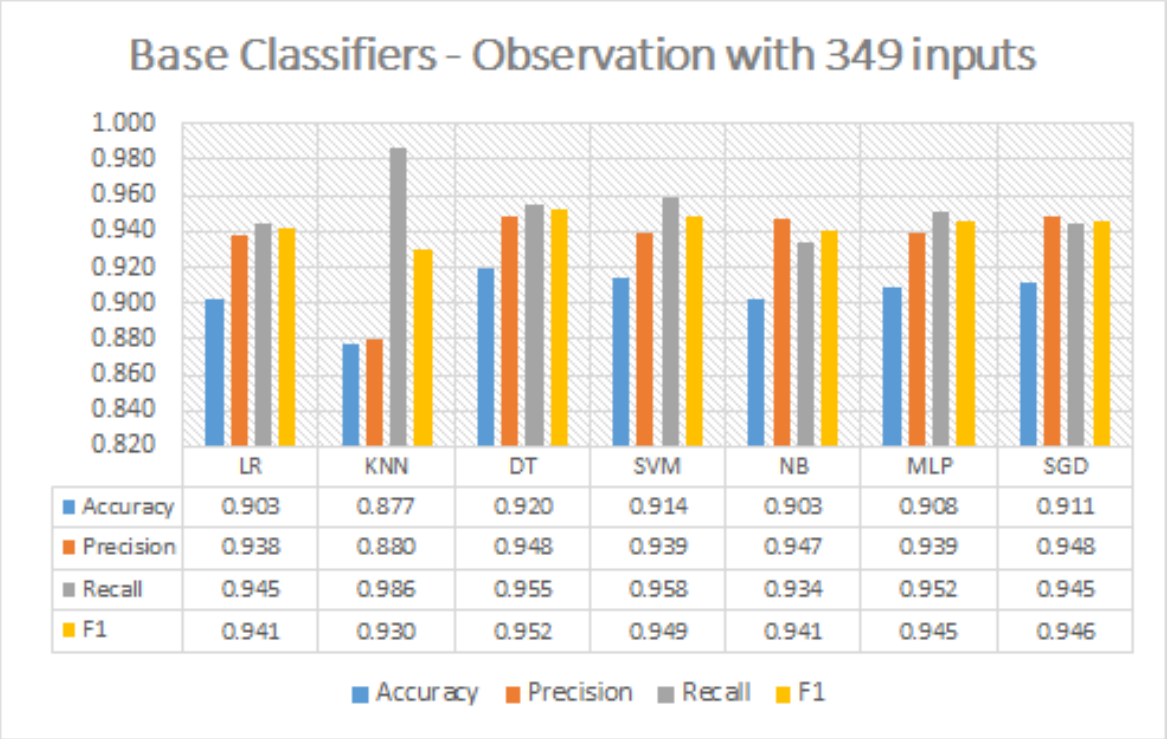
From the 3 different input observation, it appears that Decision Tree Classifier predicts better at an accuracy of 93.4%. Though SVM also predicts 93.4%, Decision Tree is better due to precision, recall and F1 score.

Fig 38, Fig 39 and Fig 40 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. 38. 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. 39. Observation with 349 inputs

