Big Data Analytics: Student Performance Prediction Using Feature Selection and Machine Learning on Microsoft Azure Platform

Wattana Punlumjeak, Nachirat Rachburee and Jedsada Arunrerk
Department of Computer Engineering,Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathumthani, Thailand.
wattana.p@en.rmutt.ac.th

***Abstract*—In recent years, big data analytics has been a new growing research area and the essence of cloud computing is used to support a shared pool of resources. In educational mining, the huge volume of student data needs analytics technologies to extract valuable knowledge. It has been recognized that a high performance accuracy of student prediction model will be helpful for student and stakeholders. In this experiment, feature selection methods were proposed to identify the most significant and intrinsic features before classification methods were used. Experiment was conducted to evaluate the performance of the prediction model. The result of the experiment showed that mutual information in feature selection method with neural network classifier gave the best overall accuracy at 90.60% for student's data at Rajamangala University of Technology Thanyaburi. This experiment is extremely useful for students, teachers and management to find useful knowledge not only in identifying the problem areas and reasons that affect student's performance, but also in understanding the feature selection and classification methods, which are the most effective way to analyze student's performance on a cloud computing environment.**

***Index Terms*—Big Data Analytics; Feature Selection; Microsoft Azure; Student Prediction.**

I. Introduction

In recent year, big data and data analytics are both hot topics in the world of science and technology. In terms of big data, a "3Vs" model is interpreted as three important characteristics: volume, velocity and variety. The large amount of data generated in real life such as bioinformatics, marketing, medicine, and social network becomes a huge volume, which is accumulated at a very fast speed and continuous streaming information. Velocity refers to the speedy and timely data collection and data analysis to utilize the commercial value. Big data in terms of variety could be defined as the various types of data stored in data lakes, which have many variety style and no prior database designed, and it can be traditional structured, semi-structured or unstructured [1].

Data science is a discipline that integrates several fields e.g. mathematics, statistics, information science, machine learning, and data mining to process in terms of extracts or generate an in-depth insight from data in various forms. Data analytics plays a key role in scientists' life. The large amount of data requires new techniques, algorithms and analytics to automatically extract valuable hidden knowledge. Gaining understanding and analyzing data using affective techniques and suitable platforms become significant in the field of big data [2-3].

Cloud computing is a paradigm for infrastructure, platforms, and software consumption, in which users consume from a shared pool of resources, such as networks, servers, storage, applications, and services that service provider manages [4]. Cloud computing technologies can be divided into three levels: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS), as shown in figure 1. Nowadays, there are many cloud providers such as Amazon Web Services (AWS), Google, Microsoft Azure and etc. There are providers who provide infrastructure as a service (IaaS) only, while there are providers who provide infrastructure as a service (IaaS) as well as platform as a service (PaaS). In public cloud environment, user has to pay for on-demand computing resources (e.g. networks, servers, storage, applications, and services) [5-6].

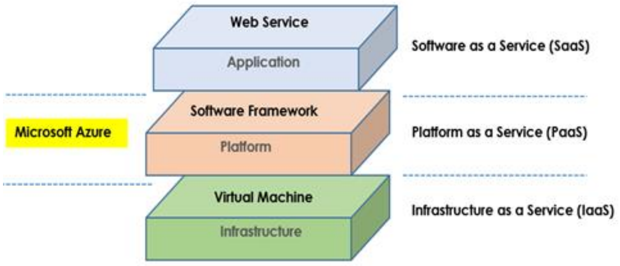


Figure 1: Cloud computing architecture

In 2015, there were more than 138 universities distributed all over every regions in Thailand. Many universities have more than one campus, serving as an equal educational opportunity to the Thai people. For example, Rajamangala University of Technology has nine campuses located in every region in Thailand, there were more than 23,748 students from 11 faculties studied n the main campus in 2015. The large amount of student data, whose volume is very high considered as big data, which can be analyzed using analytic technologies is potentially valuable. Specifically, the student data can be analyzed to obtain information not only for improving academic planning and informing valuable information for decision, but also making decisions and strategies for academic planning and management.

In this research, Microsoft azure has been proposed as the machine learning in a cloud-based platform. In the framework, feature selection method was applied to reduce some features in student data to retain high accuracy without losing any important information. Three methods in the feature selection, namely Chi-square, Pearson Correlation, and Mutual information were used to find the best set of feature for each method. The classification method in supervised learning, namely the Decision forest and neural network were trained using the feature set and then the performance was compared to make the final decision.

