International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Feature Snatching and Performance Assessment for Connoting the Admittance Likelihood of student using Principal Component Reduction

M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna

Abstract*: Recently, engineers are concentrating on designing an effective prediction model for finding the rate of student admission in order to raise the educational growth of the nation. The method to predict the student admission towards the higher education is a challenging task for any educational organization. There is a high visibility of crisis towards admission in the higher education. The admission rate of the student is the major risk to the educational society in the world. The student admission greatly affects the economic, social, academic, profit and cultural growth of the nation. The student admission rate also depends on the admission procedures and policies of the educational institutions. The chance of student admission also depends on the feedback given by all the stake holders of the educational sectors. The forecasting of the student admission is a major task for any educational institution to protect the profit and wealth of the organization. This paper attempts to analyze the performance of the student admission prediction by using machine learning dimensionality reduction algorithms. The Admission Predict dataset from Kaggle machine learning Repository is used for prediction analysis and the features are reduced by feature reduction methods. The prediction of the chance of Admit is achieved in four ways. Firstly, the correlation between each of the dataset attributes are found and depicted as a histogram. Secondly, the top most high correlated features are identified which are directly contributing to the prediction of chance of admit. Thirdly, the Admission Predict dataset is subjected to dimensionality reduction methods like principal component analysis (PCA), Sparse PCA, Incremental PCA , Kernel PCA and Mini Batch Sparse PCA. Fourth, the optimized dimensionality reduced dataset is then executed to analyze and compare the mean squared error, Mean Absolute Error and R2 Score of each method. The implementation is done by python in Anaconda Spyder Navigator Integrated Development Environment. Experimental Result shows that the CGPA, GRE Score and TOEFL Score are highly correlated features in predicting the chance of admit. The execution of performance analysis shows that Incremental PCA have achieved the effective prediction of chance of admit with minimum MSE of 0.09, MAE of 0.24 and reasonable R2 Score of 0.26.

***Index Terms*: Machine Learning, Dimensionality Reduction, MSE, MAE, R2 Score, PCA, Sparse PCA, Incremental PCA, Kernel PCA and Mini Batch Sparse PCA.**

I. INTRODUCTION

In machine learning, the prediction of the target variable can be done by applying dimensionality reduction. The optimization of the features results in reduction of the storage and execution time. The final prediction output depends on the availability of the reduced features in the dataset. So the reduced features should be less correlated for the prediction of the target variable. If the variables in the dataset are highly correlated to each other, then it results in duplication of the features. This results on high memory occupation and high response time. The difficulty of prediction increases with the increase in the number of features in the dataset. This significantly raises the need of dimensionality reduction algorithms.

The paper is organized in which the literature survey is dealt with Section 2 followed by the preliminaries in the Section 3. Proposed work is discussed in Section 4 followed by the implementation and Performance Analysis in Section 5. The paper is concluded with Section 6.

II. RELATED WORK

A. Literature Survey

The various admission methods of the students based on the stress level is explored [1]. It also considers the psychology nature of the students for predicting the admission. They also discuss that the student stress level was floating between neuroticism and conscientiousness. This procedure was carried out for the students of the psychology program at the University of Southern Denmark.

An attempt is made to analyze the solution for the strength of the colleges and universities all around the country [2]. They projected that the students accepting to continue the studies in colleges and universities is the major factor for the growth of the academic and social life of the world. They experiment machine learning for the students of Davidson College to predict the decision of students admission and found the accuracy to be 86%.

Revised Manuscript Received on 30 July 2019.

***** Correspondence Author

M. Shyamala Devi*, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Rincy Merlin Mathew, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.

R. Suguna, Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Making the high student admission in the university is really an important challenge for the management [3]. They also projected that the student admission is tough due to the increase in the competition in the market and profit nature of the company.

They provide an intelligent model to predict the student admission. They implemented Feature selection [3] to model the essential features that are needed to identify the student admission.

The admission system of the college greatly depends on the admission policies and procedures [4]. A method is designed to increase the admission rate by analyzing the relationship between the admission policies and retention rate of the organization. A predictive model is built to increase the admission yield of the university.

A statistical machine learning model namely GRADE is built for the computer science department of University of Texas at Austin [5]. This GRADE method utilizes the historical admissions data to predict the admission of the new student. It performs effective review process on the decision taken by the students towards admission.

The probability model is designed to predict the college admission and student enrolment [6]. The first method is to find the acceptance probability towards selecting the particular institution. The second method of predicting the admission depends on academic and non-academic feature of the institution. Then the probability of the student admission can be done based on the enrolment.

The admission rate is predicted by using artificial intelligence by building two stage expert system [7]. The first stage verifies the admission policies and procedures. The second stage depicts the number of degree completed students of the institution. A system is built with optimal rule induction that depends on the academic and non-academic background of the organization.

