Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

DESIGN AND DEVELOPMENT OF HYBRID PRINCIPAL COMPONENT ANALYSIS (HPCA) ALGORITHM FOR ACADEMIC PERFORMANCE PREDICTION

Chitra Mehra [1] and Dr. Rashmi Agrawal [2]
[1] Assistant Professor, Manav Rachna International Institute of Research and Studies, Faridabad.
[2]Professor, Manav Rachna International Institute of Research and Studies, Faridabad.

ABSTRACT
Data mining and its applications are ubiquitous for business purposes since its beginning. Data mining techniques are used by many fields for knowledge discovery as well as for strategic decisions. However, in the present era, some new and emerging areas like education systems are also using data mining successfully to discover meaningful patterns from the pool of data. The primary focus of all academic institutions is the prediction of student's academic performance. To achieve this, educational data mining (EDM) is used. All over the world, educational data mining (EDM) is gaining popularity among the researchers because of its need and importance for the society. To handle the complexity of large volume of educational institutions data, various informative technologies are used. Machine learning is used by many researchers to mine knowledge from the educational database for the improvement in students and instructor’s performance. The most challenging task in prediction models is to select the efficient technique by which satisfactorily results can be produced. A hybrid algorithm of principal component analysis (HPCA) in conjunction with four machines learning (ML) algorithms: random forest (RF), support vector machine (SVM), naïve Bayes (NB) of Bayes network and C5.0 of decision tree (DT)is introduced in this paper so that there is always an improvement in the performances of classification. We evaluated our proposed model on three datasets taken from kaggle. In this paper, assessment metrics of the proposed model are classification accuracy, root mean square error (RSME), precision and recall. 10-fold cross-validation is also applied on these datasets for the evaluation of predictive performance. The proposed algorithm produced satisfactorily results of prediction which shows that HPCA is best for the optimal prediction method to get good result.

Keywords: Student performance; machine learning algorithms; k-fold cross-validation; principal component analysis; Support Vector Machine, Random Forest, Naïve bayes

1. INTRODUCTION
To form successful member of society, the quality of education must be improved. Data repository of educational institutions is very significant as it assists every

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

stakeholder of an educational process. By using this data, various hidden and interesting patterns can be extracted. Evaluation of student academic performance can be done by various machine learning techniques. Academic performance prediction is a promising area of research area in the field of educational data mining (EDM). Educational data mining (EDM) uses educational data and fundamental nature of data mining techniques to discover the student's performance issues. One of the main areas of educational data mining (EDM) is prediction. Predictive analysis is required for the academic growth of students. By making early prediction and classification of student performance level, academic performance of students as well as their managerial settings will be improved.

Prediction models are also useful to monitor the student academic progress. To analyze the academic performance, prediction models used various educational data mining (EDM) techniques. These prediction models depend on the important/selected features of the dataset. Therefore, we have to consider only those factors which significantly affect student's performance. Identification of factors for academic performance prediction is a research task of intricate nature. Academic institutions repository may have many inappropriate and superfluous attributes which may affect the result of prediction. In a dataset, attribute's redundancy can be minimized and their relevancy can be maximized by using feature selection methods with zero or no loss of vital data.

Feature selection algorithms are used to select most suitable features from a dataset. These algorithms always improve the prediction result. Its main aim is to select a subset by eradicating non predictive data. To boost up the predictive accuracy of model and to diminish the complications of learned results, feature selection algorithms are used. By using feature selection techniques, the quality of students' datasets has improved, and with it, the predictive accuracy of various data mining techniques has also enhanced [7]. With the help of feature selection algorithms, student performance prediction models and their efficiency can be increased. There are three groups of feature selection techniques: filter, wrapper and embedded models. Training data and its general characteristics decide the nature of filter method. This method is not dependent on a learning algorithm and is done in the preprocessing stage. Machine learning algorithms are used by wrapper method to assess the features of the dataset. Training process of classifiers is performed by using embedded methods and these are precise to some given machine learning algorithms.

