I.J. Modern Education and Computer Science, **2023, 3, 44-54**
Published Online on June 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2023.03.04

Towards Finding a Minimal Set of Features for Predicting Students' Performance Using Educational Data Mining

Souvik Sengupta*
Aliah University, Kolkata, India
E-mail: mesouvik@hotmail.com
ORCID iD: https://orcid.org/0000-0003-1842-7523
*Corresponding Author

Received: 20 October, 2022; Revised: 24 November, 2022; Accepted: 12 January, 2023; Published: 08 June, 2023

Abstract: An early prediction of students' academic performance helps to identify at-risk students and enables management to take corrective actions to prevent them from going astray. Most of the research works in this field have used supervised machine learning approaches to their crafted datasets having numerous attributes or features. Since these datasets are not publicly available, it is hard to understand and compare the significance of the chosen features and the efficacy of the different machine learning models employed in the classification task. In this work, we analyzed 27 research papers published in the last ten tears (2011- 2021) that used machine learning models for predicting students' performance. We identify the most frequently used features in the private datasets, their interrelationships, and abstraction levels. We also explored three popular public datasets and performed statistical analysis like the Chi-square test and Person's correlation on its features. A minimal set of essential features is prepared by fusing the frequent features and the statistically significant features. We propose an algorithm for selecting a minimal set of features from any dataset with a given set of features. We compared the performance of different machine learning models on the three public datasets in two experimental setups- one with the complete feature set and the other with a minimal set of features. Compared to using the complete feature set, it is observed that most supervised models perform nearly identically and, in some cases, even better with the reduced feature set. The proposed method is capable of identifying the most essential feature set from any new dataset for predicting students' performance.

Index Terms: Educational Data Mining, Machine Learning, Students performance prediction, Feature analysis, Feature selection, Decision Support System.

1. Introduction

Educational data mining (EDM) is the application of data mining and machine learning methods to aid education technology. EDM is used to find out interesting patterns and knowledge from students' data with an aim to support decision-makers at educational institutions in better understanding students' academic progress and achievements, identifying customized learning priorities for various student groups, and developing learning strategies accordingly.

There has been a significant growth of interest among researchers in the last decade in the field of predicting students' future performance using supervised machine learning (ML) models. Most of the researchers have conducted classification tasks like selecting a student's future grade (multiclass) or detecting whether students will pass or not (binary). On the other hand, some of the researchers have performed regression tasks like predicting overall numbers or grade points. In either case, the most crucial factor is identifying the proper attributes or features from the student-related information. In the rest of the paper, we use the terms' attribute' and 'feature' interchangeably.

The prediction accuracy of a machine learning model highly depends on the features of the dataset on which it is trained. However, most of the researchers in this field have reported their work performed on private datasets having numerous attributes. The datasets are prepared from institutional LMS, college/university students' information consists

primarily of demographic and academic data. Therefore, it is challenging to analyze the influence of different predictive features and performances of different predictive models used in different works.

Traditionally educational institutions collect large volumes of student-related data according to organizational and management policy, often without any specific objective. On the other hand, Learning Management System (LMS) software tracks different student information throughout the learning sessions, including grades, attendance, activities, engagements, etc. Therefore, appropriate feature selection is vital in designing and predicting models for students' future performance. ML algorithms like Logistic Regression (LR), Decision Tree (DT), Support vector Machine (SVM), K-Nearest Neighbor (KNN), Naïve Bayes (NB), Random Forest (RF), Adaptive Boosting (Adaboost) are widely used by the researchers for predicting students' future academic performance.

This objective of this work is to develop a methodology for identifying the minimal set of features from student-related information that could be used for ML based prediction of students' performance. The methodology should be able to perform generalized feature selection across diverse type datasets. This minimal set of features could be used as a baseline to test any performance improvement by the ML model on the inclusion of any unique feature.

The rest of the paper is organized as follows. Section 2 describes the review of related works. Section 3 depicts the methodology in detail. The results and analysis of the empirical study is narrated in section 4. Finally, section 5 concludes the paper.

2. Reviewed Works

Many researchers have used ML models for predicting students' academic performance from different student-related information with or without feature analysis. Han et al. [1] compared the performance of different ML models in predicting the GPA grades of undergraduate students in Chinese Universities. They created a dataset with records of 123 undergraduate students in four-year programs with a total of 20 professional core courses. They analyzed correlation and association among the courses for feature selection and observed the highest prediction accuracy of 91.67 % by the AdaBoost classifier. Anuradha et al. [2] created a dataset from students' records of three private colleges in the Tamil Nadu state of India. They tried to predict the end-semester performance of the students from the previous semester's marks as well as the demographic and pre-collegiate characteristics of the students. The authors tested multiple ML models on the Weka platform and observed the highest accuracy of 68.3% by the KNN classifier. Osmanbegović et al. [4] tried to determine dominant factors in students' performance prediction. They prepared a private dataset of 1210 students' records with 19 attributes collected from secondary-level schools in Bucharest. They achieved the highest accuracy of 73.2% with the RF classifier. Acharya et al. [5] prepared a dataset of 403 students data with 14 attributes collected from undergraduate colleges in Kolkata, India. They used chi-square and Information Gain for feature selection. The best classification accuracy is reported as 66% by SVM with Sequential Minimal Optimization (SMO) algorithm. Amra et al. [6] presented a student performance prediction model using KNN and NB classifier. This work is based on a dataset prepared from students' records of secondary schools in the Gaza Strip. The authors performed manual preprocessing by eliminating some of the attributes. In the prediction job, the NB model achieved the highest accuracy of 93.6%. Jalota et al. [8] performed two types of feature selection- wrapper-based and correlation-based. The dataset is prepared from the users-log of an institutional LMS, consisting of 480 records with 14 attributes. The authors reported that SVM and DT worked better with a correlation-based filter, whereas NB performed well with wrapper-based feature selection.