The lack of business intelligence, adaptation of emerging technology greatly affects the admission rate of the educational organization [8]. The grade point average of the students is found by using decision tree analysis, neural network analysis and multiple regression analysis. The model is evaluated based on the average square error of the student admission.

The supervised learning methods are used to predict the decision making process of a student towards selecting the organization [9]. This problem is a binary classification problem where the different classifiers are implemented to predict the rate of student admission. They found that logistic regression classifier is found to have an accuracy of 76.9% to predict the admission offer of a student.

The prediction of the student dropout is done by using data mining methods such as K- Nearest neighbor, Naïve Bayes, Decision Tree and Neural Network. They also used genetic algorithm for analyzing the prediction rate of the student admission [10].

A review on the prediction of student dropout is done through data mining methods. It deals with the aspects considered for the prediction of student admission in the university [11]. The machine learning feature selection and extraction methods can be used for the prediction of any factor in different application can be learnt through this article [12]–[16].

III. DIMENSIONALITY REDUCTION

The minimization of high data features into low data features is done by Dimensionality reduction process. It also makes sure that the dimensionality reduction does not result in loss of information in the dataset. Let us have a look on some of the dimensionality reduction methods below.

A. Principal Component Analysis

The selection of the principal components from the data set is done in this method and the steps of PCA are as follows.

(i) Develop the data set Covariance matrix
(ii) Find the Eigen vectors of Covariance matrix
(iii) Redesign the dataset with Eigen vectors having large Eigen values
(iv) Selecting Principal components with the features having high variance.

B. Kernel PCA

The non linear selection of the principal components from the data set is done in this method and the steps of Kernel PCA are as follows.

(i) Preferring the Kernel plot
(ii) Creating the kernel matrix K of training data set
(iii) Develop the data set Covariance matrix
(iv) Find the Eigen vectors of Covariance matrix
(v) Redesign the dataset with Eigen vectors having large Eigen values
(vi) Selecting Principal components with the features having high variance

C. Sparse PCA

The selection of the principal components is done through non linear selection having maximum variance and sparsity from the data set is done in this method and the steps of sparse PCA are as follows.

(i) Develop Sparse matrix of training data set
(ii) Carry out minima problem for Sparse matrix.
(iii) Develop the data set Covariance matrix
(iv) Find the Eigen vectors of Covariance matrix
(v) Redesign the dataset with Eigen vectors having large Eigen values Step
(vi) Perform singular value decomposition for Sparse matrix
(vii) Select eigen vector component with high variance and sparsity as principal components.

D. Mini Batch Sparse PCA

The selection of the principal components is done through non linear selection of multiple batches having maximum variance and sparsity from the data set is done in this method and the steps of mini batch sparse PCA are as follows.

(i) Partitioning of data set into ‘m multiple batches
(ii) Develop Sparse matrix for ‘m multiple batches
(iii) Develop the data set Covariance matrix
(iv) Find the Eigen vectors of Covariance matrix
(v) Carry out minima problem for Sparse matrix.

(vi) Redesign the dataset with Eigen vectors having large Eigen values
(vii) Perform singular value decomposition for Sparse matrix
(viii) Select eigen vector component with high variance and sparsity as principal components.

Incremental PCA

The selection of the principal components is done through non linear selection having less rank from the data set is done in this method and the steps incremental PCA are as follows.

(i) Develop Rank matrix of training data set
(ii) Develop the data set Covariance matrix
(iii) Find the Eigen vectors of Covariance matrix
(iv) Execute rank matrix with low rank approximation
(ix) Redesign the dataset with Eigen vectors having large Eigen values
(x) Perform singular value decomposition for rank matrix
(v) Select eigen vector component with high variance and low rank as principal components.

IV. PROPOSED WORK

In our proposed work, machine learning algorithms are used to predict the chance of admit of the student. Our contribution of predicting the chance of admit is achieved in four ways.

(i) Firstly, the correlation between each of the dataset attributes are found and depicted as a histogram correlation matrix.
(ii) Secondly, the top most high correlated features are identified which are directly contributing to the prediction of chance of admit.
(iii) Thirdly, the Admission Predict dataset is subjected to dimensionality reduction methods like principal component analysis (PCA), Sparse PCA, Incremental PCA , Kernel PCA and Mini Batch Sparse PCA.
(iv) Fourth, the optimized dimensionality reduced dataset is then executed to analyze and compare the mean squared error, Mean Absolute Error and R2 Score of each method.