Quality of data that agitate the feature selection algorithms is the reason for misclassification in the classification problem. In literature survey, many researchers have focused on dimensionality reduction i.e., feature extraction and feature selection methods for improvement in classification performance as well as in prediction. A hybrid principal component analysis (HPCA) algorithm is proposed in this paper and the same is applied on three datasets for the evaluation of predictive

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

performance of the models. The proposed algorithm i.e. classification model is adopted using four baseline machine learning algorithms along with 10-fold cross-validation and principal component analysis.

1.1 Related Work

In the field of education, various tasks are performed by tools and techniques of educational data mining (EDM). To identify relevant knowledge in educational datasets, various researchers' uses educational data mining (EDM) and its techniques. Basically, raw educational data is given as input for EDM methodology and then applied learning analytics, data mining and machine learning to get useful and meaningful information while performing comparative analysis of manual and conventional data. To solve prediction and classification problems, an effective predictive model can be prepared with the help of supervised machine learning techniques.

Mehra & Agrawal [1] discussed a group of methods which can handle a variety of data, their application areas and list of data mining tasks. Various objectives of data mining tasks and techniques which are helpful for many organizations to expand their understanding about data i.e., knowledge by making modifications in the defined operations and procedures, their profitability may rise.

Hussain, Doneva & Muhsi [2] analyzed the academic performance of the students using deep learning by developing a regression model. Linear regression along with regression using deep learning is applied on the dataset. Performance of deep learning model beats the linear regression model.

Altabrawee, Ali & Qaisar [3] discussed the consequences of internet usage as a learning resource and time spent by them on social networking sites for their academic performance. After this, features were taken into consideration that determine whether the internet used by student and the time used up by students on the social networks is for their learning or not. The two main parameters for models comparison were ROC index performance measure and the classification accuracy.

Mehra & Agrawal [4] appraised the problem of imbalanced classification on the Pima Indian Diabetes dataset using weka. A variety of filters i.e. Spread Subsample filter is used for majority class's under-sampling while Resample filter is used for random over-sampling of minority class were used for the balancing of class distribution in the preprocessing stage of the above-mentioned dataset, under weka. It was clearly mentioned in the result that by using sampling-based techniques all accuracy parameters such as Recall, F1-Score, Precision, and ROC area of minority class were improved.

Jalota & Agrawal [5] discussed the importance of feature selection for a classification problem. For this purpose, the author used two approaches of filter selection i.e. wrapper-based feature selection and correlation feature selection (CFS). It is clearly visible from results that highest accuracy with the correlation feature selection

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

algorithms is achieved by using SMO and J48 whereas highest accuracy with the wrapper subset feature selection algorithms to predict various grades i.e. low, medium and high grade for the students is measured by using Naïve Bayes.

Hasan & Abbas [6] proposed a supervised data classification model with the aim of predicting student academic performance at the end of the semester. For this, eight classification algorithms along with the activities performed by students in Moodle on campus and outside campus were also used.

Jalota & Agrawal [8] discussed five machine learning algorithms i.e. Support vector Machine, J48, Random Forest, Multilayer perceptron and Naïve Bayes along with statistical techniques for the enlightenment of academic performance and learning habits of students for the prediction of their performance level. It can be seen from the results that multilayer perceptron shows best performance as compared to other classifiers.

Shukla & Malviya [9] compared various clustering techniques and conclude the best technique for prediction. While comparing numerous clustering techniques to get the clusters of student instances, the expectation maximization (EM) technique is chosen.

Amjad, Mostafa & Khaled [10] extracted a new dataset from a student information system using Decision Tree, Random Forest, Gradient Boosted Trees, Deep Learning, Naïve bayes, Logistic Regression and Generalized Linear Model to predict the students' academic performance. It is clearly visible from the results that Random Forest proves best accuracy i.e. 98%.

