Predicting students' academic grade using machine learning algorithms with hybrid feature selection approach

Snehal Paddalwar[1,*], *Vanita* Mane[2], and *Leena* Ragha[3]

[1]Department of Computer Engineering, Ramrao Adik Institute of Engineering, Dr. D.Y.Patil Deemed to be University, Nerul, Navi Mumbai, India
[2] Department of Computer Engineering, Ramrao Adik Institute of Engineering, Dr. D.Y.Patil Deemed to be University, Nerul, Navi Mumbai, India
[3] Department of Computer Engineering, Ramrao Adik Institute of Engineering, Dr. D.Y.Patil Deemed to be University, Nerul, Navi Mumbai, India

Abstract. Data plays an important role where any prediction is to be made. Due to the advancement in technology there have been significant ways of collecting students' data. Educators use lots of supportive digital system to teach and collect data, such as Learning Management System, Student Information system, etc. The data collection through these system is huge and nevertheless it can be called as a big data. Lots of research is being done in how to improve the students' performance and overall educational performance and to tap key areas that may help in student progress. Traditional analysis of data includes sampling, whether it be a climate prediction or any performance prediction. In this research, the data collected through classroom observation, academic performance or learning management systems is used to predict the performance of the students. Lots of factors were present in the collected data but the main factors that add relevance for appropriate prediction are only few. This paper proposes how more than one feature selection algorithm results can be combined to improve the predictions with machine learning algorithm.

1 Introduction

The academic or non-academic performance of the students plays a vital role on education system. The performance of students in academics can be affected by various factors. Early predictions of students' performance help in taking prior measures to improve the performance [1] and tap the key areas of students. Data sets related to students can be collected through various means. Some of them are Learning management system, daily observations, quarterly or half yearly assessments results. All these datasets come with a lot of features relating to each student. It is not always all the features that are contributing in prediction accuracy. Some of the features turn out to be irrelevant or might mislead the prediction model. Also, many of the features can act as an outlier that can degrade the accuracy of the model. Selecting the right features for the prediction model is very important to maintain the consistency and accuracy of the results.

Feature selection is selecting related attributes or features to the target results. The features that are selected are from the same dataset. Feature selection plays an important role in data mining and machine learning. On the other hand, feature extraction is different than feature selection. In feature extraction, new variables or attributes are created from the dataset that defines similarity or group of common attributes [3].

The features that are selected from the dataset should improve the accuracy of machine learning models avoiding the overfitting issues. With the help of feature selection, the dataset dimensions can be reduced to a greater extent. Thereby, reducing the time required for computation or training the model. Machine learning is studying the computer algorithms that gets improved with tine and training. Machine learning is classified into supervised, unsupervised, semi supervised and reinforcement learning. This paper focuses impact of feature selection on supervised learning algorithms [4].

2 Background

2.1. Supervised machine learning

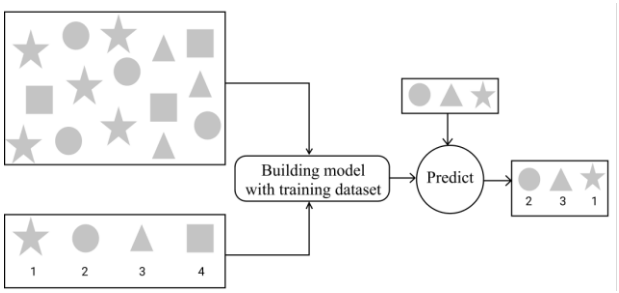


Fig. 1. Work Process of Supervised Machine learning.

* Corresponding author: snehal1335.p@gmail.com

The dataset is labelled in supervised learning. The model that is generated has knowledge of all the types of data in dataset. After the model is trained with the help of some training data, it is then tested using the test data from the same dataset.

2.1.1 Types of supervised machine learning

Supervised learning can be classified into Classification and Regression. Whenever the output of the variable is continuous regression machine learning is used. If the output or the target is discrete then classification machine learning is used [5]. Following are some of the commonly used classification machine learning algorithm in prediction of student academic performance:
- Decision Tree
- Naive Bayes
- K-Nearest Neighbour (KNN)

2.2 Feature selection and its methods

Feature selection is a process in which irrelevant and repetitive features are removed from the dataset. This is done to improve the performance of machine learning algorithms. The factors that defines the performance improvement are accuracy and time required for building the model. Feature selection methods are Filter based, Wrapper based and Embedded based. Out of these methods, filter based methods proved out to perform best when it comes to high dimensional data and time. Filter based methods works independently on each feature and thereby do not look for correlation among the features. In case of Wrapper method, the computational time is little more than filter based method but it looks for correlation between the features while interacting with the classifier [8]. Selection of relevant features is very important when it comes to improving the efficiency and accuracy.