On the other hand, some of the predictive models used all the features of the dataset without any feature selection. Kabra et al. [7] presented a dataset of 346 students from one Indian institute running an engineering program that includes demographic data like category, gender, and past performances at 10^{th} and 12^{th} standards. The prediction is performed with a single DT classifier, which achieved 69.94 % accuracy. Devasia et al. [9] worked with a dataset prepared from records of 700 students and 19 attributes with data collected from Indian Universities. This work reported that the NB classifier worked better than the other models in predicting students' performance. Bhardwaj et al. [10] prepared a dataset of 300 students' with data collected from different degree colleges in India. It recorded the highest prediction accuracy of 86.25% by the NB classifier. Pandey et al. [11] integrated three ML classifiers in predicting students' performance from students' social and educational background information. A dataset of 600 students' records was collected from different colleges in India. A voted aggregation of predictive models achieved the highest accuracy of 87.03%. Abdullah et al. [12] proposed students' performance prediction using multi-agent data mining. The authors used a tiny dataset containing 155 students' records on a single course. They compared the performance of DT with AdaBoost and observed that the ensemble method with 80% accuracy worked better than the single classifier that obtained 74% accuracy.

Although most of the works in predicting students' performance are classification jobs, some researchers tried to predict numeric marks or grade points of the student, which are considered as a regression problem. Arsad et al. [13] presented a study on ANN-based prediction of the academic performance of engineering students. The dataset is prepared from the students' records from a technical university in Malaysia. The ML model was trained on students' scores on the fundamental courses in the first semester to predict the CGPA of the final semester. The best performance is recorded as a

mean squared error (MSE) of 0.0409. Sharma et al. [14] tried to predict students' academic performance based on parental influences. The parental factors, like educational background, job, family size, etc., are included in the dataset [15, 16]. A linear regression model was used for marks prediction and achieved a root mean squared error (RMSE) score of 3.31.

Table 1. Summary of some of the reviewed works

Study by	Country of source data	No. of samples	No. of features	Attribute types	Best predictive model	Highest accuracy %
Han et al. [1]	China	123	20	Course performances	Adaboost	91.67
Anuradha et al. [2]	India	180	19	Demographic, social, family, and academic	KNN	68.3
Osmanbegović et al [4]	Romania	1210	19	Demographic, family, and academic	RF	99.0
Acharya et al. [5]	India	403	14	Demographic and activity data	SVM	66.0
Amra et al. [6]	Palestine	500	8	Course enrollment data	NB	93.6
Kabra et al. [7]	India	346	17	Demographic, family, and academic	DT	69.94
Jalota et al [8]	India	480	14	Academic performances	SVM	85.8
Bhardwaj et al. [10]	India	300	17	Demographic, social, family, and academic	NB	86.42
Pandey et al. [11]	India	960	18	Demographic and academic	DT	98.96
Abdullah et al. [12]	Saudi Arabia	175	9	LMS data	Adaboost	80.0
Ketui et al [22]	Thailand	483	9	Course performances	GradientBoost	92.62
Marbouti et al. [24]	USA	1560	14	Course-related assignments and quizzes	Ensemble of seven classifiers	84.6
Hussain et al. [25]	India	300	24	Demographic, academic, and socioeconomic	RF	99.0
Goga et al. [26]	Nigeria	5100	9	Demographic and educational data	RF	99.82
Miguéis et al [27]	Portugal	2459	15	Demographic, socioeconomic, and educational background data	RF	96.1
Kotsiantis et al. [28]	Greece	1347	4	Assignment performance	Ensemble of six classifiers	78.95

All the above-discussed works used private datasets in predicting students' performance using ML models. However, some of the research works have used public datasets available in Kaggle and UCI for the same task. This work explored three popular academic performance datasets; two are available in UCI Repository, and one in Kaggle. Two UCI repositories are i) the students' academic performance dataset – Portuguese [29], ii) the students' academic performance dataset – Math [29], both having 33 features, and 649 and 395 records, respectively. The other one, iii) student academic performance dataset – xAPI [30], is available in Kaggle Repository that has 480 records and 16 features. Nabil et al. [17] employed classical ML methods and deep learning algorithm for predicting students' academic performance on the xAPI public datasets. They used label encoding to convert categorical data into a numerical features and SMOTE for oversampling. They recoded an accuracy of 89%, 76%, and 75% with DL, SVM, and KNN models respectively. In a similar work, Ismail et al. [3] presented a comparative analysis of different ML models on the student-por and xAPI datasets with an information gain-based feature selection method. They recorded the highest accuracy of 75% and 73% on the Portuguese dataset, with RF and DT models, respectively. Similarly, SVM classifiers recorded 70% and 72% accuracy on the xAPI dataset, with and without feature selection [18,19,20,21].

