To propose Hybrid Classification model for the students' performance prediction

Manoj Kumar [1,*], Dr. Raghav Mehra[2,*]

[1]Research Scholar, Computer Science & Engineering Department, Bhagwant University Ajmer (Rajasthan)
[2]Associate Prof., Computer Science & Engineering Department, Bhagwant University Ajmer (Rajasthan)

Abstract

This Study aims to design EDM has attracted the attention of researchers to enhance the quality of education .Therefore, early prediction of performance is vital to keep students on a progressive track The hybrid classifier model has been proposed for an improvement in the performance of the students. The results of the proposed system will be analyzed with the help of various performance parameters

Keywords- EDM, Big Data in Education, Decision Tree, Methodology.

Introduction:

EDM is a growing area of research that is being used to explore educational data for different academic purposes. In the present era of a knowledge economy, the students are the key element for the socio-economic growth of any country, so keeping their performance on track is essential. .

Research Methodology

This section describes the research methodology in detail. The implementation of the data mining approach is completely described in this section.

The WEKA was used to perform data mining tasks. The research methodology consists of different phases and experiments conducted during this research.

The data collection was based on attributes suggested by researchers as the most rational attributes to predict academic performance at secondary level of education. To reduce the computational complexity while implementing the mining techniques, missing values were also removed.

The pictorial representation of research methodology is given in Figure

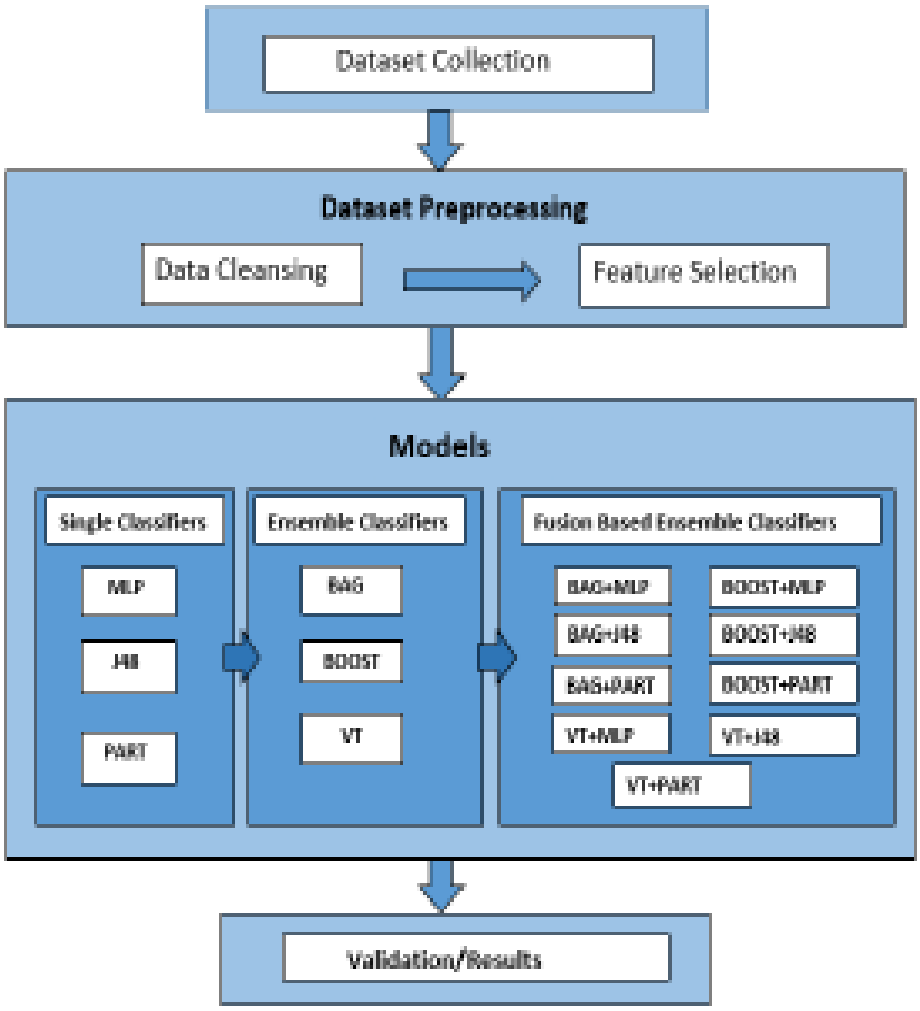


Figure 1. Proposed methodology.