2.2.1. Filter based methods

These methods are based on ranking of features. The ranking technique is easy and provides better results. Filter-based methods used the techniques of Chi Square test, Correlation coefficient, fisher score.

2.2.2. Wrapper based methods

These methods use a greedy search to retrieve the most important and appropriate features that helps in providing the best results. It is basically comprising two types of algorithm sequential and heuristic search.

The sequential search starts by considering an empty set and each feature is then added or remove from the set until the objective function is achieved. Whereas, the heuristic search algorithm basically works on evaluating the dataset to achieve the objective function [10].

3 Literature Survey

In 2019, Kumar, T. R., Vamsidhar, T., Harika, B., Kumar, T. M., & Nissy, R. has used four classification algorithms Naïve Bayes, ID3, C4.5, Random Tree, out of these four Naïve Bayes algorithm performed best in providing good classification of prediction. The main aim of the paper was to compare the classification of the methodology and to predict the student performance [5].

In 2017, Ihsan A. Abu Amra and Ashraf Y. A. Maghari has used two classification algorithm KNN and Naïve Bayes on educational dataset. The agenda was to make early prediction so that it can be used to increase student performance. They have concluded Naïve Bayes performed better than KNN algorithm. As a future work, they have suggested that more classification algorithms can be applied to different education institute datasets. [6].

In 2019, Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. have used KNN, DT and Naïve Bayes classifiers were used on the dataset of 230 students of Kabul University to predict their GPA as high, medium and low. They concluded that KNN showed the highest accuracy rate among others. [7].

In 2016, Danasingh, Asir Antony & Balamurugan, Suganya & Epiphany, Jebamalar Leavline done a literature review on various featured selection based techniques on high dimensional data. They found that filter based methods are more efficient and takes less computational time than wrapper and embedded methods. Also, the filter based method performs much better with classification algorithms [8].

In 2016, Ms.Tismy Devasia1 ,Ms.Vinushree T P2, Mr.Vinayak Hegde3 used Naïve Bayyes algorithm to predict the student performance and proved that it performed better than Decision tree, Regression methods [9].

In 2021, Kaur, A., Guleria, K., & Kumar Trivedi, N. has made a comparative study of different feature selection algorithm and presented advantages and research gaps. They have provided an insight to various feature selection methods that includes filter based, wrapper based and embedded methods. [10] .

4 Proposed Work

The proposed solution aims to overcome the drawbacks of individual feature selection methods. Two feature selection methods can be combine to overcome the limitations of each method. Based on the literature review, we have used filter based and wrapper based feature selection method as a hybrid method to predict the academic performance of students.

Algorithm:

Step 1: Select and analyse the dataset.
Step 2: Select features using Filter based feature selection approach.
Step 3: Select features using Wrapper based feature selection approach.
Step 4: Compare the selected features and choose the features that are commonly selected by both the approaches.
Step 5: Select the relevant machine learning algorithm that best fits with the type of dataset.
Step 6: Use the selected features to train the machine learning model and check the accuracy of the model.
Step 7: Compare the results of individual feature selection based approach with hybrid based approach.

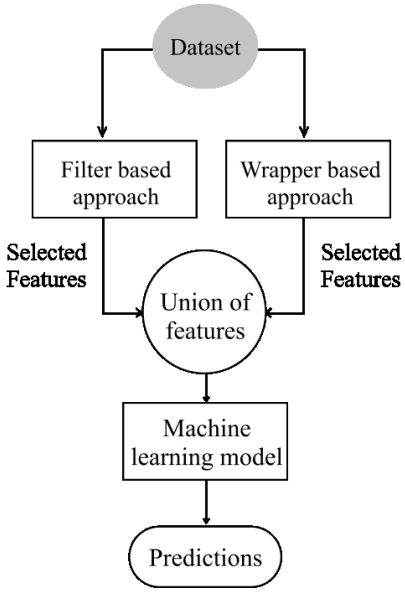


Fig. 2. Generalised Working of Proposed Method

5 Implementations and Results

5.1. Dataset overview and Pre-processing

The dataset is from the UCI Machine Learning repository. The data contains the students' performance in two subjects: Mathematics and Portuguese with additional features that may or may not impact the final grades achieved by the student. Term 1 score and Term 2 score has a strong relation with final grades, but some of the other features shows relevant impact on scoring final grade.