Fatima & Mahgoub [11] provided a concise summary about the usage of all tools and techniques with respect to the educational data mining. To predict student's academic success, a simple framework with different variables was proposed where two different algorithms: Bayesian Network and Decision Trees were used to accomplish the above mentioned task. Results of comparative analysis of accuracy shows that Bayesian Network gives better accuracy and outperforms the Decision Trees.

Salal & Kumar [12] paid attention on two major factors for academic performance prediction:

- Most significant factors for student performance prediction used by the most researcher
- Most frequent algorithm used for student performance prediction

Hamoud & Humadi [13] implemented feature selection (FS) and presented an artificial neural network (ANN) based prediction model to predict the grade (Pass (P) or Failed (F)). To prove the optimality of model, performance of this model is compared with three previous models.

Borges, Esteves. & Holanda [14] proposed Principal Component Analysis (PCA)

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

based method for students' performance prediction and for relevant patterns identification which concerns with their characteristics. The proposed method allowed us to study the predictive capability of students' performances and the usefulness of Principal Component Analysis (PCA) for patterns interpretation in educational data.

Akinrotimi & Aremu [15] used C4.5 and Random tree algorithm for the prediction of students' academic performance on the basis of factors affecting their family background and their preceding academic achievements. Comparisons of these algorithms were made by using the holdout method. From the results, it was clearly visible that C4.5 produced the highest accuracy.

Roy & Garg [16] discussed the factors that can influence a student's academic performance using three machine learning classifiers i/e/ J48, Naïve bayes (NB) and Multilayer Perceptron (MLP) under WEKA.

The rest of the paper is organized as follows: Section II describes the Machine Learning Classifiers. Proposed algorithm and its investigational results are shown in Section III. Conclusion is given in Section IV.

2. MACHINE LEARNING CLASSIFIERS

Predefined classes for target variables can be predicted and classified with the help of a common technique of machine learning i.e. classification. Several machine learning classifiers have been used in this area. In our research, we have taken four most important algorithms i.e random forest, support vector machine, C5.0 and decision tree, are as base models.

2.1 Random Forest (RF)

This algorithm belongs to the supervised learning technique of machine learning. It is an ensemble technique by which both regression and classification problems can be solved with the help of multiple decision trees and bagging. In this algorithm multiple decision trees are used as base learning models. This algorithm picks up the performance of the model and put off the problem of overfitting.

It can be seen from fig.1 that predictive accuracy of the dataset can be improved by calculating the average through random forest algorithm after applying multiple decision trees on various subsets of the given dataset.