Researchers have relied on diverse attributes in predicting students' performance; some are very common, and some are unique. Selection of attributes ranging from the dataset consisting of only previous exams' marks to the dataset that includes accounts in social media, smoking/drinking habits, number of friends, parents' feedback, etc. Since none of the reviewed works with private datasets has performed any detailed analysis or ablation study on the used features, it is hard to understand their impact and justification for using them in ML-based predictive models. On the other hand, features used in the different public datasets also vary significantly, and different researchers have reported different accuracy ranges while applying predictive models to them. Therefore, it is important to identify the essential core attributes of student-related information that should be used as an optimized feature set for any predictive model. Table 1 represents a summary of some

of the reviewed works on private datasets, detailing the origin of the dataset, numbers and types of the attributes, numbers of records, best performing ML model, and highest recorded accuracy. It can be observed that, in general ensemble methods performed better than single models. While Adaboost achieved good accuracy with smaller dataset [1,12], RandomForest showed better performance both in smaller [25] and larger dataset [26,27]. Naïve Bayes performed better than other single models like Support Vector Machine and Decision Tree while working mostly on demographic data [6,10]. Although correlation among the attributes is not reported in these works, the success of Naïve Bayes indicates that most of the attributes are mutually independent [31].

3. Methodology

In this paper, we try to address the problem of finding a minimal set of features that can cover the original feature set and predict students' performances with equally good accuracy as using a superset of features.

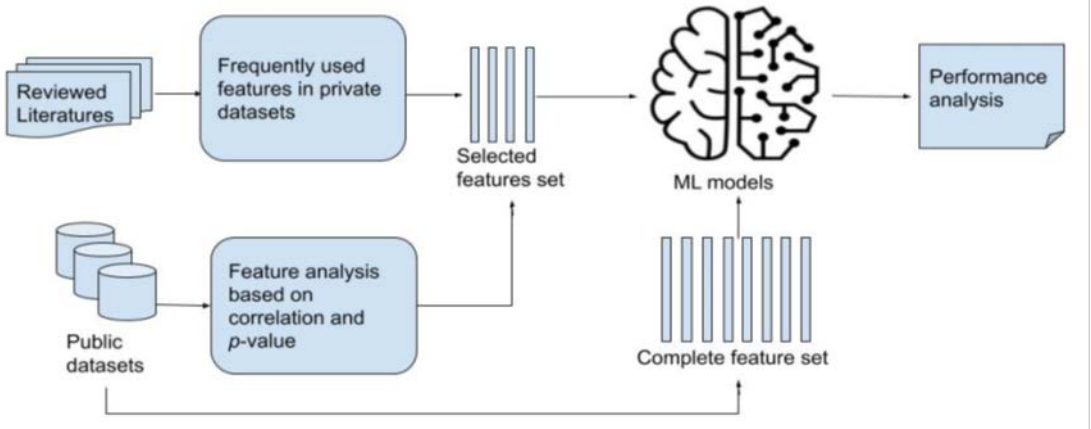

Fig. 1. Framework of the proposed work

Figure 1 represents the framework of the proposed work. It is composed of three components, namely feature selection, predicting students' performance with the entire feature set, and predicting students' performance with the minimal set of features. In the feature selection phase, we prepare two feature lists - common features (CF) and significant features (SF). CF is prepared by identifying frequently used features in the literature with more than 40% participation in all the reviewed works. From the reviewed literature, it is observed that some of these features are unique and can be found only in that particular work, e.g., *no. of friends* and *vehicle type*. On the flip side, it is also observed that certain personal details are very commonly considered, like *age*, *gender*, *father's education*, *mother's education*, *father's occupation*, *mother's occupation*, and *family income*. Academic details like *tenth standard grade* and *twelve standard grade*, and *CGPA of the previous semesters* are also commonly used. Some common social factors are identified as *locality*, *wealth*, and *category*. However, the same attributes may appear with different names, like the *father's qualification* in place of the *father's education* or the *mother's job* instead of the *mother's occupation*. In addition, two semantically same attributes could appear in different forms like *age* and *DoB*, *marks* and *grade*, *country*, and *nationality*.

Some of the attributes are culture-dependent. For example, *parents' marital status* in the USA, and Africa, *smoking habit* in Asia and Latin America, *drinking habit* in Europe and America, *category/caste/religion*, and *family type* in India. On the other hand, attributes like *marital status*, *scholarship*, *number of friends*, *number of siblings*, and *vehicle type* are some of the infrequent features used by the researchers. It is also observed that all these wide ranges of attributes used in different works are not at the same abstraction level, some are generic, and some are specific. Therefore, this work presents an attribute tree that represents levels of abstraction and interrelationships within the attributes. Figure 3 depicts the attribute tree where the attributes are categorized into biological, social, family, and academic attributes at the first abstraction level. Two further detailed layers are created where the leaf nodes represent the CF used in the reviewed datasets.

On the other hand, SF is selected from statistical analysis of the features of the three public datasets. The features available in the three public datasets are analyzed with Pearson Correlation and Chi-square tests. Pearson correlation measures the strength of the linear relationship between two numerical attributes, whereas, Chi-square measures the degree of association between two categorical attributes using the null hypothesis testing. The relationship between a feature and the target variable is estimated by *p*-value. A lesser *p*-value (preferably less than equals to 0.05) indicates the strong significance of the attributes. The categorical features are converted into numerical features using label encoding to measure correlation coefficients. A high correlation between two features implies linear dependency, i.e., two features would have almost the same impact on the dependent variable. Therefore, dropping one of them does not affect the prediction

performance of the model. After eliminating highly correlated features and features with higher *p*-values, we map the rest of the features with the already obtained set of commonly used features, CF. We then fuse the two lists, discarding the redundant features by semantic analysis, to form a minimal set of selected features (SLF). The proposed algorithm for preparing and selecting SLF from any dataset is narrated below.