5.2. Feature selection methods used

Two Feature Selection approaches were used. Filter-based and Wrapper based. One of the filter based selection method is Chi-Square test. It is applied to evaluate the correlation in the dataset. In this test, the Chi-square between each feature and the target feature is calculated where the best chi-square feature is selected [10]. Chi-square test is ranking based algorithm and hence works with individual features, without looking out for the dependency on other feature set.

$$X^2 = \frac{(O-E)^2}{E}$$

O - Observed value
E - Expected value

The other approach used was Wrapper-based feature selection. One of wrapper feature selection method RFE (Recursive Feature Elimination) method was used to select the relevant features. RFE deals with the correlation with each feature and the target feature. It is a method that helps in fitting the model by removing the weakest feature until the specific number of features are related to the target.

Initially, the data was processed to categories the final grade into 4 classes. This 'final grade' column was kept as a target feature. The rest 32 features were used for training the model.

Firstly, with 32 features, a model based on decision tree, Naïve Bayes and KNN was build. Corresponding accuracy, precision, recall and f1 scores were also calculated.

Secondly, using Chi-square test and RFE, 10 features were selected by each algorithm. Later, these 10 features of each algorithm was compared and union was formed. It is observed that 15 features were ranked and selected by both the algorithms. Using this 15 features a model based on decision tree, Naïve Bayes and KNN was build. Corresponding accuracy, precision, recall and f1 scores were also calculated.

Later, it was observed that only selection of 15 features by both the algorithm have increased the accuracy, precision, recall and f1 score.

5.3. Results

Accuracy defines how the model have behaved or performed across all classes. It is calculated by the ratio between the correct predictions made out of the total available predictions.

Dataset (No. of Features)	Decision Tree	Naïve Bayes	KNN
Mathematics (32)	80 %	70 %	57 %
Portuguese (32)	79 %	48 %	68 %

Table 1. Prediction accuracy without feature selection.

Dataset (No. of Features)	Decision Tree	Naïve Bayes	KNN
Mathematics (10)	75 %	75 %	50 %
Portuguese (10)	72 %	57 %	59 %

Table 2A. Prediction accuracy with single feature selection method (Chi2)

Dataset (No. of Features)	Decision Tree	Naïve Bayes	KNN
Mathematics (10)	69 %	55 %	53 %
Portuguese (10)	79 %	54 %	47 %

Table 2B. Prediction accuracy with single feature selection method (RFE)

Dataset (No. of Features)	Decision Tree	Naïve Bayes	KNN
Mathematics (15)	87 %	84 %	67 %
Portuguese (15)	83 %	83 %	77 %

Table 3. Prediction accuracy with feature selection (hybrid method).

Dataset: Mathematics, No. Features: 32			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	80	83	81
Naïve Bayes	67	74	68
KNN	69	45	47
Dataset: Portuguese, No. Features: 32			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	68	77	72
Naïve Bayes	45	55	41
KNN	55	41	44

Table 4. Classification report without feature selection.

Dataset: Mathematics No. Features: 10 Feature selected using one method (Chi-2)			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	74	77	75
Naïve Bayes	73	82	75
KNN	73	48	50
Dataset: Portuguese No. Features: 10 Feature selected using one method (Chi-2)			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	69	76	72
Naïve Bayes	67	54	57
KNN	78	55	59

Table 5. Classification report with single feature selection (Chi-2) method.

Dataset: Mathematics No. Features: 10 Feature selected using one method (RFE)			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	68	72	69
Naïve Bayes	65	67	65
KNN	76	50	53
Dataset: Portuguese No. Features: 10 Feature selected using one method (RFE)			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	68	71	79
Naïve Bayes	67	61	54
KNN	49	45	47

Table 6. Classification report with single feature selection (RFE) method.