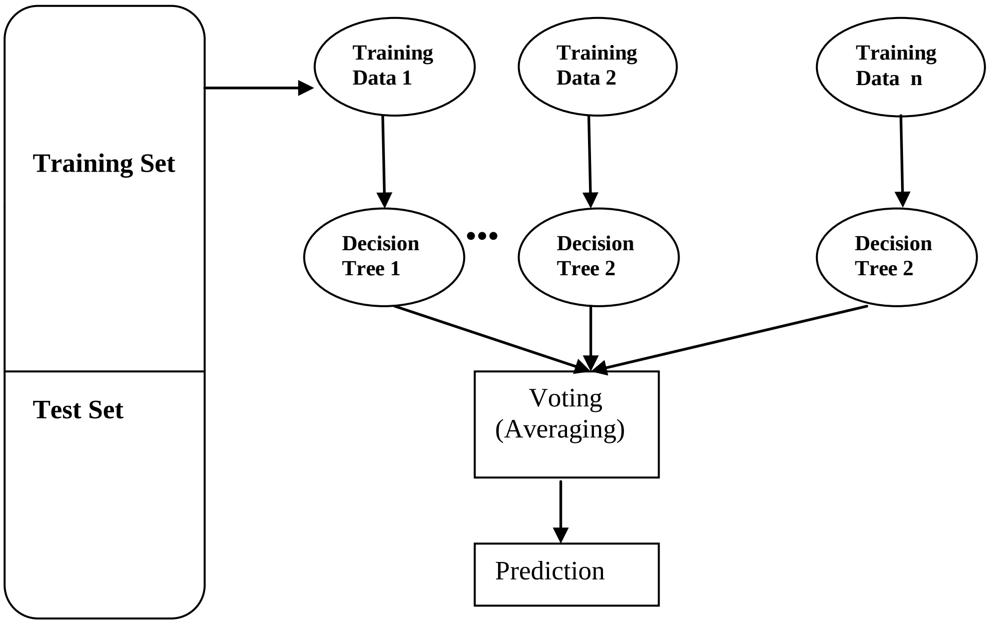

Figure 1. Random Forest Algorithm

Although all these above mentioned algorithms works satisfactorily with the datasets, but sometime due to over fitting, the accuracy of the model declines. To overcome this problem, irrelevant features are removed during the classification to improve accuracy. In real world, to handle the large volume of data more computational resources are required. Therefore, to improve accuracy and boost up the performance there should be a combination of dimensional reduction techniques with classification algorithms.

2.2 Support Vector Machine

It is a type of most accepted machine learning classifier belongs to the category of supervised learning. To classify between classes, concept of margin is used by this classifier. It is used for both types of problems i.e. classification and regression. Best line or decision boundary generation is the objective of the SVM algorithm so that n-dimensional space can be segregated into classes. Thus, there will be correct categorization of new data point in the future. So, this best line/decision boundary is called a hyper plane. hyperplane (line) can be found by the maximum distance between data points of both classes i.e. maximum margin.

To maximize the margins of hyperplanes, dissimilar kernel functions are used in SVM for the optimal solution. Sigmoid function, radial basis function and polynomial

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

function are some of the most popular kernels in support vector machine (SVM). For multi-classification problem, the most commonly used kernel is SVM with radial bias function (SVMRBF) kernel as it involves less number of parameters with respect to the polynomial kernel.

2.3 Naïve Bayes (NB)

It can be defined as an assortment of classification algorithms on the basis of Bayes' Theorem used for binary (two class) and multi classification problems. Naïve Bayes belongs to the family of 'probabilistic classifiers' with strong independence between features. In simple terms, a Naive Bayes classifier presupposes that there is no relation between features in a class i.e. a feature in a class is not related to the presence of any other feature.

For example, if a fruit is red, round, and about 3 inches in diameter, it may be considered as an apple yet these features can be:

a) Dependent on each other
b) Dependent upon the existence of the other features

It means all of these properties separately give weight age to the overall probability about the fact that this fruit is an apple and therefore it is known as 'Naive'.

For very large data sets, this model is easy particularly useful because it is easy to build. Highly sophisticated classification methods could also be performed with naive bayes.

Bayes theorem is used to calculate posterior probability as shown in the equation (1) given below:

$$P(a|b) = \frac{P(b|a)P(a)}{P(b)} \quad (1)$$

The preceding probability of class is represented by $P(a)$.

To characterize the subsequent probability of class (a, target) with the given predictor (b, attributes), we used $P(a|b)$.

The probability of predictor for a given class is $P(b|a)$

The prior probability of predictor is shown by $\mathrm{P(b)}$

2.4 Decision Tree (C5.0)

It is one of the well-known classification algorithm based on decision tree and a successor of C4.5 algorithm. To solve classification and regression problems, a supervised learning technique (decision tree) is used. In this classifier, features of a dataset can be represented by internal nodes, decision rules corresponds to the branches whereas outcome is show by each leaf node.

Selection of best attribute/feature is the main challenge for implementation of a decision tree. Best attribute for the nodes of the tree can be selected by attribute selection or ASM method. Attribute selection measure (ASM) can be calculated by:

a) Information gain or entropy
b) Gini Index

a) Information Gain I(G)

It is used to compute the information about a class provided by the feature. Splitting the node and building of decision tree will be done on the basis of significance of information gain. I(G) can be computed from the equation (2).

$$\mathrm{I(G)} = \mathit{Entropy}(s) - [(\mathit{Weighted\ Average}) * \mathit{Entropy}(\mathit{each\ faeture})] \quad (2)$$