After the introduction, materials and methods including the proposed model is presented in Section II, follow by the result and discussion in Section III. Finally, Section IV provides the conclusion of this research.

II. MATERIALS AND METHODS

A. Big Data

In cloud–based framework, data comprising more than five million from students who registered at the open university of China (OUC) indicated that there were some significant rules via K-Mean clustering and Apriori associated algorithm. The result showed that the students' learning skills have been grouped into four groups namely, attitudinal skills, cognitive skills, communication skills, and relational skills [7]. The advantage of big data did not only improve PHP Programming teaching method in vocational colleges in China, but also gave high learning efficiency and teachers' professional competency [8]. Big data analytics was discussed with the learning performance prediction system in collaborative learning to conduct the most efficient curriculum paths, course, module, and individual knowledge in order to maximize student potential. Extreme learning machine based feed forward neural networks and regression tree were used to predict group performance [9]. A real-life adult dataset from UCI machine learning repository acted as a big data to conduct a pre-clustered classification method and multiple classifier system to combine the result of clustering [10].

B. Feature Selection

Feature selection was used in data pre-processing to find out the most significant and intrinsic features. The methods used in Microsoft azure were Chi-square, Pearson correlation coefficient, and Mutual information.

a. Chi-square

In feature selection, chi-square was applied to test the independence of two kinds of comparison: The tests of independence and the tests of goodness of fit. The test of independence was assessed by chi-square and estimated whether the class label is independent of a feature. The score of chi-square score with c class and r values is defined as Equation (1) and (2).

$$x^2 = -\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}\text{-}\mu_{ij})}{\mu_{ij}} \tag{1}$$

$$\mu_{ij} = \frac{(n_{*j}n_{i*})}{n} \tag{2}$$

where: n = number for samples
n_{*j} = amount of samples in class j
n_{ij}, n_{i*} = amount of samples value with the i^th value of the feature.

Chi-square was applied in the feature selection method to separate the factors of successful from unsuccessful students in Chittagong University before using CART and CHAID tree classifiers. From their experiment, Chi-square selected the most importance factors that can separate the successful from the unsuccessful students as follows: financial support, age group and gender [11]. In the Malay opinion mining and sentiment analysis research, seven methods of feature selection e.g. principal components analysis (PCA), Gini Index, chi-square, Relief-F (RE), support vector machines (SVM) and others, were used in feature selection step on three machine learning classifiers. Although, SVM provided the highest accuracy performance in order to classify Malay sentiment comparing with other feature selection, chi-square gave more high accuracy performance than RE and PCA [12].

b. Pearson Correlation Coefficient

Pearson's linear correlation coefficient is a statistic measurement, which measures the strength of relation between the distribution of feature values x and the class label c. Pearson's linear correlation coefficient is defined as Equation (3) and (4).

$$\rho(X,C)= \frac{E(XC)\text{-}E(X)E(C)}{\sqrt{\sigma^2(X)\sigma^2(C)}} \tag{3}$$

$$\rho(X,C)= \frac{\sum_i(x_i - \bar{x_i})(c_i - \bar{c_i})}{\sqrt{\sum_i(x_i - \bar{x_i})^2 \sum_j(c_j - \bar{c_j})^2}} \tag{4}$$

where, ρ(X, C) = ± 1, if feature value x are linear dependent. Otn the other hand, uncorrelated is represented by zero.

The purpose of the research experiment was to combine several variable selection methods (e.g. Pearson coefficients, Kolmogorov-Smirnovtest and the Fisher criterion score) and dimensionality reduction algorithms e.g. PCA. The experiment composed of two steps. The first step is extracting informative features with a variable selection algorithm from the data. Then, dimensionality reduction to extract the most informative directions was applied. Their dataset come from two publicly available datasets: the YALE dataset and the AR dataset. The accuracy result makes sense only when using the same utility function in both stages [13]. A Pearson Redundancy Based Filter (PRBF) was used in feature selection method to remove redundancy in bioinformatics dataset, which is high-dimensional data. Support vector machine was used to evaluate the performance of PRBF algorithm. The result showed that PRBF algorithm works well with a linear SVM for n > 100 samples [14].