Algorithm1	**Algorithm for selecting a minimal set of features**
Intialization	**let** U ={all the attributes in all the reviewed works}
	let SF={}
	let CF={}
	let SLF={}
	let thres_corr=0.7
	let thresh_sig=0.05
	let sigma= 0.4
1.	**for each** reviewed paper R
2.	**for each** attribute A reported in R
3.	A^+= {all semantic equivalents of A}
4.	**for all** a in A^+
5.	**if** a ∉ CF **then**
6.	CF= CF ∪ A
7.	**endif**
8.	**for each** public dataset D
9.	**for each** attribute A reported in D
10.	*p*-value= chi-square (D, A)
11.	**if** *p*-value ≤ thresh_sig **then**
12.	SF= SF ∪ A
13.	**endif**
14.	**for each** attribute B in {SF-A}
15.	corr-value= pearson-correlation (D,A,B)
16.	**if** corr-value ≥ thres_corr **then**
17.	SF= SF- B
18.	**endif**
19.	**for each** attribute T in a new dataset S
20.	T^+ = {all semantic equivalents of T}
21.	**for any** attribute t in T^+
22.	**if** t ∈ CF **or** t ∈ SF **then**
23.	SLF= SLF ∪ T
24.	**endif**

Finally, an empirical study is performed with different ML models tested on three public datasets in two experimental setups - first with all the features and then with SLF. The model's efficiency in predicting student's performance is evaluated in terms of four evaluation metrices, namely, accuracy (AC), precision (PR), recall (RC), and F1-score (F1). All these metrices are calculated using four fundamental measures- true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$Accuracy(AC) = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$Precision(PR) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall(RC) = \frac{TP}{TP+FN} \quad (3)$$

$$F1\ score(F1) = 2 * \frac{PR*RC}{PR+RC} \quad (4)$$

The choice of the ML models as Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN), and ensemble methods as Random Forest (RF), and Adaptive Boosting (AdaBoost) is based on the most commonly used models in the reviewed works.

Logistic Regression: LR is a basic ML algorithm frequently used in classification tasks in education technology. The logistic function is used in the model's core to estimate the probability of a target class based on a linear combination of predictor variables or features. This work employs a cost function called Cross-Entropy, called Log Loss, to estimate prediction error. The average cost of all the predictions is referred to as the loss function (equation 5).

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n} Cost(h_\theta(x^{(i)}), y^{(i)}) \quad (5)$$
$$Cost(h_\theta(x), y) = -log(h_\theta(x)) \quad if\ y = 1$$
$$Cost(h_\theta(x), y) = -log(1 - h_\theta(x)) \quad if\ y = 0$$

K-Nearest Neighbor: KNN uses distance between all data points and searches for k nearest data point which are having minimum distance. The majority of class present in these k instances is labeled as the class of the test instance. Although KNN is the simplest ML algorithm for classification, it is called as lazy learner and has high memory cost. Proximity between the data points can be measured using Euclidean distance, Manhattan distance, and Minkowski distance. Euclidean distance is the most commonly used method (equation 6).

$$dist(p, q) = \sqrt{\sum_{i=1}^{n}(p_i^2 - q_i^2)} \quad (6)$$

Naive Bayes: NB is a probabilistic model for classification tasks . The Bayes theorem serves as the fundamental principle for this classifier. P(d|h) is the prior probability of the hypothesis h, P(d) is the probability of the data, and P(h|d) is the posterior probability. A Naive Bayes classifier assumes that the features are conditionally independent of one another. Equation 2 represents the joint probability of observing a certain combination of contextual features is expressed as equation 7.

$$P(h|d) = \frac{P(d|h)*P(h)}{P(d)} \quad (7)$$
$$p(F_1, F_2, \ldots F_n, S) = p(S)\prod_{i=1}^{n} p(F_i|S)$$

Decision Tree: DT uses a set of decisions rules to make classification. It creates a tree structure, where each internal node denotes a feature of the dataset, branches represent an outcome of a splitting criteria, and each leaf node represents a labeled class object. In every splitting it tries to minimize the Entropy (S) based on equation 8, where p_i is the probability of an element in the ith class.

$$Entropy(S) = \sum -p_i\ log\ p_i \quad (8)$$

Support Vector Machine: SVM maximizes the margins for a hyperplane in a high-dimensional feature space to separate data points into different target classes. In general a larger margin increases generalization of the classifier. The hyperplane is constructed using a kernel function, which determines the linear or non-linear characteristics of the hyperplane. The cost function of SVM with regularization parameter is shown in the equation 9.

$$min_\lambda \|w\|^2 + \sum_{i=1}^{n}(1 - y_i(x_i, w)) \quad (9)$$
$$C(x, f(y)) = \begin{cases} 0 & if\ y * f(x) \geq 0 \\ 1 - y * f(x) & else \end{cases}$$

Artificial Neural Network: ANN is the most popular and widely used for classification of non-linearlyseparable data items. A multilayer perceptron (MLP) is a fully connected feed forward ANN. The back-propagation algorithm updates the weights of the connected neurons using stochastic gradient descent (SGD) in contrast performs a parameter update for each training example $x^{(i)}$ and label $y^{(i)}$. (equation 10):