Dataset: Mathematics No. Features: 15 Features selected by hybrid approach			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	84	87	85
Naïve Bayes	86	84	85
KNN	65	59	59

Dataset: Portuguese No. Features: 15 Features selected by hybrid approach			
Dataset (No. of Features)	Precision	Recall	F1-Measure
Decision Tree	71	79	74
Naïve Bayes	70	79	74
KNN	82	81	81

Table 7. Classification report with feature selection.

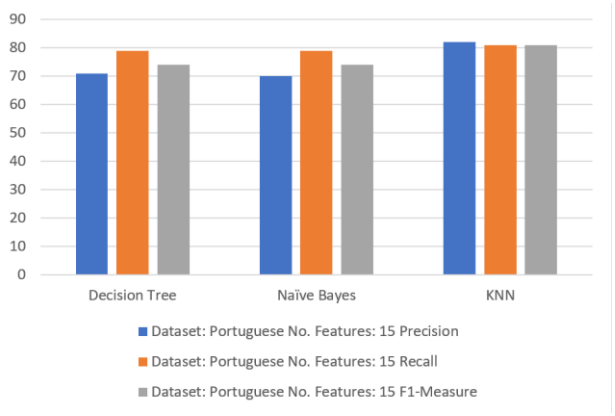
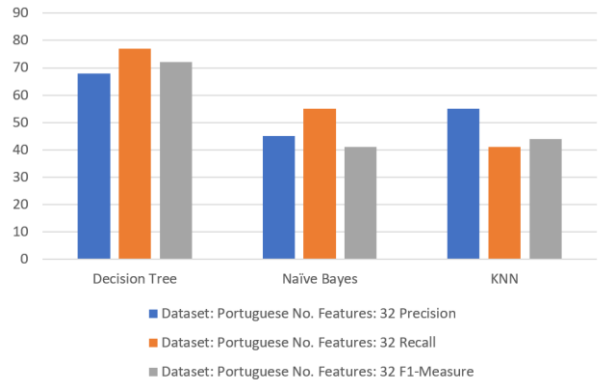
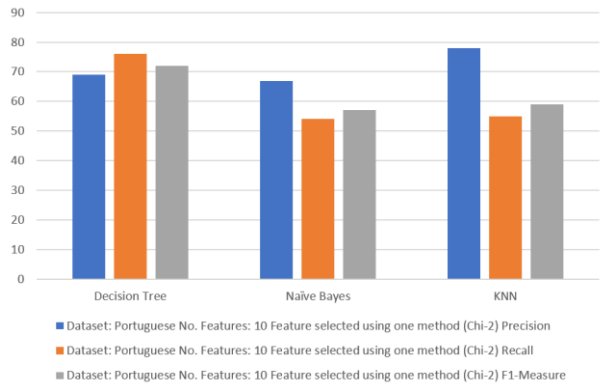
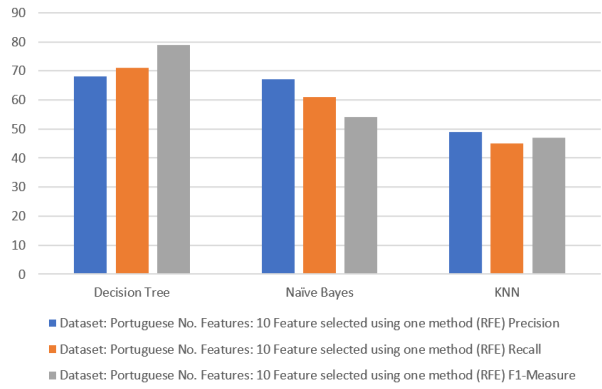
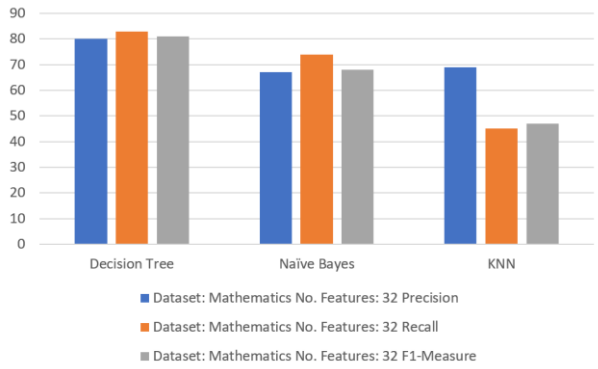