c. Mutual Information

Mutual information (MI) of two random variables is the variables' mutual dependence measurement in information theory. Two discrete random variables of X and Y in mutual information can be defined as Equation (5) and (6).

$$I(X;Y) = \sum_{x,y} Pxy(x,y) \log \frac{Pxy(x,y)}{Px(x)Py(y)} \tag{5}$$

Here, P X (x) and P Y (y) are the marginal.

$$Px(x) = \sum_{y} Pxy(x,y) \tag{6}$$

Feature selection based on mutual information was applied in four different datasets to select a good feature according the maximal statistical dependency criteria (max-dependence, max-relevance, and min-redundancy). Max-dependence (MaxRel) and max-relevance and min-redundancy (mRMR) method were used in this experiment, and the performance accuracy was evaluated by naïve bay (NB), linear discriminant analysis (LDA), and support vectors machines (SVM). The result of the mutual information based on mRMR showed outstanding high classification accuracy [15]. In the study of the students' performance in Rajamangala University of Technology Thanyaburi Thailand, several feature selection method (filter and wrapper method) were presented to find the best accuracy. mRMR, showed a minimum and significant feature subset with high accuracy of students' performance [16].

C. *The Proposed Model*

The main idea of our work is shown in the proposed model as followed:

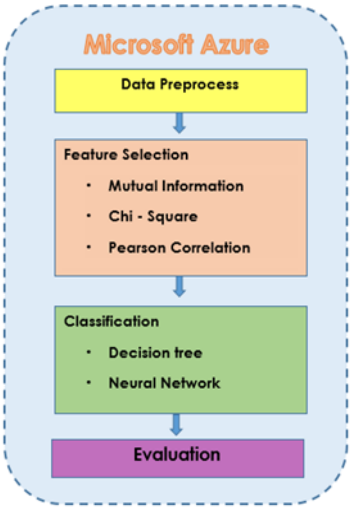


Figure 2: Proposed Framework

a. *Data Preprocessing Step*

In this research, we used data from the Faculty of Engineering, Rajamangala University of Technology Thanyaburi, Pathumthani, Thailand from year 2004 to 2012. There was a large amount of student log data registered in the subjects during their first year of study. 630,503 records had been grouped into 7,537 records based on student_ID. We were interested only the subjects registered by first year student and their admittance data. Hence, the feature dataset in this experiment consisted of 15 features as shown in Table 1.

In the data pre-processing step, data cleaning is a task to remove noisy and correct inconsistent of the data, for example student's record with grade 'W' letter, which means student withdraw that subject. Thus, student's record with grade 'W' letter shall be removed from this experiment. In student record, students who regrade the subject and get a 'B', the grade is in the form of 'B$', which means that this student has grade 'B' after regrading. Therefore, cleaning data for this record is to remove '$' letter from this attribute. In our experiment, we transformed student's GPA data in data transform step by discretizing into categorical classes. A class is divided into three classes consisting of high, medium, and low, as shown in Table 2.

Table 1
List of feature

Feature no.	Feature name	Description
1	year_entry	Student's entry year
2	year_curr	Year of curriculum
3	level	Type of student
4	program	Student program
5	admit	Type of student entry
6	cal1	Calculus1 Subject
7	cal2	Calculus2 Subject
8	cal3	Calculus3 Subject
9	phy1	Physic1 Subject
10	phy2	Physic2 Subject
11	chem	Chemistry Subject
12	com	Computer programming Subject
13	mat	Material Subject
14	mech	Mechanic Subject
15	draw	Drawing Subject

Table 2
Categorical Class

Class	Possible Value (GPA Range)
High	3.00 - 4.00
Medium	2.00 - 2.99
Low	1.00 - 1.99

b. *Feature Selection Step*

The proposed model used three feature selection methods: Chi-square, Pearson Correlation, and Mutual information. We applied each method with a full dataset to find a set of significant feature. We were interested in 5 and 10 significant features as shown in Table 3.