A. System Architecture

The architecture of proposed system is shown in Fig. 1

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Chance of Admit Prediction for Feature Extraction

The Admission Predict dataset from Kaggle Repository is used for implementation with 7 independent attribute and 1 Chance of Admit dependent attribute and they are,

1) GRE Scores (out of 340)
2) TOEFL Scores (out of 120)
3) University Rating (out of 5)
4) Statement of Purpose (out of 5)
5) Letter of Recommendation Strength (out of 5)
6) Undergraduate GPA (out of 10)
7) Research Experience (either 0 or 1)
8) Chance of Admit (0 0r 1) - Dependent Attribute

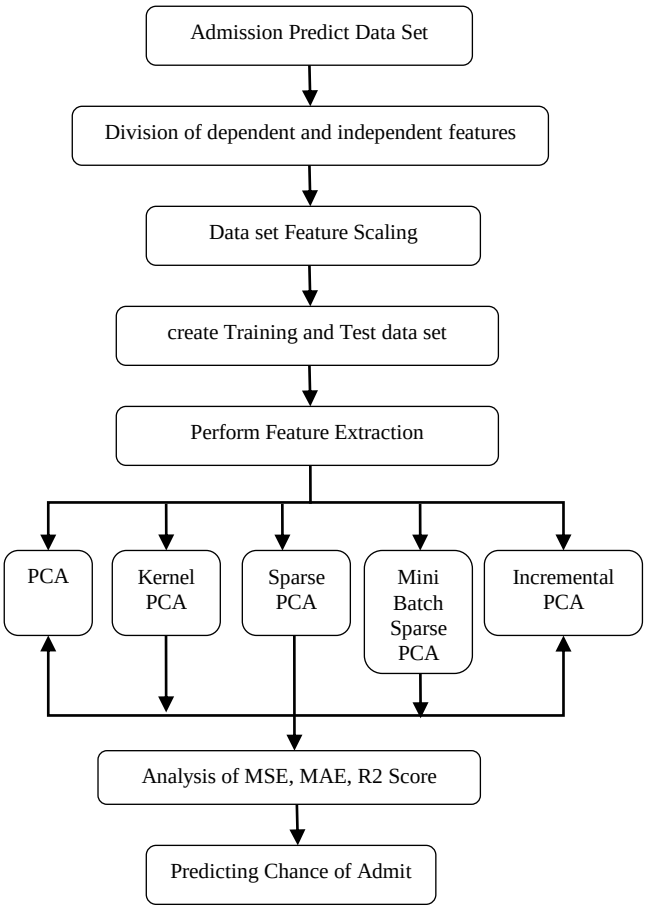

Fig. 1 System Architecture

The admission predict data set is subjected to find the correlation between each variable in the data set. The implementation is done in python scripts and the obtained correlation matrix of the dataset is shown in Fig. 2.

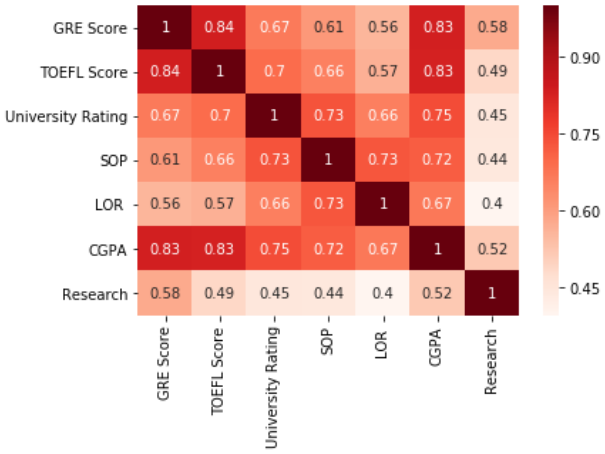

Fig. 2 Correlation matrix of admission predict dataset.

Admission predict dataset is implemented to form the histogram with its features and is shown in Fig. 3. The obtained feature importance of the variable of the Admission predict dataset is shown in Fig. 4 – Fig. 7

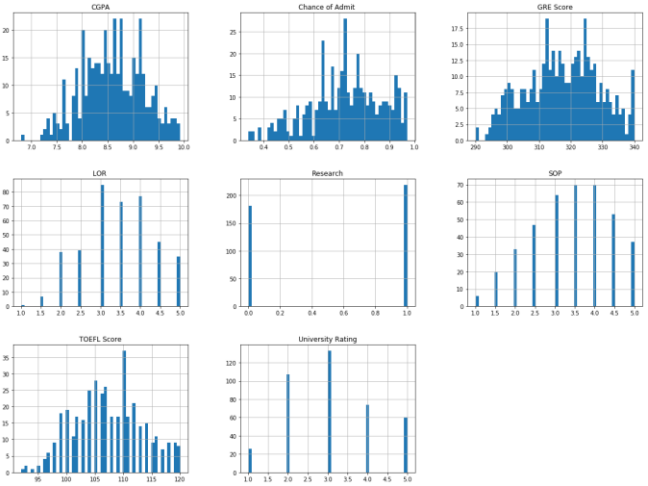

Fig. 3 Distribution of attributes in the dataset

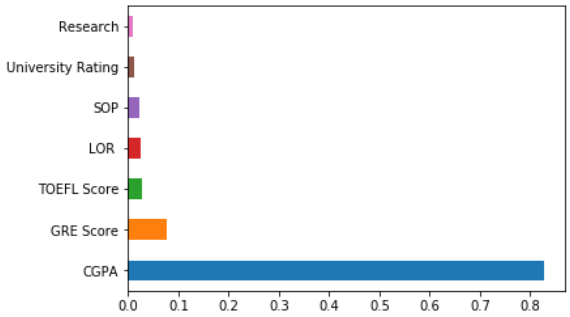

Fig. 4 Feature importance of each variable

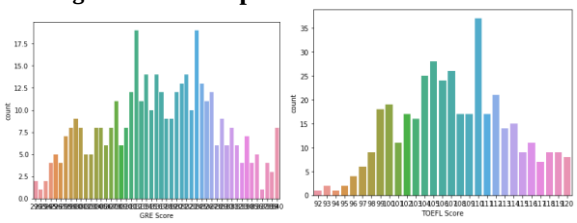

Fig. 5 Distribution of GRE and TOEFL Score

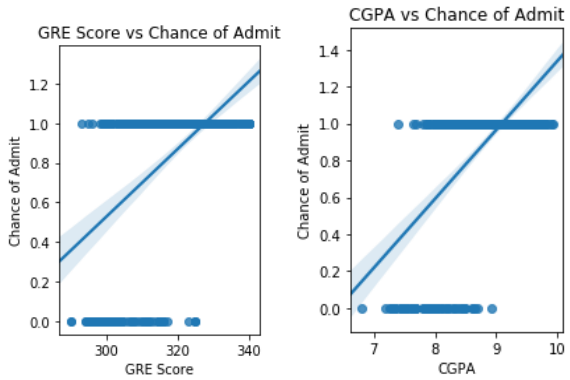

Fig. 6 GRE and CGPAScore VS Chance of Admit

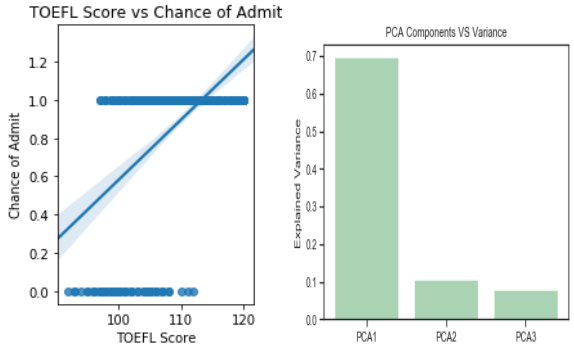

Fig. 7 TOEFL VS Chance of Admit, 3 Principal Components

Admission predict dataset is subjected to each PCA methods with 3 components and is shown in Fig. 8 - Fig 12.

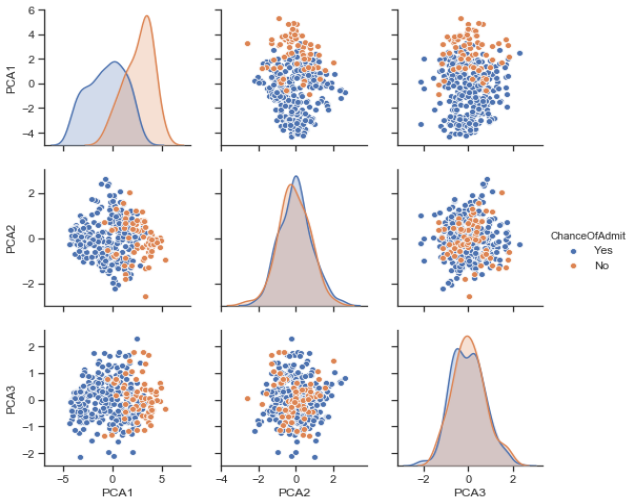

Fig. 8 Admit Chance of PCA with three Components

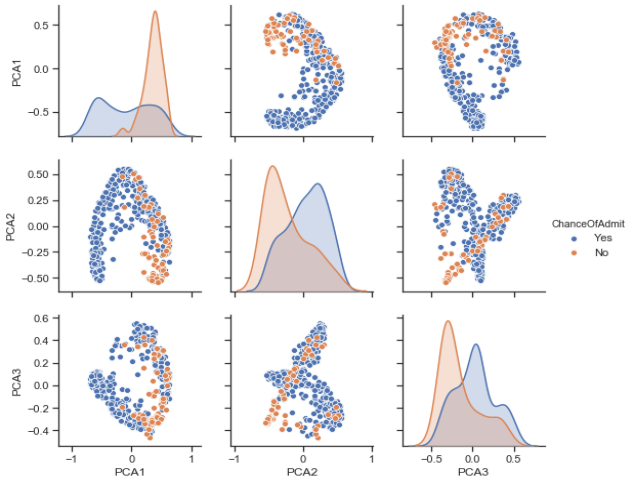

Fig. 9Admit Chance of Kernel PCA with three Components

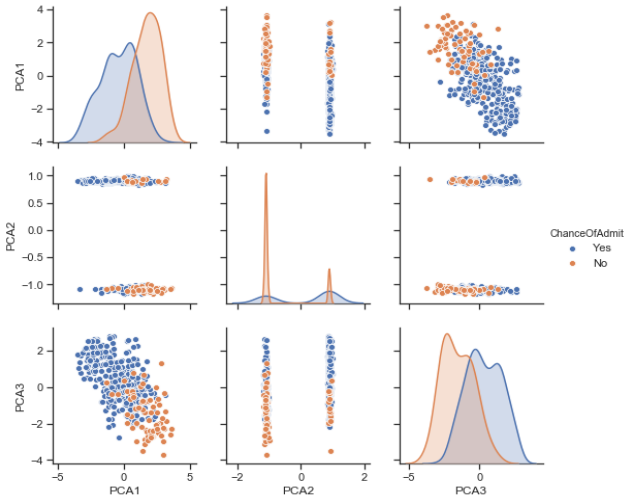

Fig. 10 Admit Chance of Sparse PCA with three Components

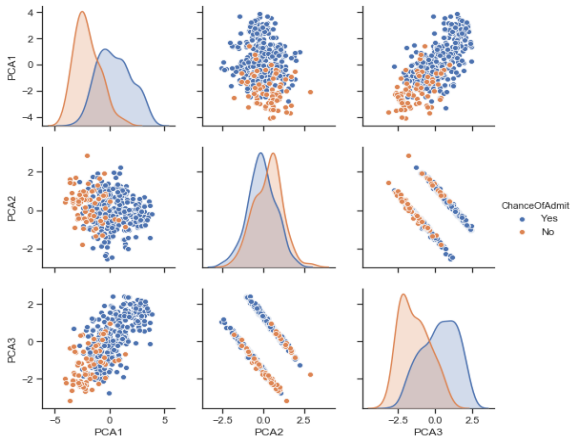

Fig. 11 Admit Chance of MiniBatch PCA with three Components

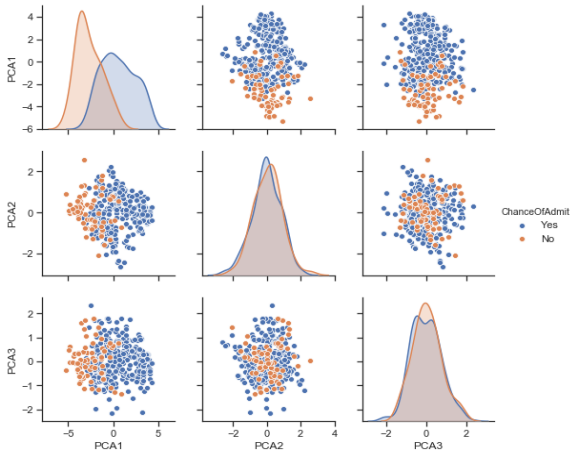

Fig. 12 Admit Chance of Incremental PCA with 3 Components

Admission predict dataset is subjected to each PCA methods with 2 components and is shown in Fig. 13 - Fig 18.