The second step in data pre-processing is feature selection which is used to reduce dimensionality in feature space and obtain better classification results [26] because training on high-dimensional data leads to over fitting of the model. The subset of original features have been picked up through feature selection method which leads to the removal of redundant and obsolete characteristics without losing any important information [26]. This study applied filter-based methods using information gain-based selection to evaluate crucial features which may help in developing good performance models. The class distribution is shown in Table .

Table 1 . Dataset description and possible values

S. No	Attributes	Description
1	GE	Gender (Male, Female)
2	HA	Home Address Urban, Rural)
3	PCA	Parent Cohabitation Status (Living together, Apart)
4	QFR	Quality of family Relationship (Very Good, Good, Not Good)
5	MJ	Mother Job (Yes, No)
6	FJ	Father Job (Yes, No)
7	ME	Mother Education (None, Elementary, Secondary, Higher)
8	FE	Father Education (None, Elementary, Secondary, Higher)
9	FS	Family Size (Less than 3, Greater or equal to 3)
10	GF	Going out with friends (Yes, No)
11	PF	Past Failures (Yes, No)
12	NS	Attended Nursery School (Yes, No)
13	HE	Want to take Higher Education (Yes, No)
14	R	Relationship (Yes, No)
15	IA	Internet access at home (Yes, No)
16	ECA	Extra-Curricular Activities (Yes, No)
17	DST	Daily Study Time (<2 h, 2 to 5 h, 5 to 10 h, >10 h)
18	HST	Home to school Travel Time (<15 min, 15 to 30 min, 30 min to 1 h, >1 h)
19	EG	8th Class Grades (A+, A, B+, B, C, D)
20	NG1	9th Class First Term Grades (A+, A, B+, B, C, D, F)
21	NG2	9th Class Final Term Grades (A+, A, B+, B, C, D, F)

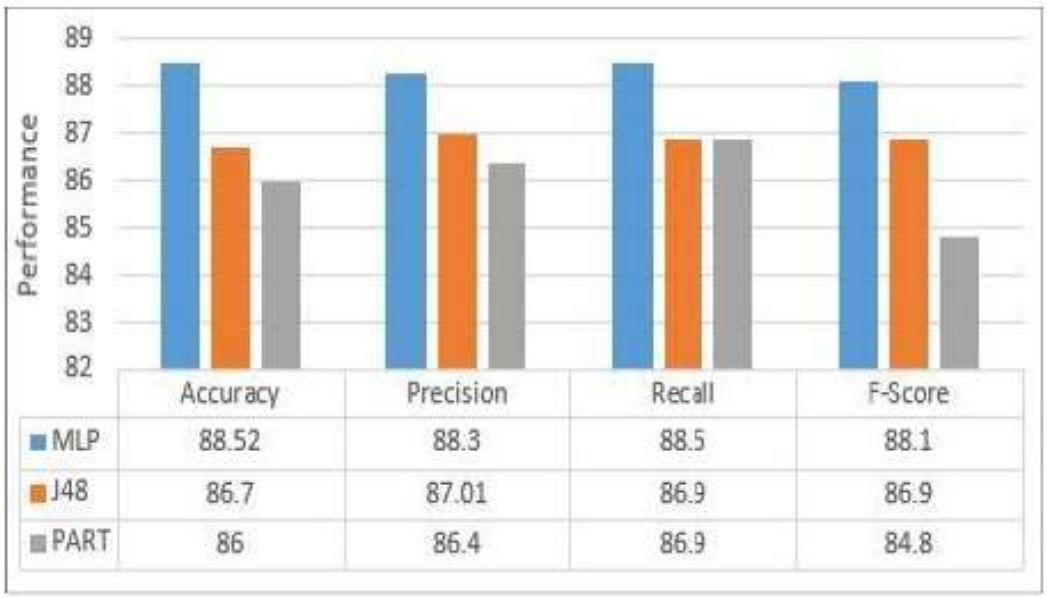
The research methodology is mainly premised on ensemble methods including bagging, boosting, and stacking, which is a different kind of ensemble method which uses a blend of models [29]. Among these methods, bagging, boosting, and stacking can be utilized for classification and prediction. Each model has some strengths and limitations, so the ultimate objective of ensemble methods is to complement the models, in order to achieve higher prediction accuracy. The bagging method is used to sort the tuples randomly into different bags while developing a model.

Experiments and Evaluation

WEKA was used to evaluate the proposed classification model and to make comparisons. In this study, different experiments were conducted sequentially to assess students' performance. The comparison was made through various single base classifiers, ensemble based Classifiers , and fusion ensemble classifiers. The time complexity of each algorithm is also represented in terms of Big O notation which plays an important role in finding the efficiency of algorithms. Additionally, a comparative analysis has been performed to discover performance improvements in different models. The experiments detected the efficient model in predicting student academic performance at the secondary level. To acquire precise results during evaluation, 10-fold cross-validation was used.

Experiments with Base Classifiers and Ensemble Base Classifiers:

The evaluation results showed that among these three base classifiers, MLP outperformed the other classifiers, achieving greater accuracy (i.e., 88.52) as other classifiers, achieving greater accuracy (i.e., 88.52) as shown in Figure 2.



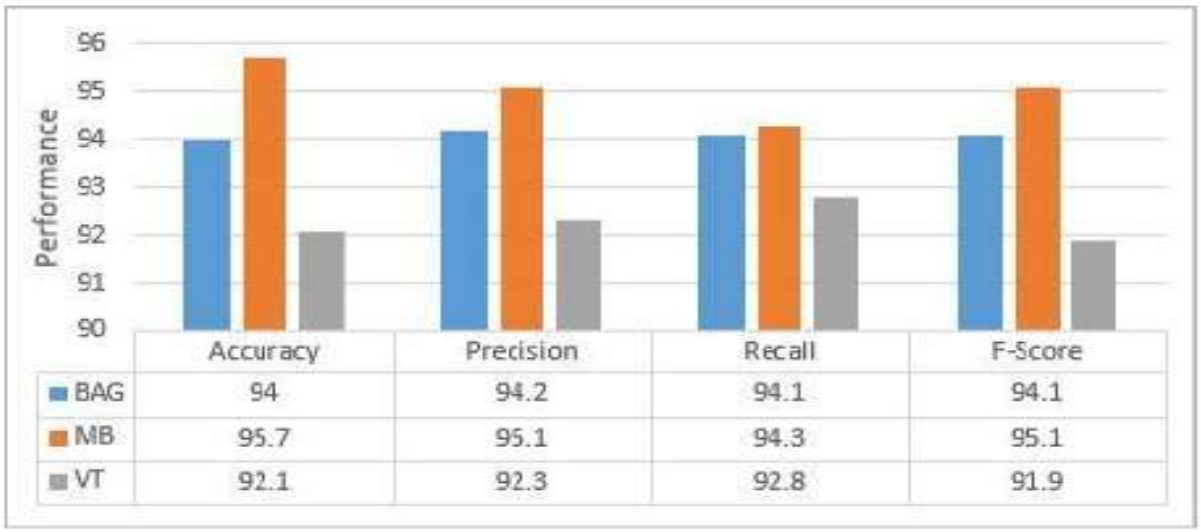
	Accuracy	Precision	Recall	F-Score
MLP	88.52	88.3	88.5	88.1
J48	86.7	87.01	86.9	86.9
PART	86	86.4	86.9	84.8

Figure 2. Single-based classifiers.

Furthermore, three different assembly classifiers including bagging, multi boost, and voting were built. amongst these three ensemble classifiers, multi boost outperformed the other classifiers, achieving higher accuracy (i.e., 95.7) as shown in Figure 3. The figure comprises two parts; in the first part, the bar chart shows the performance of classifiers in terms of accuracy, precision, recall, and F-score. The second part indicates the performance of classifiers in tabular form through the same measures. The classifier also performed better in terms of other measures such as meticulousness recall, and F-score. The time complexity of bagging is (O(klogn), where k is the number of bag.