Entropy: To compute the impurity of each attribute, entropy is used. It identifies arbitrariness in data. It can be measured using equation(3):

$$\mathbf{Entropy\ (s) = -p(yes)log2\ \ p(yes) - p(no)log2\ \ p(no)} \quad (3)$$

S= Total number of samples
P(yes)= possibility of yes
P(no)= possibility of no

b) Gini Index

Gini index is used to determine the impurity or purity at the time of decision tree creation. If an attribute is having low Gini index, it must be selected/preferred in attribute selection measure as compare to the attribute having high Gini index. Gini Index can be measured using equation (4):

$$\mathbf{Gini\ Index = 1 - \textstyle\sum_j P_j^2} \quad (4)$$

3. PROPOSED METHOD: Hybrid Principal Component Analysis (HPCA)

In this paper, we propose a novel classification algorithm **Hybrid Principal Component Analysis (HPCA)** using ensembling technique to get better performance of the above-mentioned machine learning algorithms. We performed 10 fold cross validation along with principal component analysis (PCA) in the base classifier. Principal component analysis (PCA) is a statistical method used to convert an original data set to a new dataset of with lesser number of attributes. It is one of the popular technique used for exploratory data analysis and predictive modeling and is also used to draw strong patterns from the given dataset by reducing the variances. The steps of proposed approach are as follows:

3.1 Steps:

Algorithm Hybrid principal Component Analysis Algorithm (HPCA)

Input: A d-dimensional training dataset T_0 with attributes (x) where x=1 to n and with target attribute Y

a) Scale training dataset T_0 to obtain new dataset T_i

$T_i = \frac{x-\mu}{\sigma}$

b) Calculate mean of attributes (μ)from T_0

$$\mu = \frac{1}{n}\sum_{x=1}^{n} x_i$$

c) Calculate standard deviation (σ)of attributes from T_0

$$\sigma = \frac{\sqrt{\sum(x_i - \mu)^2}}{N}$$

d) Build a covariance matrix (A) of each attribute (x) and its subsequent attribute (y) from T_i

$$Cov(x, y) = \sum \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{N - 1}$$

e) Compute eigenvalues and eigenvectors for T_i .

i) For a covariance matrix (A)

$$AV = \lambda V$$

Where V is a non zero vector and λ is a scalar.

ii) From (i)

$$AV - \lambda V = 0$$

$$\therefore V(A - \lambda) = 0$$

iii) To satisfy (i) and (ii)

$$\det(A - \lambda) = 0$$

f) Sort eigenvector with their corresponding eigenvalues in decreasing order to get orthonormal eigenvectors.

$$\lambda_{1>}\lambda_{2>\cdots\lambda_n} > 0$$

g) Find the average of eigenvalues for k^{th} attribute

$$\lambda_k = \frac{\lambda_1 + \lambda_n}{2}$$

h) Select k eigenvectors to create a feature matrix (f_m) of k features in dataset T_i

$$T_i = f_m * \sum_{1}^{k} v$$

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

i) Apply for each classifier (Random Forest, Support Vector Machine, Nayes Bayes, C5.0) -

i) Input Ti to classifier to get intermediate output (I_o)

ii) Apply 10 fold cross validation to get final output (O).

4. RESULTS AND DISCUSSION

The proposed method is equated with the existing versions to verify its efficacy and proficiency. For performance evaluation, the proposed method i.e .hybrid principal component analysis (HPCA) is conducted on 3 datasets of academic domain.

1) xAPI-Edu-Dataset: It is a dataset of educational domain gathered from Kalboard 360, a multi-agent learning management system (LMS). It was designed to assist knowledge by using leading-edge technology. It contains 16 attributes with 480 instances.

2) Student Grade Prediction Dataset: This dataset approach student achievement of two Portuguese schools in secondary education. There are two datasets for the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). It contains 33 attributes with 649 instances.

3) Students Performance Dataset: This dataset consists the student's marks in various subjects, which is available through Kaggle. It contains 8 attributes with 10000 instances.