Table 3
List of Significant Features

Method	5	10
Chi-square	cal3, phy2, chem, cal2, mat	cal3, phy2, chem, cal2, mat, com, cal1, mech, phy1, draw
Pearson Correlation	cal3, cal1, cal2, mat, phy2	cal3, cal1, cal2, mat, phy2, chem, mech, com, phy1, draw
Mutual information	cal3, phy2, cal2, com, mat	cal3, phy2, cal2, com, mat, chem, cal1, mech, phy1, draw

c. *Classification Step*

We selected a classical classification model in supervised method to evaluate a performance of machine learning model. Decision forest and neural network were selected to do this work.

d. *Evaluation Step*

In machine learning, a specific table named as confusion matrix represents a visualization of the performance of supervised learning. Confusion matrix is composed of two rows and two columns that reports the numbers of true positives, true negatives, false positives, and false negatives as shown in Table 4.

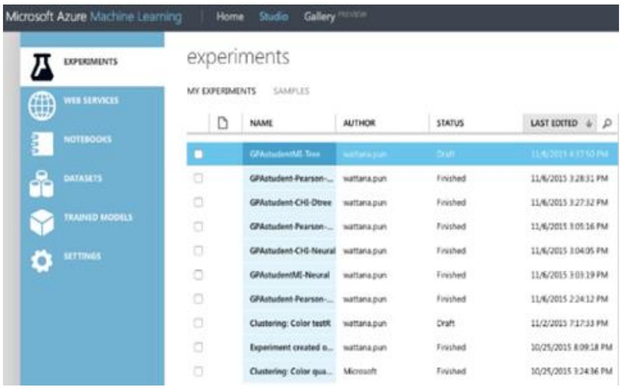
Table 4
Confusion Matrix

	Predicted Class	
Actual Class	true positives	false positives
	false negatives	true negatives

The overall performance accuracy of this model is verified by ten-fold cross validation test. 70% of the training dataset were randomly used and 30% were for testing.

III. Results And Discussion

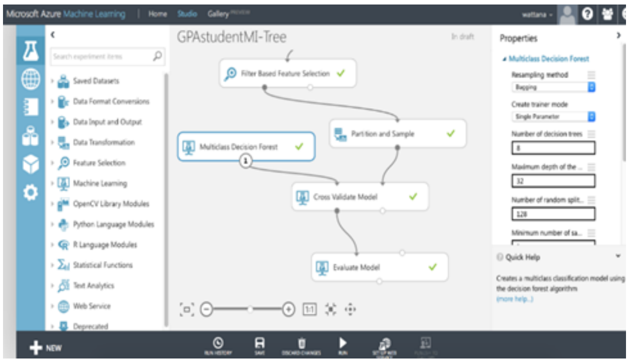
We applied an experiment on Microsoft Azure platform by creating a machine learning workspace, as shown in Figure 3.



NAME	AUTHOR	STATUS	LAST EDITED
GPAstudentMI-Tree	wattana.pun	Draft	11/6/2015 4:17:10 PM
GPAstudent-Pearson-...	wattana.pun	Finished	11/6/2015 3:28:31 PM
GPAstudent-CHI-Dtree	wattana.pun	Finished	11/6/2015 3:27:32 PM
GPAstudent-Pearson-...	wattana.pun	Finished	11/6/2015 3:05:16 PM
GPAstudent-CHI-Neural	wattana.pun	Finished	11/6/2015 3:04:05 PM
GPAstudentMI-Neural	wattana.pun	Finished	11/6/2015 3:03:19 PM
GPAstudent-Pearson-...	wattana.pun	Finished	11/6/2015 2:24:12 PM
Clustering: Color testR	wattana.pun	Draft	11/2/2015 7:17:33 PM
Experiment created o...	wattana.pun	Finished	10/25/2015 8:09:18 PM
Clustering: Color qua...	Microsoft	Finished	10/25/2015 3:24:36 PM

Figure 3: Example of machine Learning Workspace

We created each classification model in Azure ML-studio, selected an item on the left hand using the drag and drop function. After that, the properties were set for each item using default parameter in Azure ML-studio. An example of the model of classification created in Azure ML-studio is as shown in Figure 4.



Multiclass Decision Forest
Resampling method: Bagging
Create trainer mode: Single Parameter
Number of decision trees: 8
Maximum depth of the ...: 32
Number of random split...: 128
Minimum number of sa...