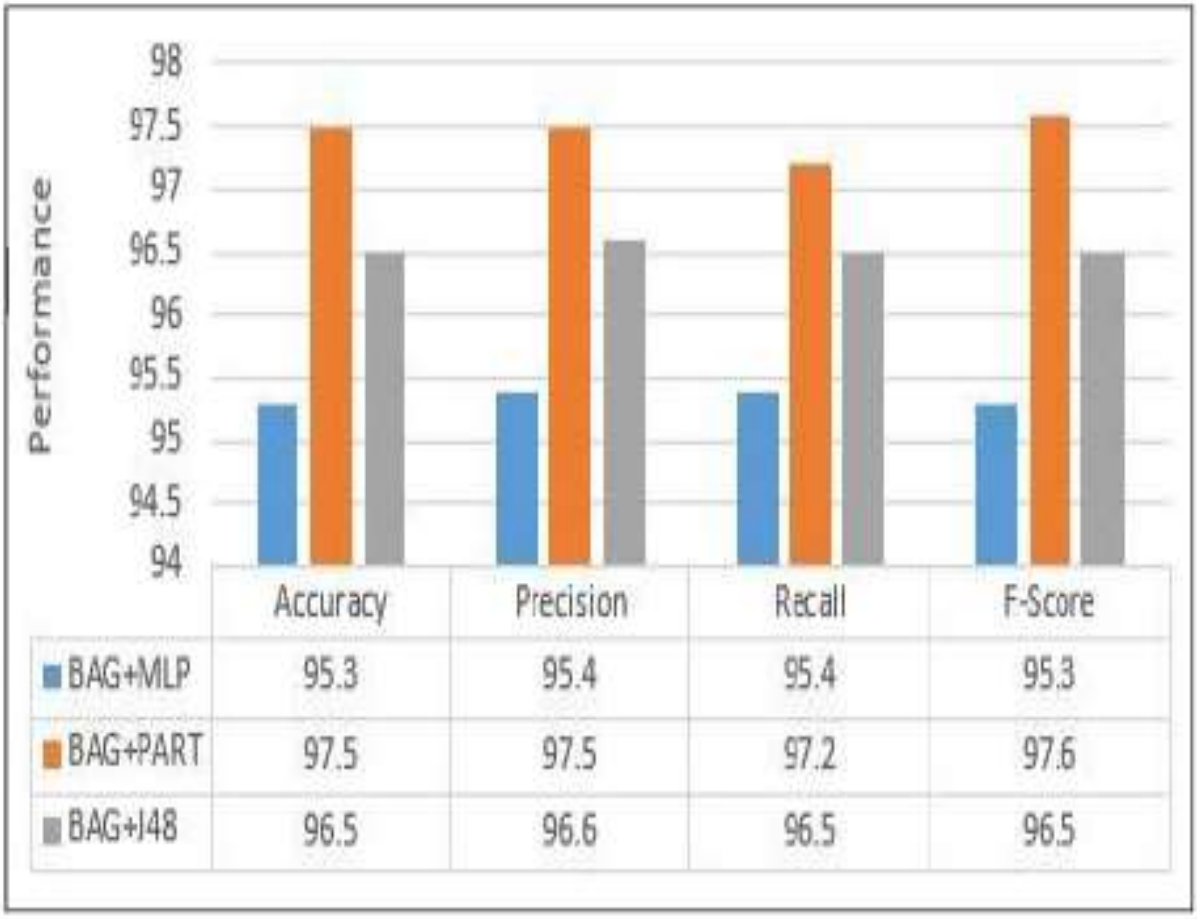
Experiments with Fusion Ensemble-Based Models:

The aim of this phase was to develop hybridization of ensemble classifiers with single based classifiers. The results of these models are shown in Figures 4-6.



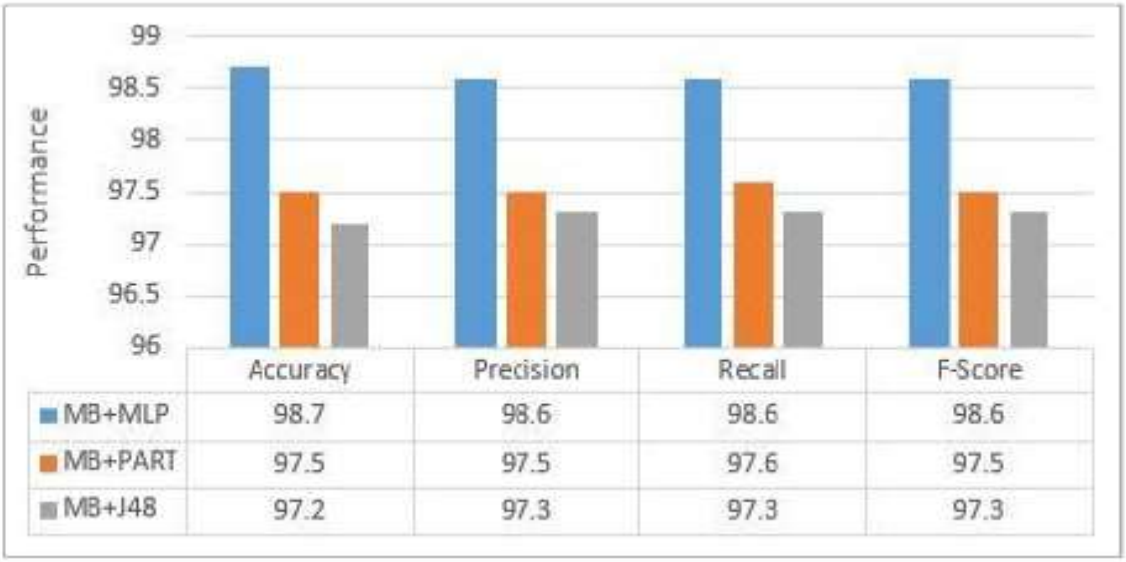
	Accuracy	Precision	Recall	F-Score
BAG	94	94.2	94.1	94.1
MB	95.7	95.1	94.3	95.1
VT	92.1	92.3	92.8	91.9