All of the three datasets obtained from kaggle. We evaluated the performance of our algorithm using the following evaluation metrics:

i) Accuracy: It is used to quantify the percentage of correct predictions for the test data. Accuracy of our algorithm can be calculated as:

$$Accuracy = \frac{correctly\ predicted\ class}{total\ testing\ class} * 100\%$$

ii) Root Mean Square Error(RMSE): For evaluating the quality of predictions, we used Root Mean Square Error(RMSE). By using RMSE, it is easy to measure that how prediction is close to the performance level of students. The RMSE can be computed as:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n}(Predicted_i - Actual_i)^2}}{N}$$

iii) Precision: It measures the count of positive class predictions that really belongs to the positive class.

$$Precision = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

iv) Recall: It is used to measure the accuracy of proposed model for relevant data identification. Another term used for recall is sensitivity or True Positive Rate

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

In order to evaluate the effectiveness of hybrid principal component analysis algorithm (HPCA), we first applied base classifiers (Random Forest, Support Vector Machine, Naïve Bayes and C5.0) on 3 above-mentioned datasets.

We also applied 10-fold cross validation on (RF, SVM, NB and C5.0) and later we compare the results with HPCA algorithm. The performance of base classifiers, base classifiers+10-fold cross validation and HPCA on first dataset has been shown in Table 3.

TABLE 3: Performance of base classifiers, base classifiers+ 10-fold cross validation and HPCA on xAPI-Edu-Data.csv dataset

Classifiers	Accuracy	RMSE	Precision	Recall
RF	69.9	0.431	71.3	69.9
RF+ 10 CV	76.6	0.143	76.6	76.7
HPCA+RF	98.92	0.056	97.5	97.8
SVM	75.4	0.516	76.1	75.5
SVM+10 CV	77	0.496	78.8	78.8
HPCA+SVM	78.7	0.414	79.1	79.3
NB	64.4	1.191	64.1	64.4
NB+10 CV	67.7	0.645	67.5	67.7
HPCA+NB	86.2	0.521	86.6	86.1
C5.0	73.6	0.487	73.8	73.6
C5.0+10 CV	75.8	0.185	76.0	75.8
HPCA+C5.0	98.32	0.067	98.1	97.9

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

TABLE 4: Performance of base classifiers, base classifiers+ 10-fold cross validation and HPCA on STUDENT-MAT.CSV DATASET

Classifiers	Accuracy	RMSE	Precision	Recall
RF	90.37	0.321	90.0	89.9
RF+ 10 CV	96.58	0.147	96.6	96.7
HPCA+RF	98.09	0.057	97.9	98.0
SVM	75.52	0.489	75.3	74.8
SVM+10 CV	94.15	0.274	93.8	93.8
HPCA+SVM	96.32	0.182	96.3	96.1
NB	67.68	0.664	67.4	67.4
NB+10 CV	76.47	0.498	75.8	76.2
HPCA+NB	91.42	0.321	90.66	91.1
C5.0	86.18	0.372	85.9	85.8
C5.0+10 CV	95.6	0.174	95.2	95
HPCA+C5.0	98.62	0.067	98.4	98.1

TABLE 5: Performance of base classifiers, base classifiers+ 10-fold cross validation and HPCA on STUDENTS PERFORMANCE IN EXAM DATASET

Classifiers	Accuracy	RMSE	Precision	Recall
RF	89.23	0.561	88.9	88.9
RF+ 10 CV	98.22	0.113	98.1	98.0
HPCA+RF	99.72	0.041	99.7	99.5
SVM	86.44	0.823	86.4	86.3
SVM+10 CV	90.67	0.678	90.4	90.5
HPCA+SVM	97.01	0.178	96.8	96.9
NB	65.02	1.016	65	65.1
NB+10 CV	94.82	0.154	94.61	94.8
HPCA+NB	98.94	0.145	98.8	98.8
C5.0	76.55	0.845	76.4	76.5
C5.0+10 CV	97.54	0.114	97.4	97.4
HPCA+C5.0	99.25	0.073	99	99.1

Performance of the proposed model on three datasets (xAPI-Edu-Dataset, Student-mat.csv and StudentsPerformnce.csv) is shown in table 3,4 and 5 and their corresponding visualization is given in fig. 3,4 and 5. It is clearly visible from fig. 3

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

that our proposed hybrid model boosts the accuracy of RF from 69.9% to 98.92%, SVM from 75.4% to 78.7%, NB from 64.4% to 86.2% and C5.0 from 73.6% to 98.32% on XAPI-Edu-Dataset.

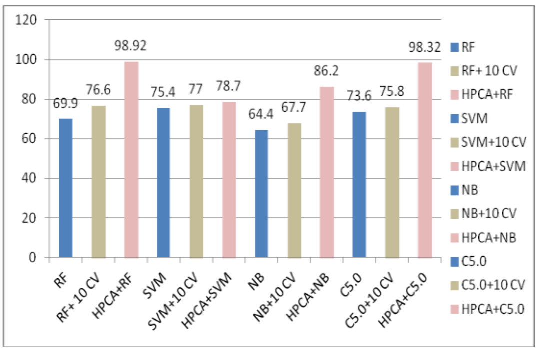

Fig. 3 Performance-based on the accuracy of xAPI-Edu-Dataset

In Fig.4 the hybrid models improved RF, SVM, NB and C5.0, and with accuracies of 8%. 21%, 24% and 12% respectively.

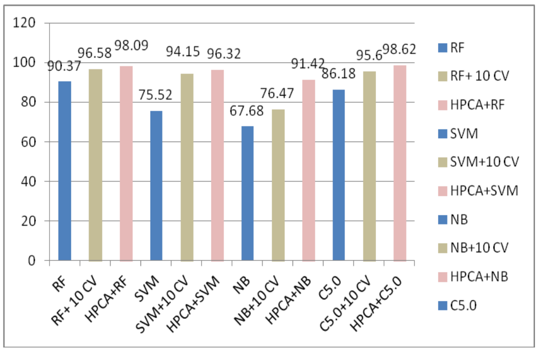

Fig. 4 Performance-based on the accuracy of Student-mat Dataset

In Fig. 5, the classification accuracy is improved i.e. from 86.44% to 97.01% by using proposed hybrid SVM. There is an improvement in the accuracy i.e. 33% by classification through hybrid naïve bayes rather than baseline naïve bayes. There is an improvement in the accuracies of Random Forest and C5.0 and Random Forest i.e. 99.72% and 99.25% respectively.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

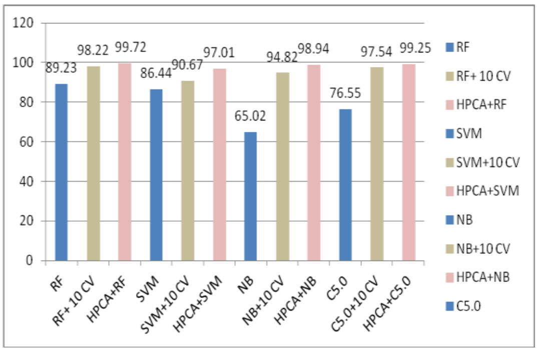

Fig. 5 Performance-based on the accuracy of StudentsPerformance.csv

It is clearly visible from fig. 3,4 and 5 that accuracy of baseline models i.e. (Support Vector Machine, Naïve Bayes, C 5.0 and Random Forest) is much improved by using hybrid principal component analysis algorithm (HPCA). The proposed algorithm Hybrid principal Component Analysis (HPCA) boosts up the classification accuracy with the noticeable difference. Thus to solve prediction and classification problems, the proposed hybrid principal component analysis (HPCA) can be considered as a best possible prediction model.