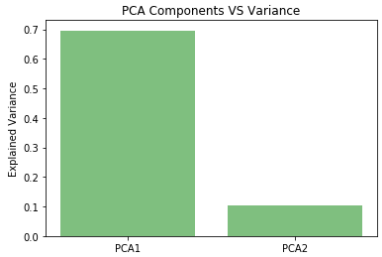

Fig. 13 Feature extraction with two components

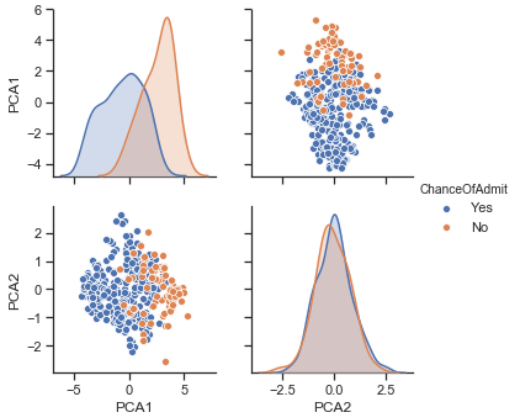

Fig. 14 Admit Chance of PCA with two Components

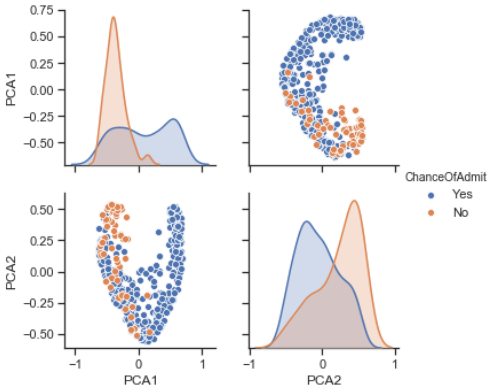

Fig. 15 Admit Chance of Kernel PCA with two Components

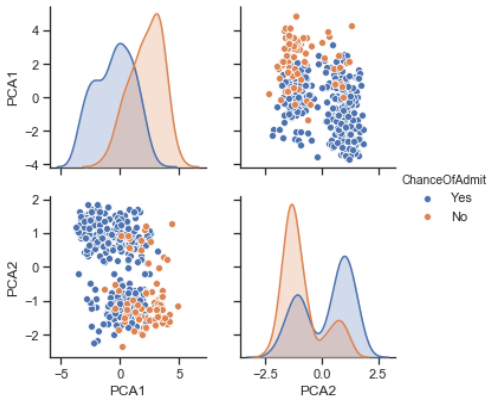

Fig. 16 Admit Chance of Sparse PCA with two Components

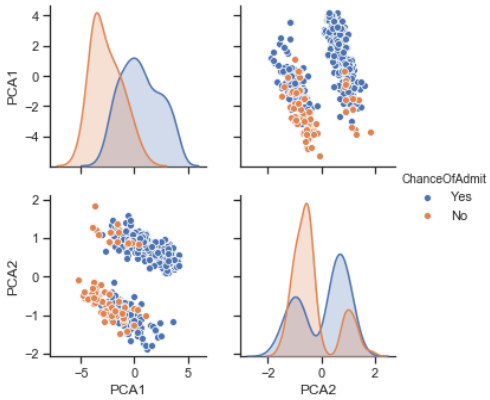

Fig. 17 Admit Chance of Mini Batch PCA with two Components

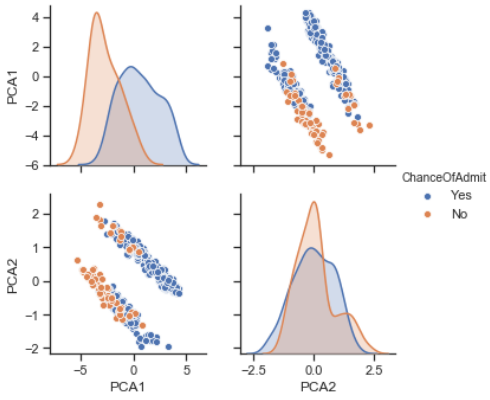

Fig. 18 Admit Chance of Incremental PCA with two Components

B. Performance Analysis of PCA

The mean squared error, mean absolute error and R2 score of each method is compared to analyze the performance and is shown in Table. 1 and Table. 2.

Table. 1. Comparison of Evaluation Parameters for PCA with three Components

Admission Predict Dataset	Number of Components = 3		
Feature Extraction Methods	Mean Squared Error	Mean Absolute Error	R2 Score
PCA	0.10	0.26	0.26
Kernel PCA	0.11	0.22	0.27
Sparse PCA	0.11	0.26	0.26
Mini Batch Sparse PCA	0.10	0.26	0.27
Incremental PCA	0.09	0.24	0.26

Table. 1. Comparison of Evaluation Parameters for PCA with two Components

Admission Predict Dataset	Number of Components = 2		
Feature Extraction Methods	Mean Squared Error	Mean Absolute Error	R2 Score
PCA	0.11	0.27	0.28
Kernel PCA	0.11	0.26	0.22
Sparse PCA	0.11	0.26	0.27
Mini Batch Sparse PCA	0.10	0.26	0.27
Incremental PCA	0.10	0.27	0.27

The two component PCA dimensionality reduction of the training and test dataset of each method is shown below in Fig. 19 - Fig 18.