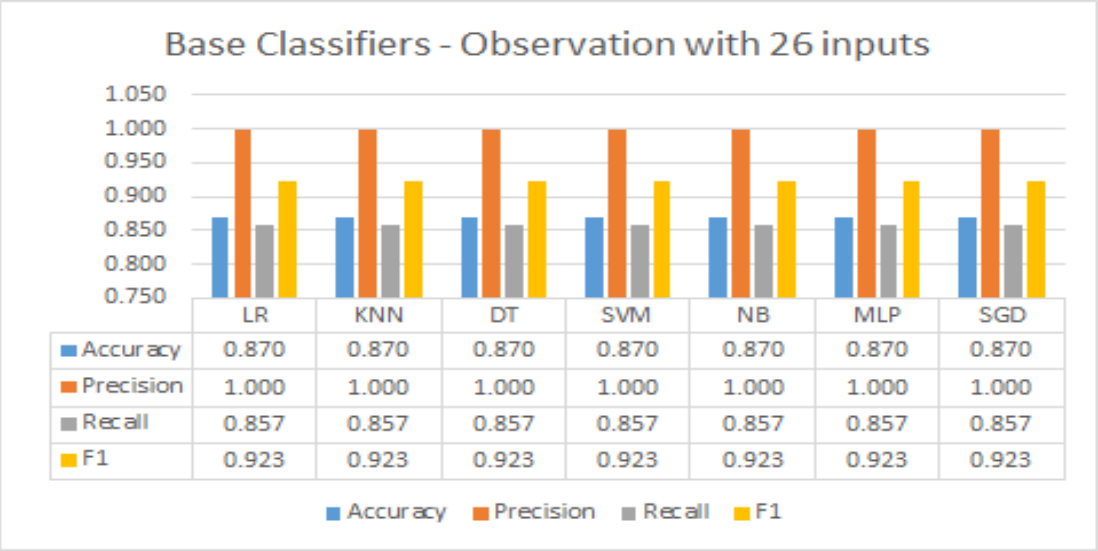


Fig. 40. Observation with 26 inputs

4.7. 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 41, Fig 42 and Fig 43 outlines the observations for each input set.