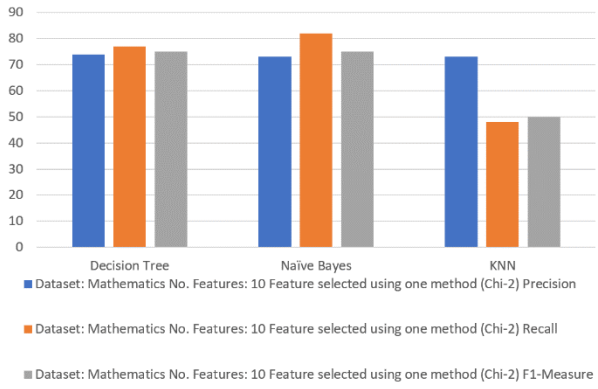
Fig. 3. Classification report comparative graph for Mathematics dataset

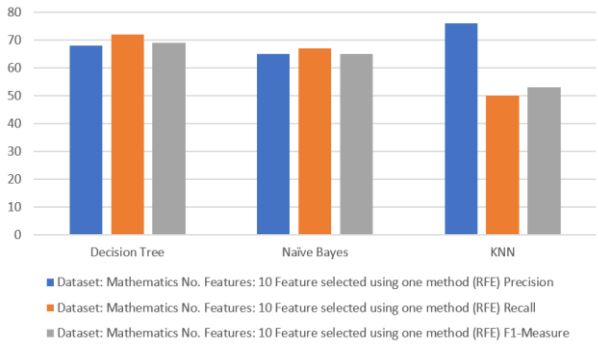












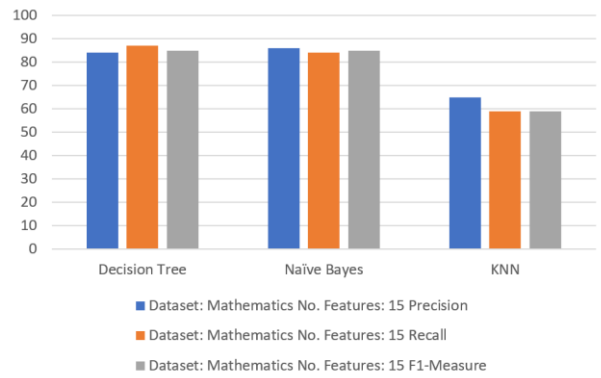


Fig. 4. Classification report comparative graph for Portuguese dataset

Conclusion

Feature selection plays a very important role in machine learning and prediction. With hybrid feature selection the number of features were reduced to more than 50%. Also the accuracy of the different machine learning algorithm prediction has increased to a greater extend. Hence by considering two different feature selection output and providing a union of these output to a machine learning algorithm will increase the prediction accuracy.

References

[1] Gull, H., Saqib, M., Iqbal, S. Z., & Saeed, S. (2020). Improving Learning Experience of Students by Early Prediction of Student Performance using Machine Learning. 2020 IEEE International Conference for Innovation in Technology (INOCON).

[2] Thomas, R. N., & Gupta, R. (2020). Feature Selection Techniques and its Importance in Machine Learning: A Survey. 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS).

[3] Aparna U.R. and S. Paul, "Feature selection and extraction in data mining," 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-3.

[4] T. R. N and R. Gupta, "A Survey on Machine Learning Approaches and Its Techniques:" 2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2020, pp. 1-6.

[5] Kumar, T. R., Vamsidhar, T., Harika, B., Kumar, T. M., & Nissy, R. (2019). Students Performance Prediction Using Data Mining Techniques. 2019 International Conference on Intelligent Sustainable Systems (ICISS).

[6] Amra, I. A. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. 2017 8th International Conference on Information Technology (ICIT).

[7] Mohammadi, M., Dawodi, M., Tomohisa, W., & Ahmadi, N. (2019). Comparative study of supervised learning algorithms for student performance prediction. 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).

[8] Danasingh, Asir Antony & Balamurugan, Suganya & EPIPHANY, JEBAMALAR LEAVLINE. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. International Journal of Computer Applications.

[9] Prediction of students' performance using Educational Data Mining. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE).

[10] Kaur, A., Guleria, K., & Kumar Trivedi, N. (2021). Feature Selection in Machine Learning: Methods and Comparison. 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE).