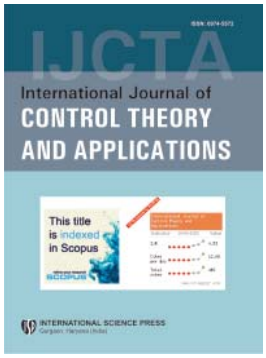

International Journal of Control Theory and Applications

ISSN : 0974–5572

Volume 9 • Number 42 • 2016

A Comparative Study of Principal Component Analysis vs Wrapper Method, an Overview of Dimensionality Reduction Techniques Applied in Developing an Undergraduate Student Dropout Model

Vinayak Hegde[a] Shruthi G. Kini[a] and Sahana A[a]

[a]Department of Computer Science Amrita School of Arts and Sciences Amrita Vishwa Vidyapeetham, Amrita University Mysuru Campus Mysuru, Karnataka, India

E-mail: vinayakhegde92@gmail.com, shruthigkini@gmail.com, sahanapadmanab@gmail.com

Abstract: Data generated in real-time will be erroneous in nature with multiple dimensions and hence demands for dimensionality reduction. Dimensionality reduction is the process of extracting significant information from higher dimensions reduced into their lower form. They are broadly classified into feature selection and feature extraction approach. Feature selection is a procedure of selecting the right subset of attributes based on the learner's objective and feature extraction involves transforming the data of higher dimensions to lower dimensions. The Principal component analysis which is the feature extraction approach is a procedure of variable reduction by selecting the most significant variables with high variance and less collinearity whereas wrapper method is a procedure of training the new feature based on classifier's intended. We are developing an undergraduate student dropout model. Dropout is a process wherein a person decides to drop from the academics before the completion of a program of studies. This paper intends to provide a novel approach in comparing the efficiency of the two methods, that is, principal component analysis and the wrapper method by choosing the best attribute selection algorithm for developing a student dropout model which helps to provide efficiency in analysing and predicting the student dropout, all of which can aid various educational institutions.

Keywords: *Dimensionality reduction, Feature selection, Feature extraction, Principal component analysis, Wrapper method*

1. INTRODUCTION

Data is generated as a by-product of tweets, retweets, blogs, streaming of data, daily chats, web documents, web usage data, stock rate information, image sharing, instant messaging, call logs and the like. These data are accumulated very rapidly and may prove to be a challenge to maintain as such data may be prone to errors. Data which contains large dimensions are beyond human interpretation as they do not fetch valuable information leading to time and space complexity along with the undue influence on cost for processing them. To solve the problem of multiple dimensions, we need to choose the attributes which are most appropriate from the larger spaces. Reducing the high dimensional data into their lower forms while upholding the intrinsic information

extracted from the larger spaces is what is come to be known as dimensionality reduction. Dimensionality reduction is used for various purposes such as visualizing the data in 2D or 3D, compressing the data for effective analysis and noise removal to ensure that accurate data is obtained in response to a query, which is important as the noisy data contribute a negative impact on the accuracy of the model. We have compared the best algorithm of dimensionality reduction technique and adopted them in the student dropout model to analyse and predict the dropout. The student dropout is depending on academic performance. It can be analysed using z factor [15] In short; dimensionality reduction provides a solution by improving the analysis process by reducing the complexity.

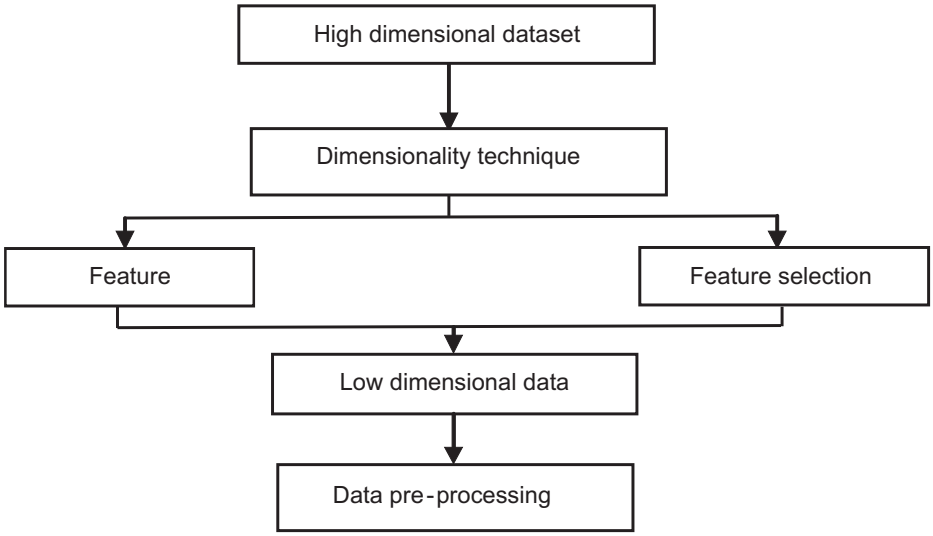

Figure 1: Various levels of dimensionality reduction

2. RELATIVE WORK

V. Arul Kumar et al [1] This paper says suggests to apply the techniques of dimensionality reduction for reducing multiple dimensions with in detail explanation. Cattell, R.B. et al [2] has focused PCA to be the best approach and says that covariance matrix, eigen values are the best pursuits for dimensionality reduction is the most widely used dimensionality reduction technique and in the PCA for minimizing the problem we use covariance matrix, eigenvectors and also Eigen values. PCA can be often used in reducing dimensionality, Multiple linear regression and also several regression techniques projection pursuit and independent component analysis. Muhammad Shakil Pervez et al [3] has completely reviewed the various techniques of feature selection and differentiated them. Albert Bifet et al [4] To reduce the multi-dimensionality author has compared a genetic algorithm with existing PCA technique. The genetic algorithm is used as a search engine then the strategy that has minimum validation error is selected finally the summation of all separate results are taken for final accuracy. Later this accuracy is compared with PCA accuracy.

3. MOTIVATION

To ease the problem of multi-dimensionality we approach for comparing the two methods, that is principal component analysis method and wrapper method. Principal component analysis is a feature extraction approach that upholds the attributes with higher variance and lower collinearity existing between them whereas wrapper method of feature selection approach helps in selecting the best subset the best subset of feature based on the learner's objective. Comparing their efficient we propose an efficient subset selection algorithm as a way solve high dimensional data.

4. METHODOLOGY

To develop a model for predicting the student dropout. A step by step procedure was followed.

4.1. Pre-Survey

In this phase, we perform a survey on the students who have enrolled themselves in a program with the educational institution. This survey is mainly conducted to analyse the student's prior history such as his/her marks in English, Mathematics and Science to analyse the student's systematic and organizing thinking, logical skills and communication skills.

4.2. Post-Survey

In this phase, we perform a survey on the students at the middle of the program for analysing the student on various aspects which depends on the financial status, academic performance, number of FA's (fail to attend), number of current backlogs, psychological factors and so on.

4.3. Pre-Processing the Data by Converting Text to Ordinal Data

Once the data is collected through pre-survey and post-survey. The data will be stored in the excel file. The data so collected will be converted into their ordinal values. As the survey is conducted through online we have no null values to be pre-processed, as every field was made mandatory to be filled.

4.4. Convert the Excel File to CSV File

The data saved in the excel file will be converted into CSV file for loading them into R or WEKA environment.

4.5. Loading the CSV Data for Analysis in R Environment and Computations Using Principal Component Analysis

Principal component analysis is a feature extraction approach. Feature extraction is a process of retaining most significant features from large spaces whenever a similar set of features are encountered that cause redundancy, they are eliminated. PCA is used for reducing large dimensional dataset into a new set called as principal components. Each component holds highest variance that forms orthogonal to the previous component represented which are uncorrelated in nature. PCA in other words is a linear combination of variables.