All the above-discussed works used private datasets in predicting students' performance using ML models. However, some of the research works have used public datasets available in Kaggle and UCI for the same task. This work explored three popular academic performance datasets; two are available in UCI Repository, and one in Kaggle. Two UCI repositories are i) the students' academic performance dataset – Portuguese [29], ii) the students' academic performance dataset – Math [29], both having 33 features, and 649 and 395 records, respectively. The other one, iii) student academic performance dataset – xAPI [30], is available in Kaggle Repository that has 480 records and 16 features. Nabil et al. [17] employed classical ML methods and deep learning algorithm for predicting students' academic performance on the xAPI public datasets. They used label encoding to convert categorical data into a numerical features and SMOTE for

oversampling. They recoded an accuracy of 89%, 76%, and 75% with DL, SVM, and KNN models respectively. In a similar work, Ismail et al. [3] presented a comparative analysis of different ML models on the student-por and xAPI datasets with an information gain-based feature selection method. They recorded the highest accuracy of 75% and 73% on the Portuguese dataset, with RF and DT models, respectively. Similarly, SVM classifiers recorded 70% and 72% accuracy on the xAPI dataset, with and without feature selection.

$$\theta = \theta - \eta \nabla_{\theta} J\left(\theta; x^{(i)}; y^{(i)}\right) \tag{10}$$

Random forest: RF is composed of multiple decision trees. Each tree is trained on randomly selected subset of data. It also may not consider all of the features available in dataset. The final prediction of the model is a voting average of predictions of each tree.

Adaptive Boost: AdaBoost is another ensemble algorithm for classification tasks. It combines many weak classifiers (decision trees) and leverages bagging and boosting methods to create one strong classifier. Unlike RF, AdaBoost creates a forest of stumps rather than trees.

4. Result Analysis

In feature extraction, we reviewed 27 research works in the last ten years (2011-2021). Figure 2 shows the frequency distribution of CF having support greater than 40%. Figure 3 depicts the attribute tree prepared from the CF list. It helps in understanding the abstraction level of the attributes and also the interrelationships among them.

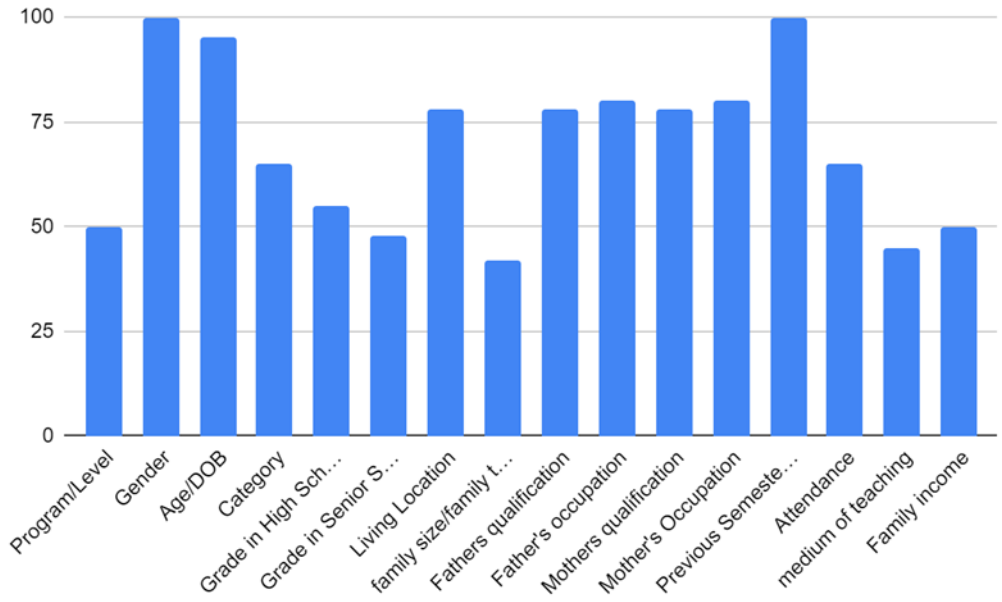

Fig. 2. Frequency distribution of commonly used features

We explored three public datasets, student-por [29], student-math [29], and xAPI [30].

The xAPI dataset is composed of 480 records and 16 features. This educational data set is collected from the log of a learning management system (LMS) using an activity tracker tool called experience API (xAPI). The target class has three values, low (score in the range of 0 to 69), medium (score in the range of 70 to 89), and high (score in the range of 90 to 100).

After applying the SLF algorithm on xAPI features, four attributes, *Gender*, *PlaceofBirth*, *Nationality*, and *StudentAbsenceDays* are selected from mapping with the CF. The Chi-square test yields five more attributes: *Raisedhands*, *VisitedResources*, *Discussion*, *ParentAnsweringSurvey*, and *ParentschoolSatisfaction*. Figure 4 depicts the correlation among the xAPI attributes. From correlation analysis, we discarded *PlaceofBirth* and *ParentAnsweringSurvey* as they are highly correlated with *Nationality* and *ParentschoolSatisfaction*. Finally, 7 attributes out of 16 are grouped as SF.