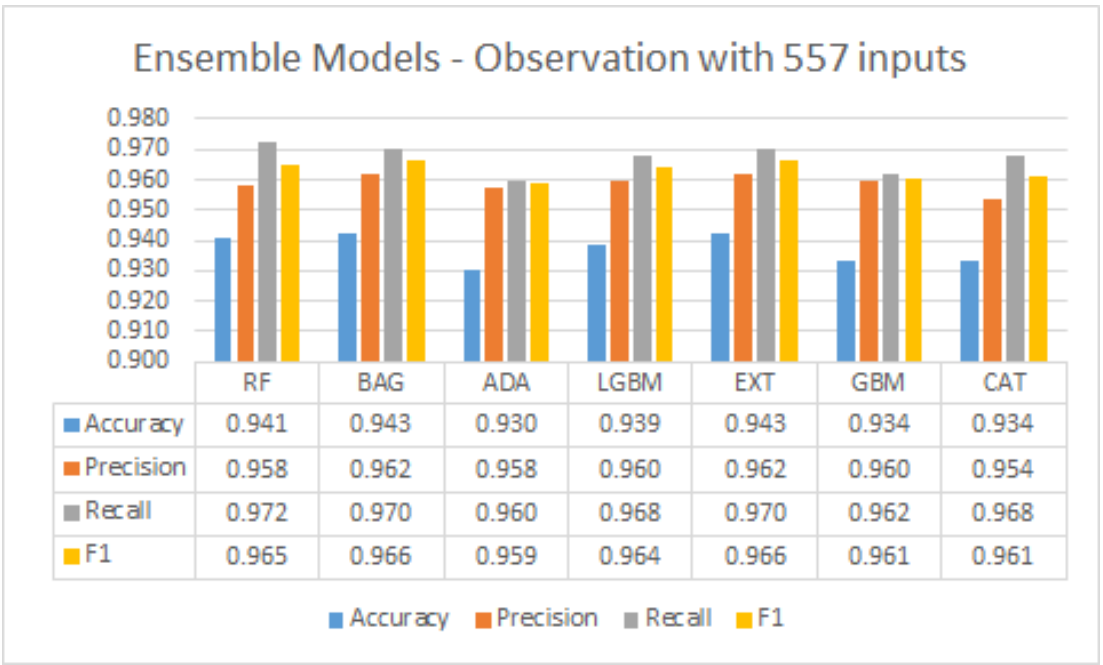


Fig. 41. Ensemble Models with 557 inputs

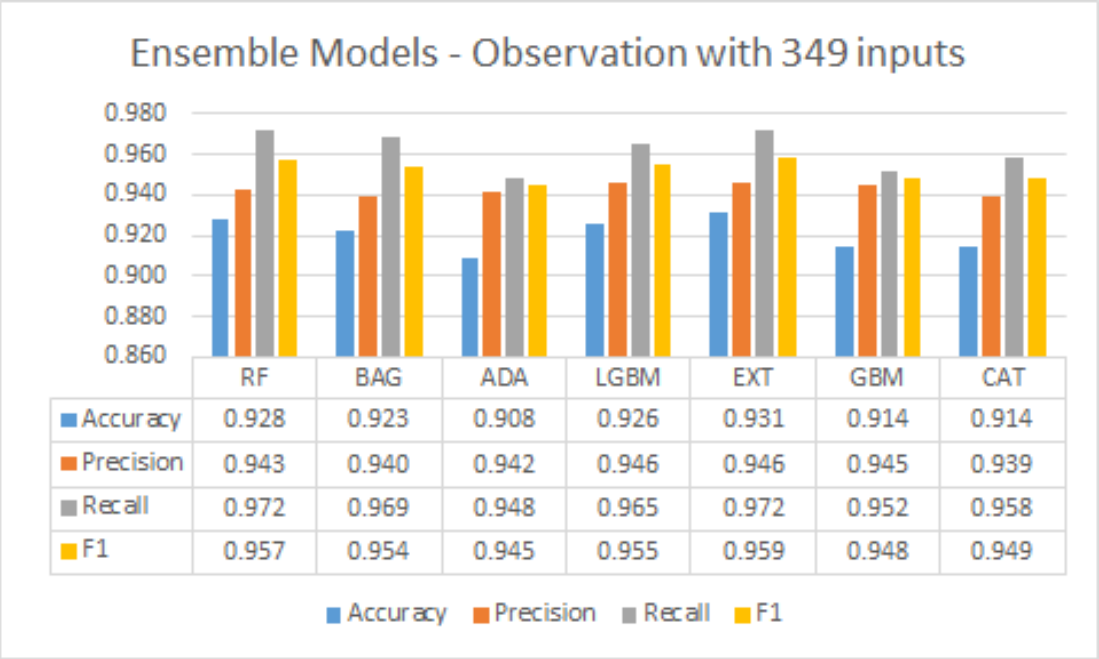


Fig. 42. Ensemble Models with 349 inputs

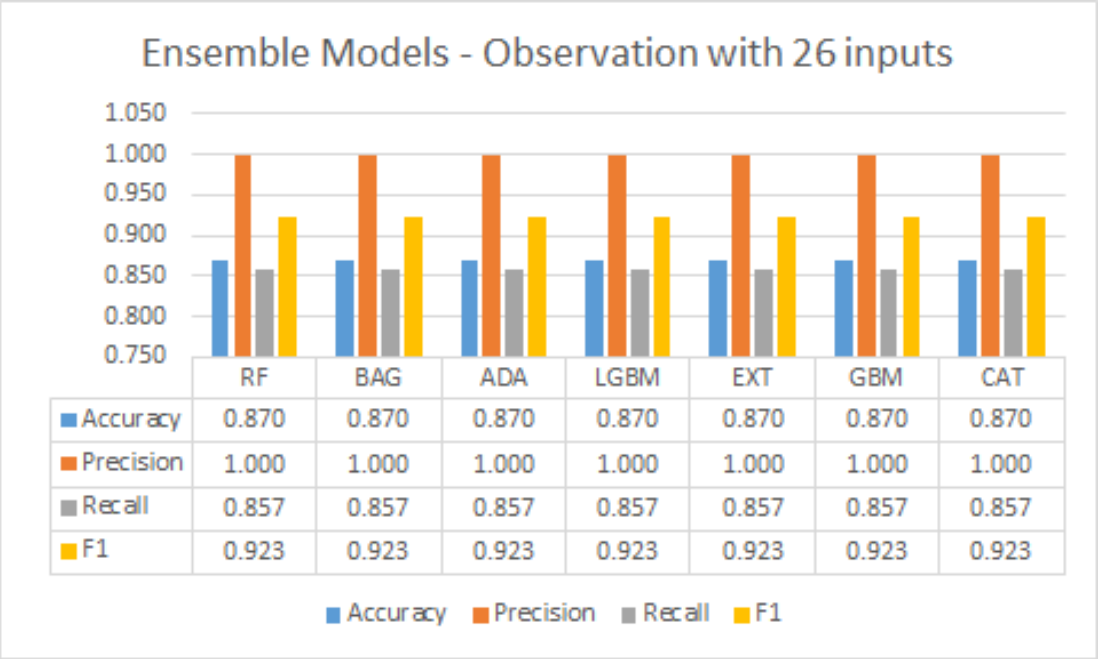


Fig. 43. Ensemble Models with 26 inputs

4.8. Hybrid Model Result

Though ensemble models gave ~94.3% prediction accuracy, the purpose of creating hybrid model is still test tosee if the score can improve by 0.1 or 0.2 % at the least. Here we used stacking and voting approach to determine the final prediction