Steps involved in the algorithm:

1. **Find the Mean of the Ordinal Data:** Mean is calculated for the ordinal data using formula:[11]

$$\bar{x} = \sum_{i=1}^{n} X_i$$

Where $\bar{x}$ indicates the mean set of X. Mean is the sum of observations by total number of observations.

We compute the variance to find the measure of spread of the data which helps us to understand covariance. The formula for computing variance is [11]:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

where is the variance that denotes the squared standard deviation.

2. **Calculate the Covariance to Form Covariance Matrix:** Next step involves calculating the relation between a pair of variables, it helps in checking how each dimension vary from the mean with respect to each other. [11]. There are two types of covariance, that is, positive covariance which indicates

that the changes made to one variable have same impact with other, in the sense that these variables increase or decrease together and another type of covariance is the negative covariance which has negative impact in the sense that they move in different directions, say with increase of one variable we have decrease in another variable. The formula for covariance is as follows [11]:

$$\mathrm{COV}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

We can represent covariance in other terms the correlation in the form of a covariance matrix.

3. **Calculate the Eigen Values and the Eigen Vectors:** Eigen values is the value that depicts the variance present in the data and the eigen vectors depicts the direction of the line for the data values. The eigen vector with the highest eigen value constitutes a principal component. The formula for computing eigen value is [11]:

$$AY = Y$$

Where A is a square matrix, λ is an eigen value and Y is a non-zero factor [11]. Eigen values are depicted from a square matrix. Square matrix is represented in terms of *m*n* form. For any value of λ for which the above equation has a solution is called the eigen value of A and the vector v which corresponds to the eigen value is the eigen vector of A. We can depict them as [11],

$$\begin{aligned} A*V &= \lambda \\ A*V - \lambda*T*V &= 0 \\ (A - \lambda*I)*V &= 0 \end{aligned}$$

By finding the roots of $|A - \lambda*I|$ we will get eigen values and each of these eigen values are our eigen vectors Eigen vector are plotted as diagonal dotted lines on the plot [11].

4. **Transpose the Data:** We transpose the matrix by replacing all rows with columns and vice-versa using the formula [11]:

$$\begin{aligned} A &= [a_{ij}]_{m*n} \\ A^T, A^1 &= [a_{ji}]_{m*n} \end{aligned}$$

where A^T represents the transpose of a matrix

Derive New Dimensional Data: After performing transpose of data we obtain new dimension data which are equivalent to the original dataset but the unrelated dimensions are removed.

4.6. Loading the CSV Data for Analysis in Weka and Computations Using Wrapper Method

Wrapper method is a feature selection approach. Wrapper method chooses a new subset of features by training a new model for each subset based on the learner's objective. Wrapper method was implemented in WEKA to analyse the efficiency. We have selected Naïve Bayesian classifier for classifier subset evaluation under attribute evaluator. Naïve Bayesian classifier is based on Bayesian probability model. It considers that the value of one attribute is independent of value of another attribute and set of attributes help in recognizing a specific feature. We have selected best first search under search method. Best first search spans through unvisited node and directs the next node to be visited.

Steps involved in the algorithm :

1. **Convert the Excel File to CSV File :** The first step is to convert the collected survey data in excel format To CSV format.

2. **Uploading the File :** Using the latest version of weka upload the CSV file and span through select attributes options and select the choice for attribute evaluator and search method as mentioned above.
3. **Computing results :** The final step is to obtain the results by following the above 1 and 2 procedures and start the process. A new feature subset will be obtained.

5. EXPERIMENTAL STUDY

5.1. First Phase of Conducting the Survey

We had conducted a survey on 1st and 3rd semester undergraduate students. In total 54 questions were posed which spanned through the areas of academics, demographical factors, psychological factors, social integration, social media and General information. Table I depicts the count of number of questions concentrated on each area.

Table 1
Tabular Representation of Questions Framed from Multiple Disciplines for an Online Survey

Areas concentrated	*Count of questions*
Academics	17
Demographical factors	5
Psychological factors	14
Health	1
Social integration	7
Social media	5
General information	5
Total	54

5.2. Variables considered for the research

We have considered 53 different dimensions, that is, variables based on the multiple behavior of the student for student dropout prediction. Table II depicts a sample of 10 variables.

Table 2
Sample of Variables Considered for the Computation

Variable_number	*Variable*	*Variable_number*	*Variable*
VAR1	Gender	VAR6	Percentage in 10th
VAR2	Age group	VAR7	Percentage in 12th
VAR3	Place of residence	VAR8	Describe me
VAR4	Km from college	VAR9	Financial assistance from outside
VAR5	PUC educ. Board	VAR10	Stay

5.3. Loading CSV File for Computing the Principal Component Analysis Results

CSV file was loaded to the R environment for computing principal component analysis and following results were obtained.

1. **Covariance matrix:** Correlation/ covariance is the measure used whenever we want to measure the variables in different scales.

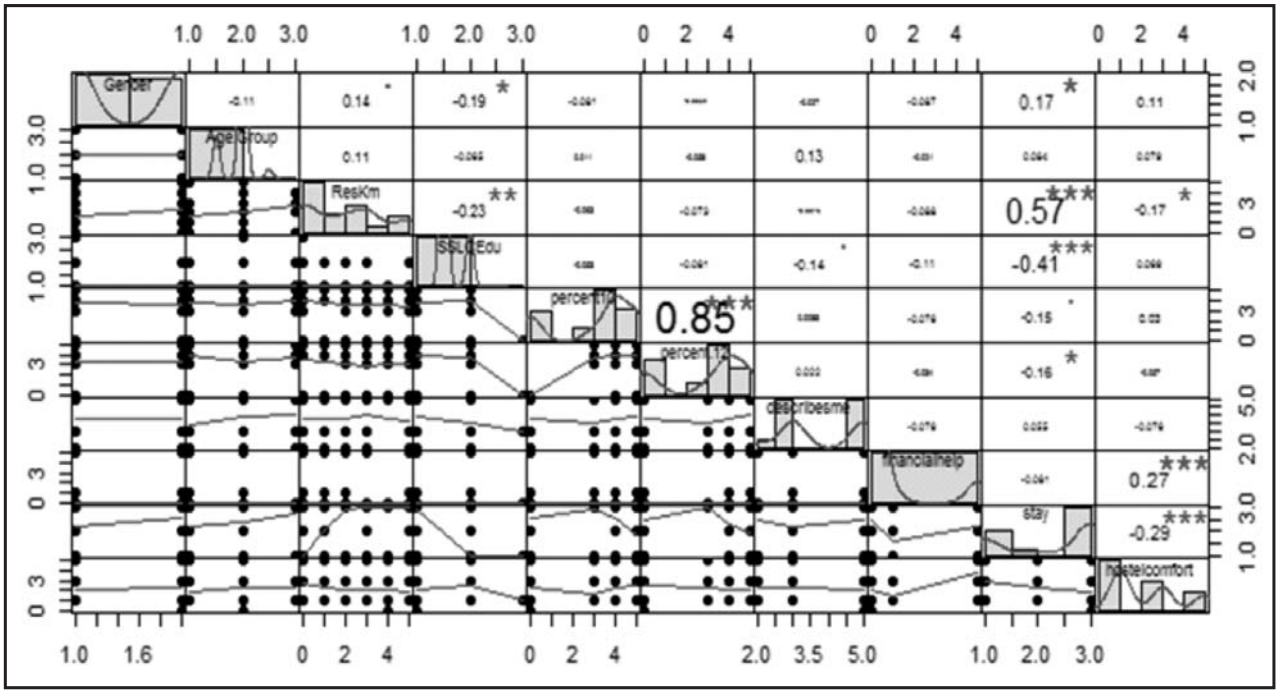

Figure 2: Figure depicts the covariance matrix of the original dataset collected from an online survey

2. **Scree plot:** It helps us to display the eigen values associated with a component in the descending order. They help us to visually analyse which components has the highest variance in the data.

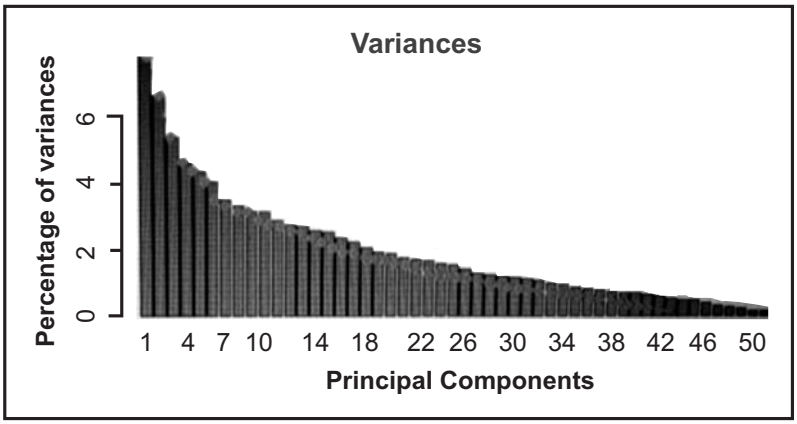

Figure 3: Figure depicts the scree plot representation of principal components in 'X' and percentage of variance in 'Y' in the original dataset collected from an online survey

5.4. Loading CSV File for Computing the Wrapper Method Results

CSV file was loaded to the R environment for computing wrapper method and following results were obtained.

1. **Attribute Selection in WEKA for Implementing Wrapper method:** Uploading the file in and selecting the naïve Bayesian classifier and best search method to get the optimal attributes fetched by WEKA.

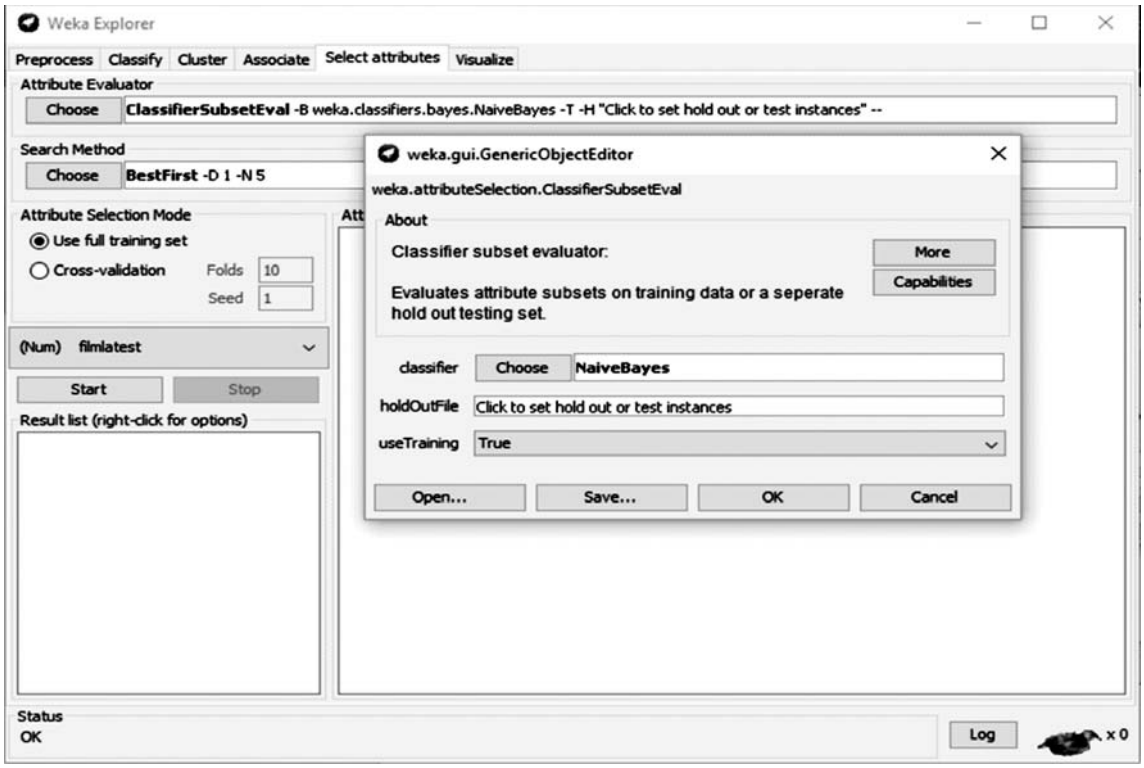

Figure 5: Figure depicts the approach chosen for attribute selector and search method

2. **Computation of output using wrapper method:** Using the selection procedure the approach fetches 4 most suitable variables with the highest variance.

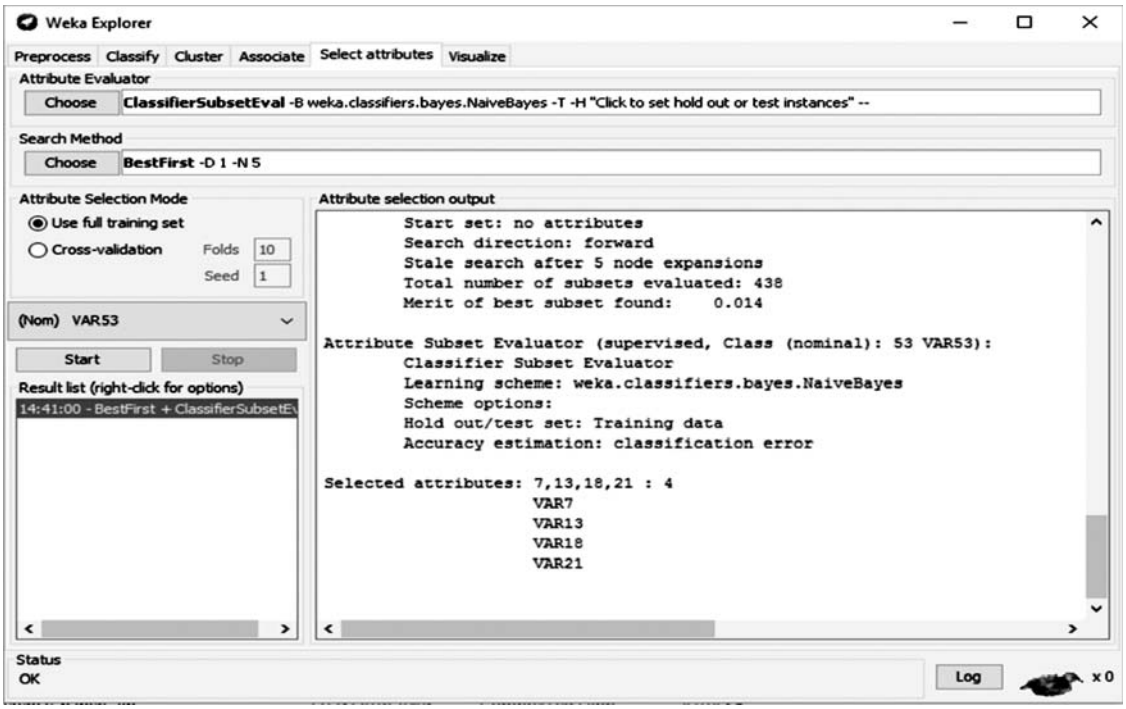

Figure 6: Figure depicts the biplot representation of the original dataset collected from an online survey

6. RESULT AND DISCUSSION

In our analysis between comparison of 2 algorithms, principal component analysis fetched 12 variables, that is, Gender, Km from college, percentage in 12th, Stay, comfortability, Hostel return back, Fail, financial problem in continuing education, satisfied with course, distraction, strive for personal attention (Variable 1,4, 7,10, 14, 18, 22, 26,30, 34, 38, 42) whereas wrapper method fetched 4 variables, that is, Percentage in 12th, enjoy studying, hostel return, sem2 CGPA (Variable 7, 13, 18,21). As principal component analysis fetched more variables it was considered to be the best subset selection algorithm over wrapper method as per our computations.

7. IMPLEMENTATION ISSUES

As computations are done in two different environments, that is PCA in R environment and wrapper method in WEKA, the computations are possible in the same environment but implementing wrapper method in R environment requires caret package which was incompatible with the latest version of R. Again, PCA which provides efficiency in selecting variables are represented in an orthogonal manner but are not suitable if they are not linearly correlated.

8. FUTURE WORK

Implementing wrapper method in R environment with the help of caret packages and comparing the efficiency of features obtained between WEKA's computation with the computations obtained in R environment is proposed to be done as a part of future work.

9. CONCLUSION

Huge increase in data often demands reducing dimensions and hence a dimensionality reduction is a useful tool. It helps us in analysing the original dataset in 2 or 3 dimensions. For developing the undergraduate student dropout model, PCA was concluded to be efficient in selecting the most appropriate subset of features by fetching 12 variables of highest variance and low collinearity. Thus, best algorithm selection provides efficiency in analysis and prediction.

10. ACKNOWLEDGMENT

First and foremost, we feel deeply indebted to her holiness most revered Mata Amritanandamayi Devi (AMMA) for her inspiration and guidance both in unseen and unconcealed ways. I would like to thank my parents and friends for being a supporting pillar in our work. I would like to express our heartfelt gratitude to our guide Mr. Vinayak Hegde, Asst. Professor, Department of Computer Science for his valuable suggestions and excellent guidance rendered throughout this project.

REFERENCES

[1] Tian, C., Wang, Y., Lin, X., Lin, J., & Hong, J. (2016). Research on High-Dimensional Data Reduction. International Journal of Database Theory and Application, 9(1), 87-96.

[2] Pervez, M. S., & Farid, D. M. (2015). Literature Review of Feature Selection for Mining Tasks. International Journal of Computer Applications, 116(21).

[3] Porkodi, R. (2014). comparison of filter based feature selection algorithms: An overview. international journal of innovative research in technology& science, 2(2), 108-113.

[4] Sorzano, C. O. S., Vargas, J., & Montano, A. P. (2014). A survey of dimensionality reduction techniques. arXiv preprint arXiv:1403.2877.

[5] Liu, H., Motoda, H., Setiono, R., & Zhao, Z. (2010). Feature Selection: An Ever Evolving Frontier in Data Mining. FSDM, 10, 4-13.

[6] Bifet, A. (2009). Adaptive learning and mining for data streams and frequent patterns. ACM SIGKDD Explorations Newsletter, 11(1), 55-56.

[7] Janecek, A. G., & Gansterer, W. N. (2008). A Comparison of Classiffication Accuracy Achieved with Wrappers, Filters and PCA. In Workshop on New Challenges for Feature Selection in Data Mining and Knowledge Discovery.

[8] Janecek, A., Gansterer, W. N., Demel, M., & Ecker, G. (2008, September). On the Relationship Between Feature Selection and Classification Accuracy. In FSDM (pp. 90-105).

[9] Inza, I., Larrañaga, P., Blanco, R., & Cerrolaza, A. J. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial intelligence in medicine, 31(2), 91-103.

[10] Jolliffe, I. (2002). Principal component analysis. John Wiley & Sons, Ltd.

[11] Smith, L. I. (2002). A tutorial on principal components analysis. Cornell University, USA, 51, 52.

[12] Hall, M. A. (2000). Correlation-based feature selection of discrete and numeric class machine learning.

[13] Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1), 273-324.

[14] Ramadevi, G., & Usharani, K. Study on Dimensionality Reduction Techniques and Applications.

[15] V. Hegde and M. S. Pallavi, "Descriptive analytical approach to analyze the student performance by comparative study using Z score factor through R language," 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, 2015, pp. 1-4.