Fig 3 Ensemble based classifiers



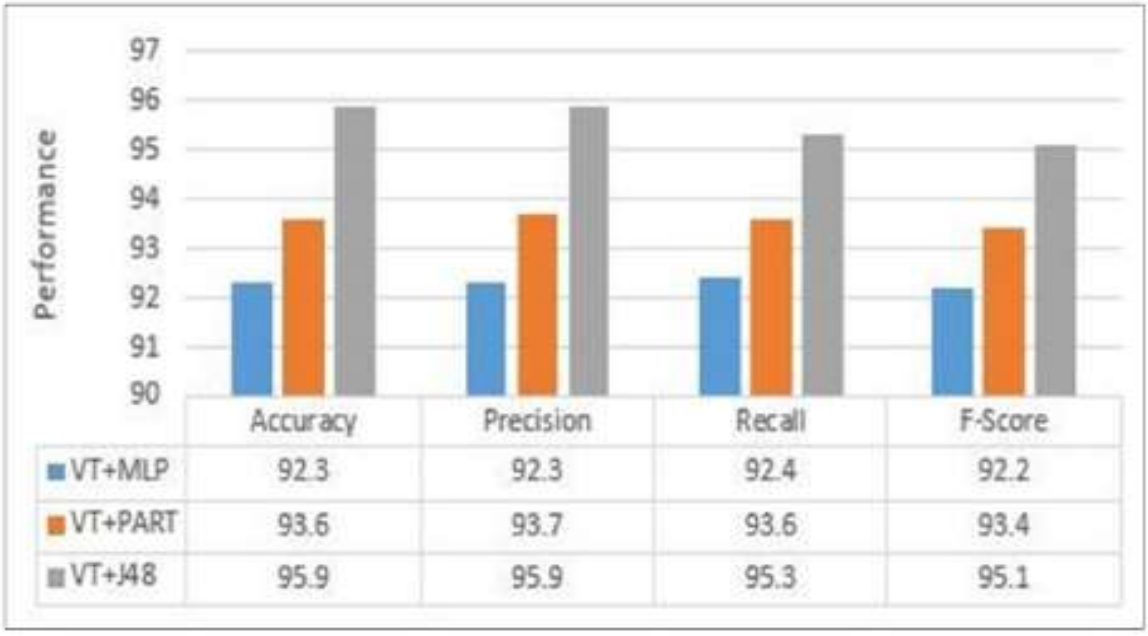
	Accuracy	Precision	Recall	F-Score
BAG+MLP	95.3	95.4	95.4	95.3
BAG+PART	97.5	97.5	97.2	97.6
BAG+I48	96.5	96.6	96.5	96.5

Fig 4 Bagging With Single Based Classifier



	Accuracy	Precision	Recall	F-Score
MB+MLP	98.7	98.6	98.6	98.6
MB+PART	97.5	97.5	97.6	97.5
MB+J48	97.2	97.3	97.3	97.3

Figure 5. MultiBoostAB with single-based classifiers.



	Accuracy	Precision	Recall	F-Score
VT+MLP	92.3	92.3	92.4	92.2
VT+PART	93.6	93.7	93.6	93.4
VT+J48	95.9	95.9	95.3	95.1

Figure 6. Voting with single-based classifiers.

The evaluation results related to BAG fusion with PART showed the highest accuracy (i.e., 97.50%). The model also performed very well with respect to precision, F-score and recall, as shown in Figure 4. This model also showed better performance in terms of precision, recall, and F-score.