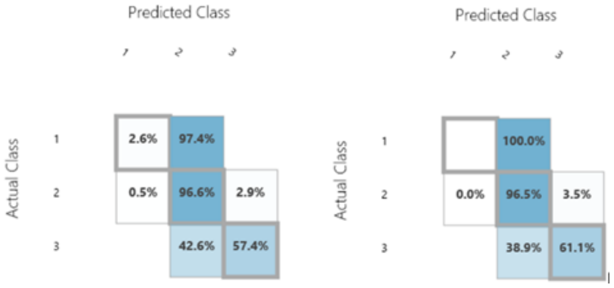
Creates a multiclass classification model using the decision forest algorithm
(more help...)

Figure 4: Example of classification model in ML-studio

After we set all the experiments in every method we used in our research, we run Azure in all the experiments on the cloud computing. An example of the results of this experiment is as shown in the confusion matrix with 10 features subset for Chi-Dtree, Chi-NN, PCC-Dtree, PCC-NN, MI-Dtree, and MI-NN method in Figure 5-7 respectively.

From Figure 5-7, the number of students in class 1, class 2, and 3 were 114, 6442, and 941 respectively. True positive in class 2 and class 3 showed high percentage in every model, which means that the models can be predicted the class correctly. On the other hand, the percentage of true positive in class 1 showed something unsatisfied.

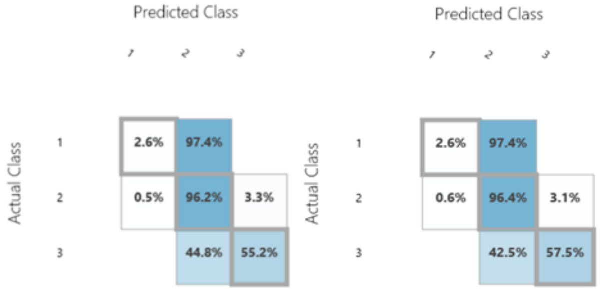
The overall accuracy of prediction model, which has three feature selection methods (Chi-square, Pearson correlation coefficient, and Mutual information) and two classification methods (D-Tree and Neural network) with 5, 10 and all features is shown in Table 5 and Figure 8:



	1	2	3
1	2.6%	97.4%	
2	0.5%	96.6%	2.9%
3		42.6%	57.4%

	1	2	3
1		100.0%	
2	0.0%	96.5%	3.5%
3		38.9%	61.1%

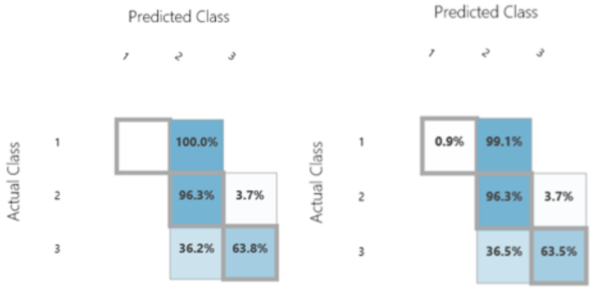
Figure 5: Confusion Matrix of Chi-Dtree and Chi-NN



	1	2	3
1	2.6%	97.4%	
2	0.5%	96.2%	3.3%
3		44.8%	55.2%

	1	2	3
1	2.6%	97.4%	
2	0.6%	96.4%	3.1%
3		42.5%	57.5%

Figure 6: Confusion Matrix of Pearson-Dtree and Pearson-NN



	1	2	3
1		100.0%	
2		96.3%	3.7%
3		36.2%	63.8%

	1	2	3
1	0.9%	99.1%	
2		96.3%	3.7%
3		36.5%	63.5%

Figure 7: NN Confusion Matrix of Mi-Dtree and Mi-NN

Table 5
Overall Accuracy

	Overall Accuracy		
Method	5	10	All
Chi -Dtree	88.31	90.07	90.38
Chi-NN	89.94	90.39	89.9
PCC-Dtree	88.43	89.46	90.38
PCC-NN	90.03	90.57	89.9
MI-Dtree	88.78	89.87	90.38
MI-NN	90.32	90.6	89.9

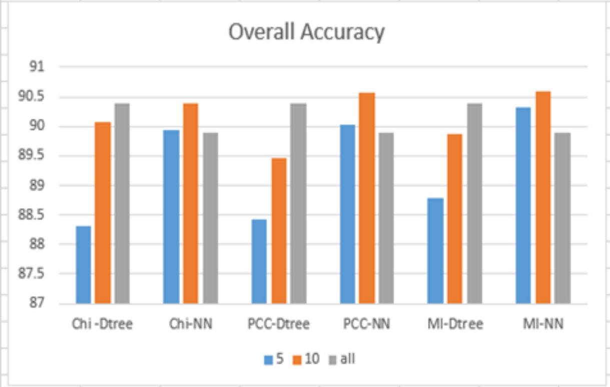


Figure 8: Comparative results of Chi, PCC, and MI

In this experiment, the overall accuracy showed that mutual information with 10 significant features in feature selection method with neural network classifier gave the best accuracy at 90.60%.