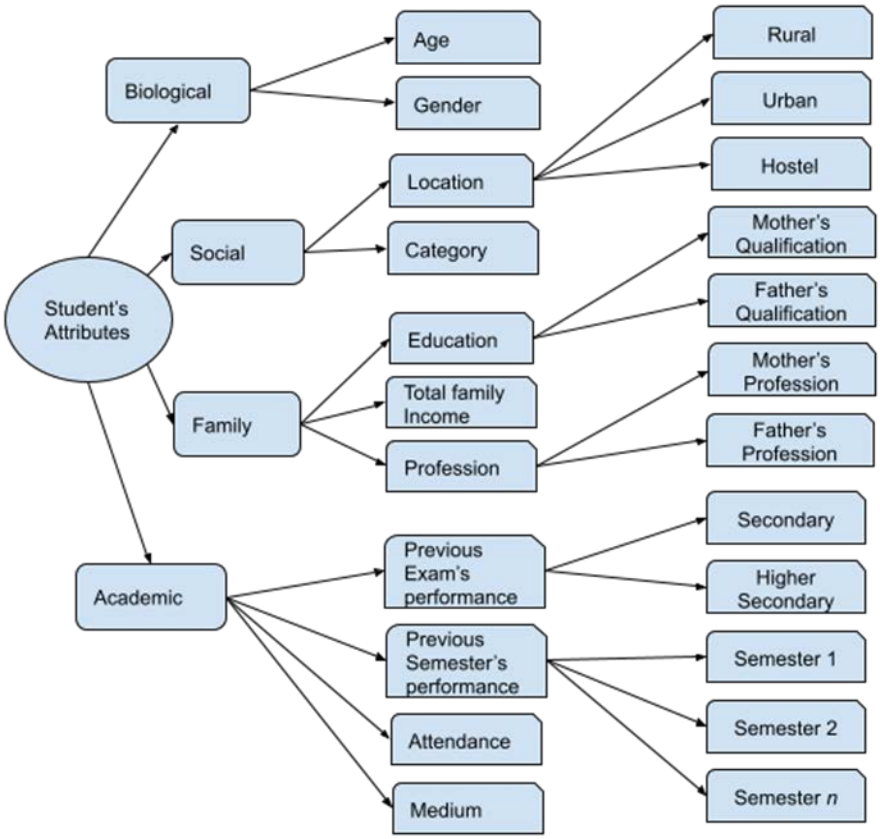

Fig. 3. Attribute tree of the commonly used features

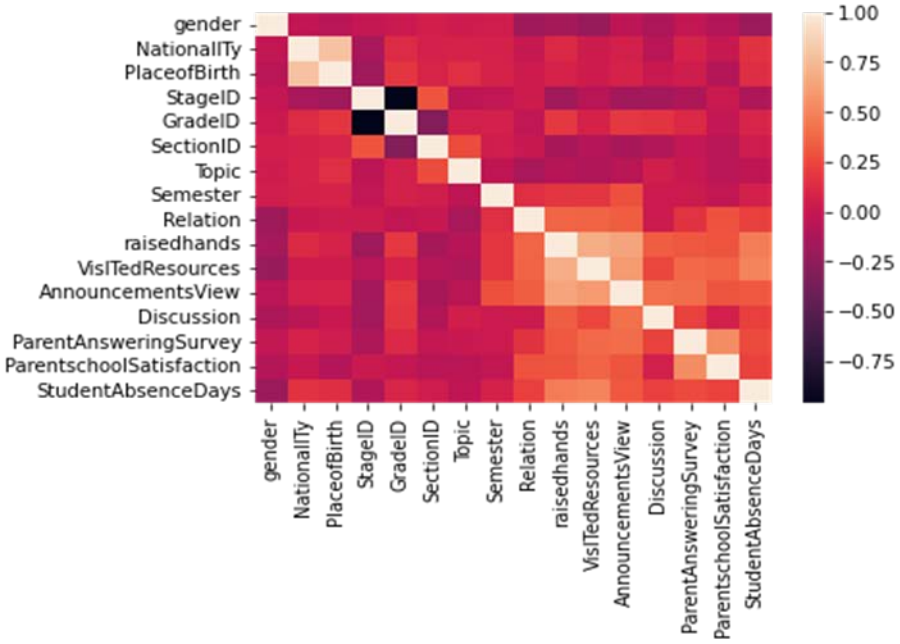

Fig. 4. Correlation among features in xAPI

The other two datasets, student-por and student-mat, are collections of students' performance in secondary education at Portuguese schools in two distinct subjects, mathematics (student-mat) and Portuguese language (student-por). Both datasets use the same set of 32 attributes, while student-mat has 395 and student-por has 650 records. On applying the SLF algorithm on student-mat, we selected 14 attributes. From mapping with the CF list, we obtained - *sex*, *age*, *address*, *Medu*,

Fedu, *Mjob*, *Fjob*, *StudentAbsenceDays*, *walc*, *absences*, *G1*, and *G2*. From the *p*-value analysis, we added *schoolsup*, *reason*, and *failures*. The target attribute *G3* is the final grade that shows a strong correlation with previous grades *G1* and *G2*. However, as this is a trivial correlation, as mentioned in the original documentation of the dataset, they are not considered. Therefore, a total of 14 features are kept as SLF from the original list of 32. Similarly, for the students-por dataset, applying the SLF algorithm yields all the features of student-mat along with two new attributes, *schools* and *dalc*, being included and *address* being discarded. Therefore, a total of 15 features are selected as SLF for this dataset.