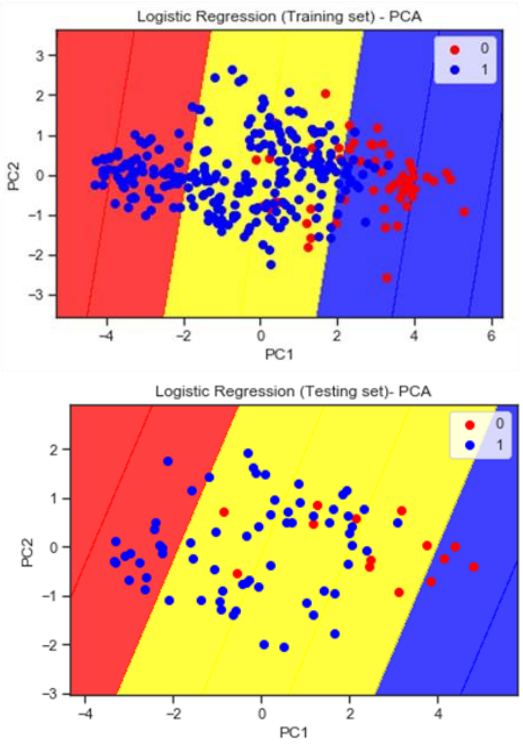

Fig. 19 Logistic Regression for PCA with two Components

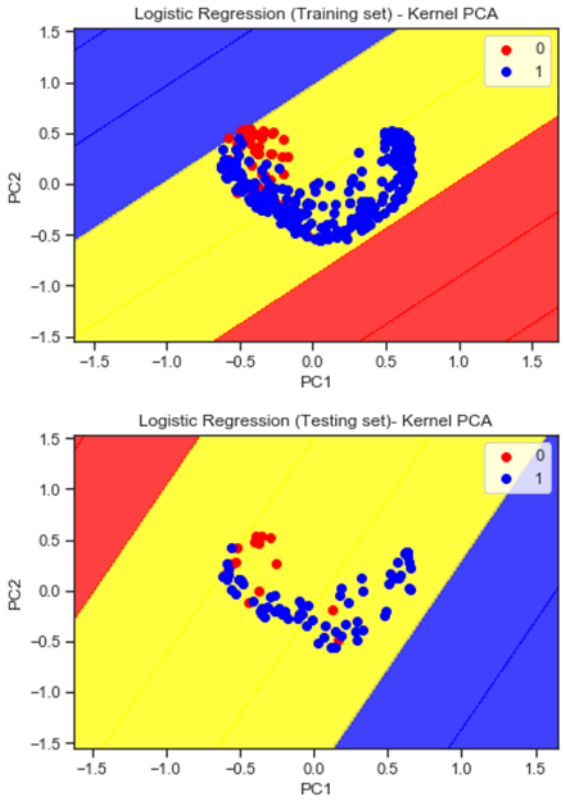

Fig.20 Logistic Regression for Kernel PCA with two Components

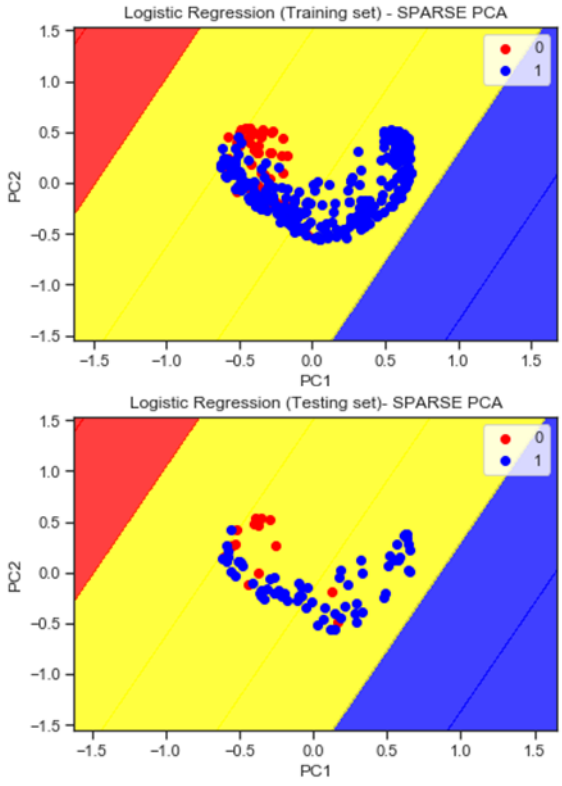

Fig. 21 Logistic Regression for Sparse PCA with two Components

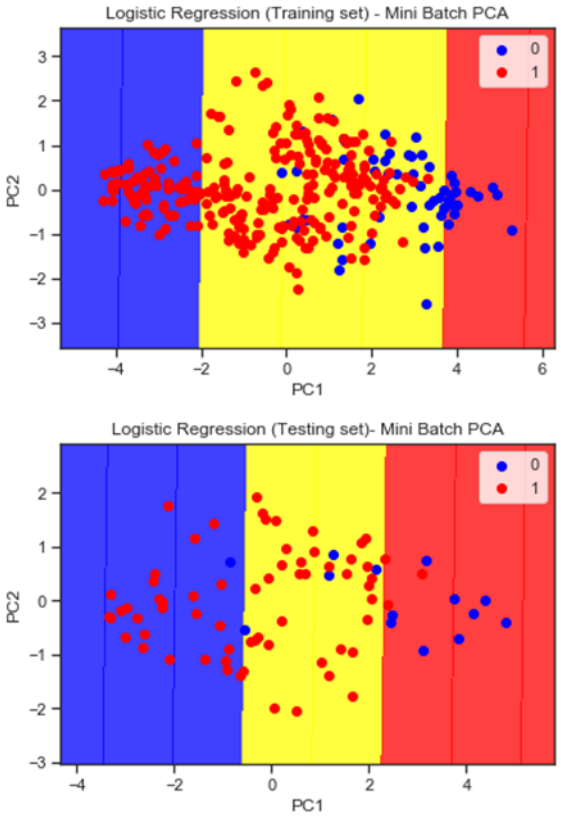

Fig. 22 Logistic Regression for Mini Batch Sparse PCA with two Components

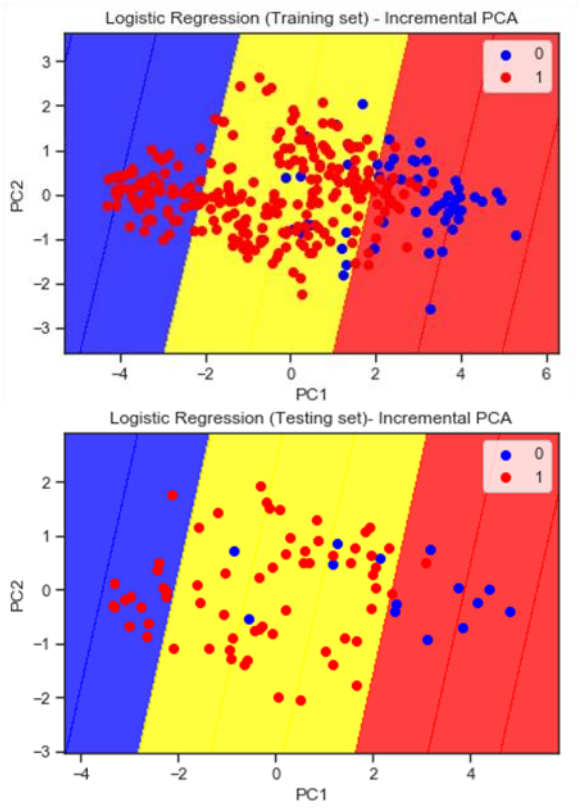

Fig. 23 Logistic Regression for Incremental PCA with two Components

VI. CONCLUSION

This paper attempts to predict the chance of Admit for the Admission predict dataset by implementing the correlation matrix between each of the dataset attributes. The top most high correlated features are identified which are directly contributing to the prediction of chance of admit. Admission Predict dataset is subjected to dimensionality reduction methods like PCA, Sparse PCA, Incremental PCA , Kernel PCA and Mini Batch Sparse PCA. Optimized dimensionality reduced dataset is then executed to analyze and compare the mean squared error, Mean Absolute Error and R2 Score of each method. Experimental Result shows that the CGPA, GRE Score and TOEFL Score are highly correlated features in predicting the chance of admit. The execution of performance analysis shows that Incremental PCA have achieved the effective prediction of chance of admit with minimum MSE of 0.09, MAE of 0.24 and R2 Score of 0.26.

REFERENCES

1. Lau Lilleholt, Anders Aaby, and Guido Makransky, " Students admitted to university based on a coginitive test and MMI are less stressed than students admitted based on CGPA", Studies in Education Evaluation, Elsevier., vol. 61, Jun. 2019, pp. 170-175.
2. Joseph Janison, "Applying Machine Learning to Predict Davidson College's Admissions Yield", proceedings of the ACM SIGCSE Technical Symposium., 2017.
3. S. Maldonado, G. Armelini, and A. Guevara, "Assessing university enrollment and admission efforts via hierarchical classification and feature selection", Intelligent Data Analysis., vol. 21, no. 4, 2017.