5. CONCLUSION

In this paper, we have used machine learning classifiers (Random Forest, Support Vector Machine, Naïve bayes and C5.0) to predict student performance on three datasets. Evaluation of performance is done using 3 steps in which performance of only baseline/machine is analyzed in the first step. In the second step, we applied 10 fold cross validation with these classifiers to improve the performance of these baseline classifiers. Lastly, we applied HPCA on baseline classifiers using 10 fold cross validation to improve the classification accuracy. On the basis of accuracy, computed using RMSE, precision and recall it shows that the proposed algorithm i.e. hybrid principal component analysis (HPCA) produced the satisfactorily results. In wrapping up, by combining the baseline models 10-fold cross-validation and hybrid principal component analysis (HPCA), the proposed algorithm computes high performance for prediction and classification problem which reveals that it is a prospective algorithm.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

REFERENCES

[1] Mehra C. & Agrawal R, " Educational data mining approaches, challenges and goals: A Review", *International Journal of Information, Communication and Computation Technology,* 2021; 8(2): 442-447.

[2] Hussain S., Doneva R., & Muhsi Z.Regression. "Analysis of student academic performance using Deep Learning", *Springer Journal of education and information technologies,* 2021; 26(2): 783-798.

[3] Altabrawee H., Ali O. & Qaisar S., "Predicting Students' Performance Using Machine Learning Techniques", *Journal of University of Babylon for Pure and Applied Sciences*, 2021; 27(1): 194-205.

[4] Mehra C. & Agrawal R., "Comparative study of resampling techniques of imbalanced dataset", *International Journal of Advance Science and Technology*, 2020; 29(3): 12699-12710.

[5] Jalota, C. & Agrawal R., "Feature Selection Algorithms and Student Academic Performance: A Study", in *International conference on innovative computing and communication (ICICC-2020),* 2020, pp. 317-328.

[6] Hasan R. & Abbas A., "Predicting Student Performance in Higher Educational Institutions Using Video Learning Analytics and Data Mining Techniques", *Journal of Applied Sciences*, 2020; 10(11): 3894-3904.

[7] Orsango S. & Chakraborty S., "Feature Selection Algorithms For Predicting Students Academic Performance Using Data Mining Techniques", *International Journal of Scientific and Technology Research,* 2020; 9(4): 1-5.

[8] Jalota C., & Agrawal R., "Analysis of data mining using classification", in *International conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCON-2019),* 2019, pp. 247-253.

[9] Shukla C. & Malviya A., "Modified Classification and Prediction Model for Improving Accuracy of Student Placement Prediction", in *International Conference on Advanced Computing and Software Engineering (ICACSE) 2019.*

[10] Amjad A.S., Mostafa Al & Khaled S., "Mining Student Information System Records to Predict Students Academic Performance", in *International Conference on advance machine learning technologies and applications*, 2019, pp.229-239.

[11] Fatima S. & Mahgoub S., "Predicting Student's Performance in Education using Data Mining Techniques", *International Journal of Computer Applications*, 2019; 177(19): 14-20.

[12] Salal Y. & Kumar M., "Systematic Review of Predicting Student's Performance in Academics", *International Journal of Engineering and Advanced Technology*, 2019; 8(3): 64-73.

ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:06:2021
DOI 10.17605/OSF.IO/SZUAD

[13] Hamoud A. & Humadi A., “Student’s success prediction model based on artificial neural network and a combination of feature selection methods”, *Journal of southwest jiaotong University*, 2019; 54(3): 1-19

[14] Borges R.P., Esteves L. & Holanda M., “Using Principal Component Analysis to support students' performance prediction and data analysis”, *Brazilian Symposium on Computers in Education,* 2018.

[15] Akinrotimi A.O, & Aremu D.R., “Student Performance Prediction using Random tree and C4.5”, *Journal of Digital Innovation, Contemporary Research, Engineering and Technology*, 2018;6(3): 23-34.

[16] Roy S. & Garg A., “Predicting academic performance of students using classification techniques”, in *4th IEEE Uttar Pradesh Section International Conference on Electrical,Computer and Electronics (UPCON)*, 2018.