Decision Tree: This algorithm has provided a completely orthogonal approach for tackling the issue of classification with the utilization of a tree structure so that the observation input is drawn to a classification outcome. A model is developed on the basis of a tree structure by the means of C4.5 algorithm. A test is exposed on attributes via every internal node; an outcome of the test is demonstrated using every branch and each leaf node is employed to denote a class label. In the training, the finest attribute is discovered in this algorithm for dividing the data at a given node on the basis of relative information gain ratio. The division is going on until the node is turned into a leaf node [13]. Information gain ratio assists in evaluating the correlation of a feature label with class label. I general case, two discrete random variables X and Y are used and the gain ratio is expressed as

$$GAINRATIO(X|Y) = \frac{H(X) - H(X|Y)}{H(X)}$$

Where

$$H(X) = -\sum_{x_i} p(x_i) log p(x_i)$$

And

$$H(X|Y) = -\sum_{j} p(y_i) \sum_{i} p(x_i|y_j) log p(x_i|y_i)$$

A number of improvements are integrated through the version including pruning which has major aim to mitigate the data over-fitting. has been widely used for student performance prediction. This algorithm performs efficiently concerning recall and precision for almost all scenarios.

Result and Discussion:

Python can be easily used by Rapid Application Development. The scripting in this tool is used to interlink the already existing components.

Dataset Characteristics	**Multivariate**
Number of Instances	649
Attribute Characteristics	Integer, Real
Associate Tasks	Classification
Dataset Characteristics	**Multivariate**
Number of Attributes	33
Missing Values	No
Area	Business
Date Donated	2016-01-26

Table 2: Dataset Description

To evaluate the performance of proposed work performance parameters like accuracy, recall, execution time.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	True Negative (TN)

Table 3: Confusion Matrix

Term	**Description**
Positive (P)	When an observation is positive this term is used.
Negative (N)	When observation is not positive, this term is used.
True Positive (TP)	When the observation is positive and is also predicted to be positive, the term TP is used.
False Negative (FN)	When the observation is positive however, it is predicted to be negative, this term is used.
True Negative (TN)	When the observation is negative and it is predicted to be negative, this term is used.
False Positive (FP)	When the observation is negative however, it is predictive to be positive, this term is used.

Table 4: Description of Common Terms Used in Confusion Matrix

The important performance metrics are:

a. Recall: The ratio of number of times the model predicts positive cases correctly to the total number of actual positive cases is known as recall.

$$Recall = \frac{TP}{TP+FN} \quad \ldots (1)$$

b. Accuracy: The ratio of number of times all correct predictions to the total number of inputs is called Accuracy.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots (2)$$

c. Precision: The ratio of number of times the model correctly predicts positive cases to the total number of positive cases predicted by it is called precision.