4. William Eberle, Douglas Talbert, Erik Simpson, Larry Roberts, and Alexis pope, " Using Machine Learning and Predictive Modeling to Assess Admission Policies and Standards", Proceedings of the 9th Annual National Symposium., 2013.
5. Austin Waters, and Risto Miikkulainen, "GRADE: Machine Learning Support for Graduate Admissions", Association for the Advancement of Artificial Intelligence, Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference., 2013.
6. T. H. Bruggink, and V. Gambhir, "Statistical models for college admission and enrollment: A case study for a selective liberal arts college", Research in Higher Education., vol.37, no. 2, 1996, pp. 221–240.
7. J. S. Moore, "An expert system approach to graduate school admission decisions and academic performance prediction", Omega., vol. 26, no. 5, 1998, pp. 659–670.
8. W. O. Dale Amburgey, and John Yi, "Using Business Intelligence in College Admissions, Principles and Applications of Business Intelligence Research", 2013, pp. 1-16.
9. Kanadpriya Basu, Treena Basu, Ron Buckmire, and Nishu Lal, "Predictive Models of Student College Commitment Decisions Using Machine Learning", Data., vol. 4, no. 2, 2019, pp. 65.
10. E. Yukselturk, S. Ozekes, and Y. K. Türel, "Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program", European Journal of Open Distance E-Learning., vol. 17, 2014, pp.118–133.
11. Mayra Alban, and David Mauricio," Predicting University Dropout through Data Mining: A Systematic Literature", Indian Journal of Science and Technology., vol. 14, no. 4, 2019.
12. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification (Accepted for publication)", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, 2019 to be published.

13. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers(Accepted for publication)", Springer's book series entitled "Learning and Analytics in Intelligent Systems, Springer, 2019 to be published.
14. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification , Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739. DOI: 10.1080/09720510.2019.1609729
15. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,"Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
16. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, " Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.