percentage. In hybrid approach, only the top four models we used for evaluation purpose.

4.8.1. Voting Approach

Taking the top four model as input for voting hybrid approach, the prediction percentage has increased to 94.25%along with 96+% precision, recall and F1 score.

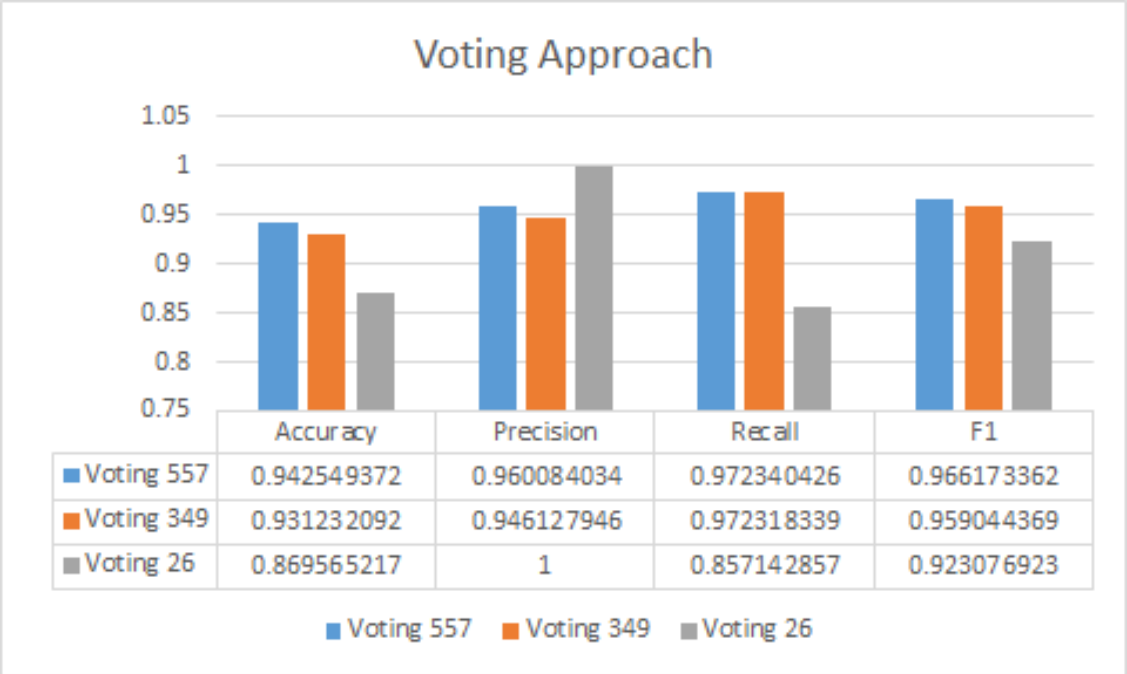


Fig. 44. Voting Approach

4.8.2. Stacking Approach

Taking the top four model as input for voting hybrid approach, the prediction percentage has increased to 94.25%along with 96+% precision, recall and F1 score.