$$Precision = \frac{TP}{TP+FP}$$

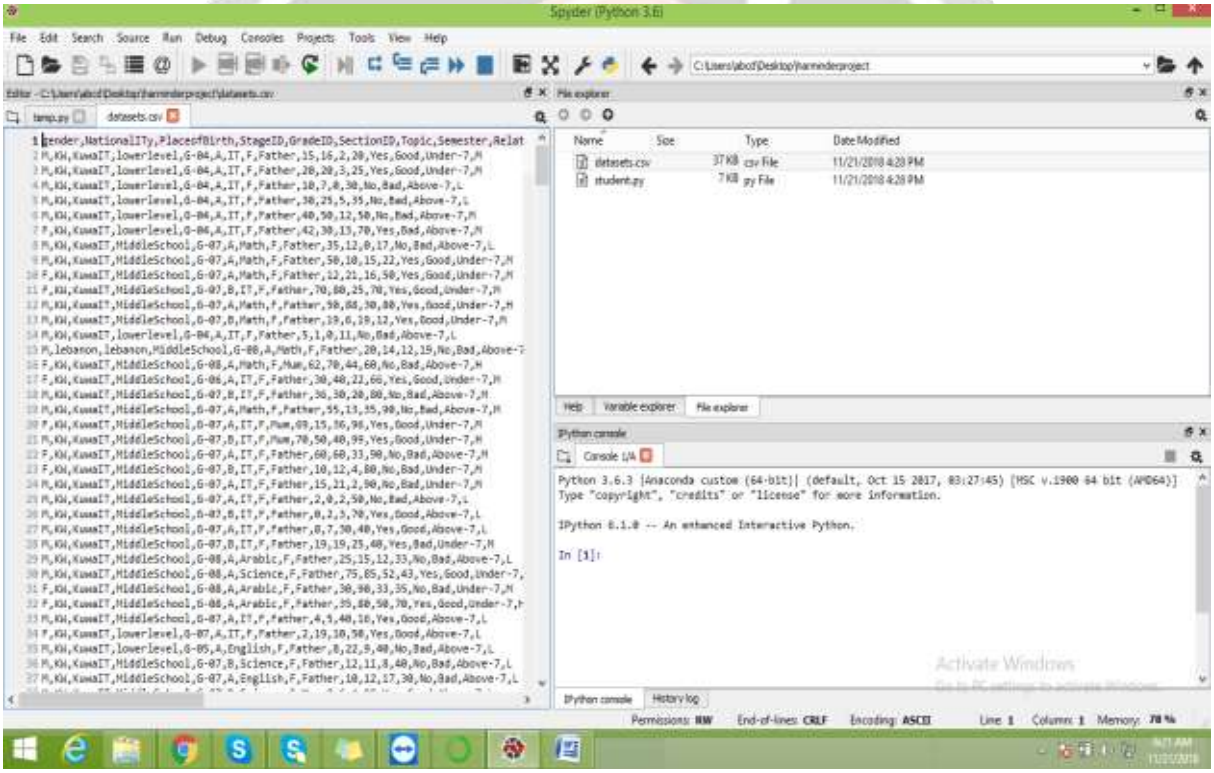


Fig 7: Anaconda default interface

Figure 7 shows the default interface of anaconda. Here, a console, editor and interface of anaconda are shown in the default interface.

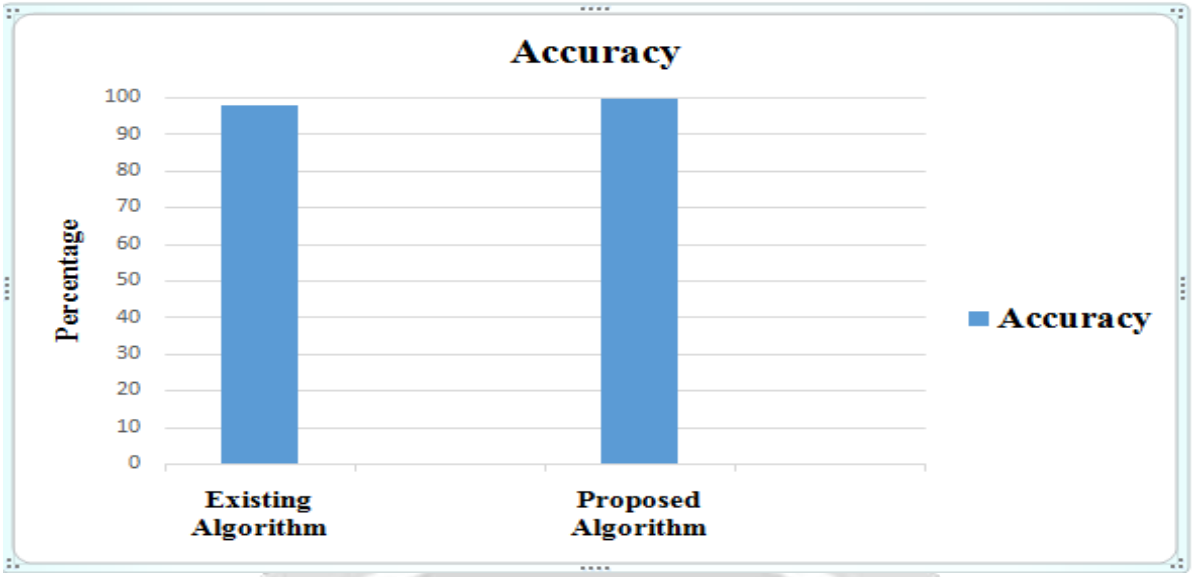


Fig 8: Accuracy Comparison

Figure 8 shows the comparative analysis of proposed and existing algorithms in terms of accuracy. Here, in comparison to existing algorithm, higher accuracy is achieved by applying proposed algorithm.