Performances of the ML models on the three public datasets are depicted in Table 2. Six single models and two ensemble models are tested in two experimental setups. As the performances of ML models are highly dependent on the appropriate choice of their hyper-parameters, we used GridSearchCV to find out the best combination of hyper-parameters for each model. It is the process of tuning different values of the hyper-parameters to obtain the optimal accuracy for a given model. We used 5-fold cross-validation for all the setups. The best results are obtained with - (penalty='*l2*', solver='*lbfgs*') for LR, (n_neighbors=*3*, metric='*minkowski*') for KNN, (kernel='*rbf*') for SVM, (solver='*adam*', learning_rate='*0.001'*, activation='*relu*') for ANN (*GaussianNB*) for NB, (criterion='*entropy*') for DT. In ensemble models, (n_estimators=*50*, criterion='*gini*') for RF, and (n_estimators=*50*, learning_rate=*0.1*) for AdaBoost, produced best results. It is observed from the results that LR performed best in the xAPI dataset, whereas SVM outperformed other models in both student-por and student-mat datasets. Among ensemble methods, RF performed better than AdaBoost in almost all cases. RF also outperformed all the single models in all the experimental setups. All the best-obtained accuracies are highlighted in bold. Interestingly, in 29.16% of the total cases, the models with the selected feature set performed better than those using the entire feature set (cases underlined). It is observed from table 2 that the average performance increase is 2.36%. In 19.04% of the total cases, there is no change in the performance. For the rest of the cases, the performance drop is 1.92%, which is insignificant considering the fact that only 45% of the average original features are used. It proves that our proposed approach is capable of extracting the minimal set of features from any new dataset, which can be used as a baseline to understand the significance of adding any other feature to the dataset.

Table 2. Performance Analysis of ML models

Model	**Evaluation Metrics**	**xAPI dataset**		**Student-Portuguese dataset**		**Student-Math dataset**	
		Full Feature set	**Selected feature set**	**Full Feature set**	**Selected feature set**	**Full Feature set**	**Selected feature set**
LR	AC	0.7917	0.7722	0.7692	0.7641	0.7983	0.7731
	PR	0.8102	0.7847	0.7776	0.7676	0.8032	0.7762
	RC	0.7795	0.7551	0.7722	0.7761	0.8006	0.7774
	F1	0.7915	0.7785	0.7745	0.7711	0.8010	0.7767
KNN	AC	0.6250	0.6250	0.7385	0.7744	0.8151	0.7731
	PR	0.6618	0.6460	0.7363	0.7767	0.8190	0.7762
	RC	0.6169	0.6133	0.7530	0.7812	0.8207	0.7847
	F1	0.6285	0.6246	0.7427	0.7778	0.8181	0.7780
NB	AC	0.7639	0.7453	0.5692	0.5385	0.7731	0.7983
	PR	0.8125	0.7955	0.6094	0.5822	0.7762	0.8000
	RC	0.7582	0.7393	0.6330	0.6005	0.7767	0.8063
	F1	0.7707	0.7542	0.5535	0.5161	0.7762	0.8019
DT	AC	0.7083	0.6806	0.7692	0.7641	0.8067	0.7983
	PR	0.7609	0.7375	0.7766	0.7688	0.8063	0.7984
	RC	0.7085	0.6811	0.7739	0.7690	0.8325	0.8209
	F1	0.7172	0.6839	0.7748	0.7689	0.8116	0.8034
SVM	AC	0.7014	0.6828	0.8308	0.8256	0.8151	0.8235
	PR	0.7468	0.7252	0.8297	0.8229	0.8159	0.8222
	RC	0.6934	0.6716	0.8410	0.8383	0.8186	0.8317
	F1	0.7079	0.6852	0.8338	0.8275	0.8167	0.8255
ANN	AC	0.7569	0.7236	0.7231	0.7333	0.7731	0.8151
	PR	0.7773	0.7571	0.7287	0.7425	0.7794	0.8175
	RC	0.7463	0.7267	0.7259	0.7351	0.7703	0.8121
	F1	0.7588	0.7216	0.7272	0.7381	0.7727	0.8125
RF	AC	0.8472	0.8100	0.8410	0.8513	0.8067	0.8403
	PR	0.8602	0.8274	0.8430	0.8484	0.8079	0.8429
	RC	0.8374	0.8103	0.8487	0.8671	0.8105	0.8451
	F1	0.8472	0.8035	0.8444	0.8544	0.8080	0.8419
AdaBoost	AC	0.6944	0.6750	0.7538	0.7231	0.8151	0.8151
	PR	0.7370	0.7016	0.7640	0.7354	0.8175	0.8175
	RC	0.7023	0.6898	0.7573	0.7271	0.8133	0.8133
	F1	0.7077	0.6792	0.7602	0.7280	0.8136	0.8136

5. Conclusions

In this work, an endeavor is made towards finding the minimal set of features essential for predicting students' future performance using machine learning models. More than 27 research papers have been investigated to find out the most frequent features used in the private datasets used by contemporary works. Features of three public data repositories are also analyzed to find out the correlation and dependency among the features of these datasets. A novel algorithm for finding a minimal set of features from the fusion of frequent feature lists and statistically selected features list is also proposed. Six single ML models and two ensemble models are tested in predicting students' performance on three public datasets in two experimental setups- first with the complete feature set and then with the selected feature set. It is observed from the experimental results that, even with discarding more than half of the original features, the difference in performance accuracy in the two experimental setups is insignificant in most of the cases, whereas it is improved in almost one-third of the cases. Therefore, using this proposed methodology, an essential feature set can be extracted from any new dataset and used as a baseline to evaluate the impact of any additional feature to be included in the dataset. It would be helpful for future researchers to identify the correct attributes for preparing their dataset for predicting students' academic performance.

References

[1] Han, M., Tong, M., Chen, M., Liu, J., & Liu, C. (2017, July). Application of ensemble algorithm in students' performance prediction. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 735-740). IEEE.