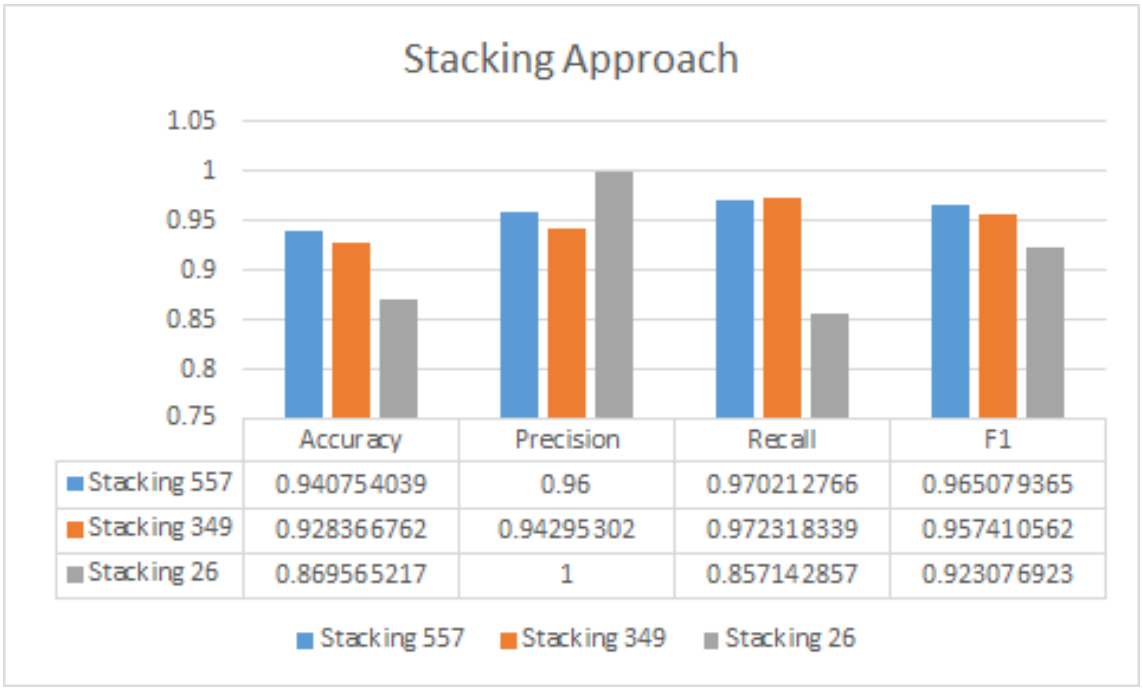


Fig. 45. Stacking Approach

5. DISCUSSION

Extracting the features that influence the student performance is a difficult task. Using algorithms to supplement machine learning techniques used for school students' academic performance prediction (SAPP), feature extraction has been done to analyze the impact on the final prediction. After training and testing various machine learning models that belong to standard (base) classification algorithms, ensemble techniques and hybrid of mixing classification models, it is observed that based on the input size, the prediction accuracy varies. From two samples of 557 and 349 inputs used for prediction, it appears that **Extra Tree Ensemble model** gives a better prediction.

In the voting hybrid approach, it has to be noted that only the top four standalone models were considered for finalprediction as other will not have impact on the final prediction. In fact it could lead to less accurate prediction.

This research experiment result shows that hybrid model didn't significantly improve the score. When such situation happens, it is better to go with individual model itself. From the final comparison, **Extra Tree Ensemblemodel** stands out the winner with 94.3% accuracy.

TABLE VII. FINAL COMPARISON

Model Name	**Type**	**Accuracy**
Decision Tree	Base Classifier	93.4%
Extra Tree Classifier	**Ensemble Classifier**	**94.3%**
Voting	Hybrid	94.25%
Stacking	Hybrid	94.0%

Limitations:

1. The sample dataset used is commonly available for any research and includes only the features more applicablefor 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.