Precision-Recall Analysis: - The precision-recall is other parameters which define the accurate prediction of the target set. The value of precision-recall is shown in figure below

Fig 9 Precision-Recall Analysis

As shown in figure 5, the precision-recall value of existing algorithm and proposed algorithm is compared for the performance analysis. It is analyzed that proposed algorithm has more precision-recall value as compared to existing algorithm

F Measure: - The F measure is the parameter which defines the average value of the precision and recall. The F measure of the proposed algorithm is shown in figure given below

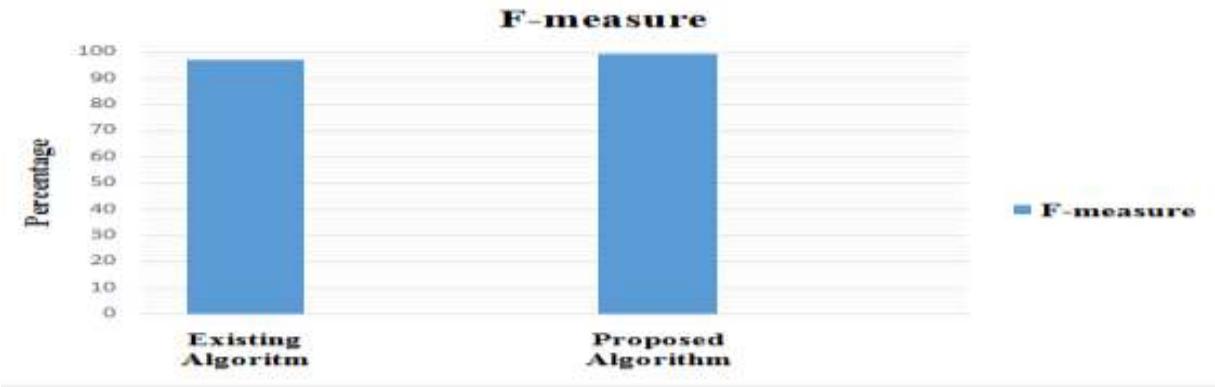


Fig 10F1 Measure

As shown in figure 10, the existing and proposed algorithms are compared in terms of F measure. The existing algorithm has low F measure score and compared to proposed algorithm

Parameter	*Existing Algorithm*	*Proposed Algorithm*
Accuracy	*97.86 %*	*99.5 %*
Precision	*0.97*	*0.99*
Recall	*0.96*	*0.98*
F Measure	*97 %*	*99 %*

Table 3: Performance Analysis

Table 3 shows the comparative analysis of existing and proposed algorithms with respect to various performance metrics in a tabular form. Here, it is concluded that the overall performance of proposed method is better than the existing algorithm.

Conclusion:

The prediction analysis can be applied with the approach classification. The classification techniques can classify data into certain target sets. In this existing system, the technique of back propagation is applied for the student performance prediction. In this research work, hybrid classification approach will be designed based on the decision tree and random forest classifier. The decision tree classifier will works like the Meta classifier and random forest will works like base classifier. The proposed algorithm will be implemented in Python and results will be analyzed in terms of accuracy, precision, recall and f measure.

References:

1. Raheela Asif, Agathe Merceron, Syed Abbas Ali, Najmi Ghani Haider, "Analyzing Undergraduate Students' Performance Using Educational Data Mining", Computers & Education, Volume 113, October 2017, Pages 177-194
2. Sneha Chandra, Maneet Kaur," Enhancement of Classification Accuracy of our Adaptive Classifier using Image Processing Techniques in the Field of Medical Data Mining", 2015, IEEE
3. Yomna M. ElBarawy, Ramadan F. Mohamedt and Neveen I. Ghali," Improving Social Network Community Detection Using DBSCAN Algorithm", 2014, IEEE
4. Dominik Fisch, Edgar Kalkowski, Bernhard Sick," Knowledge Fusion for Probabilistic Generative Classifiers with Data Mining Applications", 2018, IEEE
5. Dianwei Han, Ankit Agrawal, Wei–keng Liao, Alok Choudhary," A novel scalable DBSCAN algorithm with Spark", 2016 IEEE International Parallel and Distributed Processing Symposium

Workshops
6. Li, F.; Zhang, Y.; Chen, M.; Gao, K. Which Factors Have the Greatest Impact on Student's Performance. J. Phys. Conf. Ser. 2019,1288, 012077. [CrossRef]
7. Francis, B.K.; Babu, S.S. Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. J. Med. Syst. 2019,
8. Md Zubair Rahman, A.M.J. Model of Tuned J48 Classification and Analysis of Performance Prediction in Educational Data Mining. 2018.