[2] Anuradha, C., & Velmurugan, T. (2015). A comparative analysis on the evaluation of classification algorithms in the prediction of students performance. Indian Journal of Science and Technology, 8(15), 1-12.

[3] Ismail, L., Materwala, H., & Hennebelle, A. (2021, February). Comparative Analysis of Machine Learning Models for Students' Performance Prediction. In International Conference on Advances in Digital Science (pp. 149-160). Springer, Cham.

[4] Osmanbegović, E., Suljić, M., & Agić, H. (2014). Determining dominant factor for students performance prediction by using data mining classification algorithms. Tranzicija, 16(34), 147-158.

[5] Acharya, A., & Sinha, D. (2014). Early prediction of students performance using machine learning techniques. International Journal of Computer Applications, 107(1).

[6] Amra, I. A. A., & Maghari, A. Y. (2017, May). Students performance prediction using KNN and Naïve Bayesian. In 2017 8th International Conference on Information Technology (ICIT) (pp. 909-913). IEEE.

[7] Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. International Journal of computer applications, 36(11), 8-12.

[8] Jalota, C., & Agrawal, R. (2021). Feature selection algorithms and student academic performance: A study. In International Conference on Innovative Computing and Communications (pp. 317-328). Springer, Singapore.

[9] Devasia, T., Vinushree, T. P., & Hegde, V. (2016, March). Prediction of students performance using Educational Data Mining. In 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE) (pp. 91-95). IEEE.

[10] Bhardwaj, B. K., & Pal, S. (2012). Data Mining: A prediction for performance improvement using classification. arXiv preprint arXiv:1201.3418.

[11] Pandey, M., & Taruna, S. (2016). Towards the integration of multiple classifier pertaining to the student's performance prediction. Perspectives in Science, 8, 364-366.

[12] Abdullah, A. L., Malibari, A., & Alkhozae, M. (2014). STUDENTS PERFORMANCE PREDICTION SYSTEM USING MULTI AGENT DATA MINING TECHNIQUE. International Journal of Data Mining & Knowledge Management Process, 4(5), 1.

[13] Arsad, P. M., & Buniyamin, N. (2013, November). A neural network students' performance prediction model (NNSPPM). In 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA) (pp. 1-5). IEEE.

[14] Sharma, D., & Aggarwal, D. (2021). A Predictive Approach to Academic Performance Analysis of Students Based on Parental Influence. In International Conference on Innovative Computing and Communications (pp. 75-84). Springer, Singapore.

[15] Alshabandar, R., Hussain, A., Keight, R., & Khan, W. (2020, July). Students performance prediction in online courses using machine learning algorithms. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

[16] Ahmad, S., El-Affendi, M. A., Anwar, M. S., & Iqbal, R. (2022). Potential Future Directions in Optimization of Students' Performance Prediction System. Computational Intelligence and Neuroscience, 2022.

[17] Nabil, A., Seyam, M., & Abou-Elfetouh, A. (2021). Prediction of students' academic performance based on courses' grades using deep neural networks. IEEE Access, 9, 140731-140746.

[18] Shingari, I., Kumar, D., & Khetan, M. (2017). A review of applications of data mining techniques for prediction of students' performance in higher education. Journal of Statistics and Management Systems, 20(4), 713-722.

[19] Sarker, F., Tiropanis, T., & Davis, H. C. (2013). Students' performance prediction by using institutional internal and external open data sources.

[20] Raut, A. B., & Nichat, M. A. A. (2017). Students performance prediction using decision tree. International Journal of Computational Intelligence Research, 13(7), 1735-1741.

[21] Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student performance prediction using machine learning techniques. Education Sciences, 11(9), 552.

[22] Ketui, N., Wisomka, W., & Homjun, K. (2019). Using classification data mining techniques for students performance prediction. In 2019 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON) (pp. 359-363). IEEE.
[23] Saa, A. A. (2016). Educational data mining & students' performance prediction. International Journal of Advanced Computer Science and Applications, 7(5).
[24] Marbouti, F., Diefes-Dux, H. A., & Madhavan, K. (2016). Models for early prediction of at-risk students in a course using standards-based grading. *Computers & Education*, *103*, 1-15.
[25] Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447-459.
[26] Goga, M., Kuyoro, S., & Goga, N. (2015). A recommender for improving the student academic performance. *Procedia-Social and Behavioral Sciences*, *180*, 1481-1488.
[27] Miguéis, V. L., Freitas, A., Garcia, P. J., & Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, *115*, 36-51.
[28] Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. Knowledge-Based Systems, 23(6), 529-535.
[29] UCI Student Performance Data Set, *url: https://archive.ics.uci.edu/ml/machine-learning-databases/00320/*
[30] Kaggle Students' Academic Performance Dataset *url: https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data*
[31] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Authors' Profiles

Dr. Souvik Sengupta is currently an Associate Professor, Department of Computer Science and Engineering, Aliah University, Kolkata, India. He received his PhD (Tech) from the University of Calcutta in 2017. His research interests include Data Science, Machine Learning, Bio-medical image analysis, NLP, and Education Technology.

How to cite this paper: Souvik Sengupta, "Towards Finding a Minimal Set of Features for Predicting Students' Performance Using Educational Data Mining", International Journal of Modern Education and Computer Science(IJMECS), Vol.15, No.3, pp. 44-54, 2023. DOI:10.5815/ijmecs.2023.03.04