6. CONCLUSION

Over several years various researches has been done in education sector to identify student success as early as possible in their academic career. Early prediction of a school student's academic performance is therefore of utmost importance in order to enable the educational institutions to develop a strategy and plan to implement necessary proactive measures to enable the student to succeed. Through technology, in particular Machine Learning, it is possible to process small to huge datasets to predict academic performance of students.

Though it is widely observed that the same machine learning models gives different accuracy levels for various researchers, it is has been proven that the features or attributes chosen, play a crucial role in calculating the final accuracy. Therefore, it can be inferred that same machine learning model gives different accuracy due to the usage of varying datasets and versatile attributes.

Overall, this research, after conducting various trial and error experiments has provided support to the world of Educational Data Mining that, data, when used with Machine Learning Technology can help identify school students for whom educators can provide proactive, early intervention to ensure academic success.

7. FUTURE WORK

This research paper has provided a simulative study using various ML models that are commonly used for School Academic Performance Prediction. In this initial study, out of the diverse methods used to measure student performance, the **Extra Tree Classifier** has proved to be best model for prediction with an accuracy score of **94.3%.**

In this research, 8000 observations were used for analysis. 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 outcome instead 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.

3. New novelty approach can be devised to improve the prediction accuracy.

REFERENCES

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

[2] Kazeem Moses Abiodun, Emmanuel Abidemi Adeniyi, Dayo Ruben Aremu, Joseph Bamidele Awotunde, Emmanuel Ogbuji: Predicting Students Performance in Examination Using Supervised Data Mining (2022).

[3] 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).

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

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

[6] 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).

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

[8] Sarah Alturki, Nazik Alturki: Using Educational Data Mining To Predict Students' Academic Performance For Applying Early Interventions (2021)

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

[10] 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).

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

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

[13] Derinsha Canagareddy, Khuslendra Subarayadu, and Visham Hurbungs: A Machine Learning Model to Predict the Performance of University Students (2021)

[14] 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).

[15] Leila Ismail, Huned Materwala, Alain Hennebelle: Comparative Analysis of Machine Learning Models for Students' Performance Prediction (2021).

[16] Jitendra Darji, Tulsidas Nakrani: Enhance Student Learning Experience By Using Machine Learning To Predict Student Performance In Advance (2021)

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

[18] Eyman Alyahyan, Dilek Düştegör: Predicting academic success in higher education: literature review and best practices (2020).

[19] Fatima Alshareef, Hosam Alhakami, Tahani Alsubait, Abdullah Baz: Educational Data Mining Applications and Techniques (2020).

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

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

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

[23] Samuel-Soma M Ajibade, Nor Bahiah Binti Ahmad, Siti Mariyam Shamsuddin: Educational

Data Mining: Enhancement of Student Performance model using Ensemble Methods (2019).

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

[25] Vaibhav Kumar and M. L. Garg: Comparison of Machine Learning Models in Student Result Prediction (2019).

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

[27] Ferda Ünal: Data Mining for Student Performance Prediction in Education (2018).

[28] Raza Hasan, Sellappan Palaniappan, Abdul Rafiez Abdul Raziff, Salman Mahmood, Kamal Uddin Sarker: Student Academic Performance Prediction by using Decision Tree Algorithm (2018).

[29] Ismail Almuniri, Aiman Moyaid Said: Predicting the performance of school: Case study in Sultanate of Oman(2018).

[30] S.A. Oloruntoba, J.L.Akinode: Student Academic Performance Prediction Using Support Vector Machine (2017).

[31] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi, Sunday O. Olatunji: Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbor (2017).

[32] Lovenoor Aulck, Nishant Velagapudi, Joshua Blumenstock, Jevin West: Predicting Student Dropout in Higher Education (2016).

[33] Mashael A. Al-Barrak and Muna Al-Razgan: Predicting Students' final GPA using decision trees (2016).

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

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