IV. CONCLUSION

We applied a large student data set, as a big data, to find a prediction model to classify the students' performance on Microsoft Azure platform. In the pre-processing step, we used three feature selection methods, namely Chi-square, Pearson correlation coefficient, and mutual information to find a significant feature in terms of each method. Then, the classification mining technique was applied on the dataset. The result of this experiment was evaluated by the confusion matrix and the overall accuracy with ten-fold cross validation. Mutual information in the feature selection method with neural network classifier gave the best overall accuracy at 90.60%.

Although, the overall accuracy of the proposed model was high, the confusion matrix seemed to be in conflict in some classes e.g. class 1(class low). The imbalance of the data may be the cause of this assumption. Therefore, we will study this conflict in our future work.

ACKNOWLEDGMENT

We thanks to Rajamangala University of Technology Thanyaburi, Pathumthani, Thailand for providing the student data.

REFERENCES

[1] Chen, M., S. Mao, and Y. Liu, Y., "Big Data: A Survey", *Mobile Networks and Applications*, vol.19, no.2, 2014, pp. 171-209.

[2] Swan, M., "Philosophy of Big Data: Expanding the Human-Data Relation with Big Data Science Services", *IEEE First International Conference on Big Data Computing Service and Applications*, 2015, pp. 468-477.

[3] Meetali, "From Big Data to Big Values: A Big Science Leading to a Revolution", *2015 2nd International Conference on Computing for Sustainable Global Development*, 2015, pp.56-59.

[4] Khurana, A., "Bringing Big Data Systems to the Cloud", *IEEE Cloud Computing*, (3), 2014, pp.72-75.

[5] Roloff, E., and et al., "Evaluating high performance computing on the windows azure platform", *2012 IEEE 5th International Conference on Cloud Computing*, 2012, pp.803-810.

[6] Akinbi, A., E. Pereira, and C. Beaumont, "Evaluating Security Mechanisms Implemented On Public Platform-As-A-Service Cloud Environments Case Study: Windows Azure", *2013 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, 2013, pp.162-167.

[7] Yi, J., and et al., "Cloud-Based Educational Big Data Application of Apriori Algorithm and K-Means Clustering Algorithm Based on Students' Information", *2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, 2014, pp.151-158.

[8] Li, X., and Y. Li., "The Reform of Vocational Colleges' Teaching Method in the Age of Big Data--Based on PHP Programming", *2015 IEEE Fifth International Conference on Big Data and Cloud Computing*, 2015, pp. 99-103.

[9] Cen, L., D. Ruta, and J. Ng., "Big education: Opportunities for Big Data analytics", *2015 IEEE International Conference on Digital Signal Processing*, 2015, pp.502-506.

[10] Liu, W. and D. Chen., "Big Data Classification Based On Multi-View Method", *2015 International Conference on Wavelet Analysis and Pattern Recognition*, 2015, pp.165-170.

[11] Mustafa, M. N., L. Chowdhury, and M. S. Kamal., "Students Dropout Prediction For Intelligent System From Tertiary Level In Developing Country", *2012 International Conference on Informatics, Electronics & Vision*, 2012, pp.113-118.

[12] Alsaffar, A. and N. Omar., "Study on feature selection and machine learning algorithms for Malay sentiment classification", *2014 International Conference on Information Technology and Multimedia*, 2014, pp.270-275.

[13] Wolf, L. and S. Bileschi., "Combining variable selection with dimensionality reduction", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2005, pp. 801-806.

[14] Biesiada, J. and W. Duch., "Feature Selection For High-Dimensional Data: A Pearson Redundancy Based Filter", *Computer Recognition Systems 2. Springer Berlin Heidelberg*, 2007, pp.242-249.

[15] Peng, H., F. Long, and C. Ding., "Feature Selection Based On Mutual Information Criteria Of Max-Dependency, Max-Relevance, And Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.27, no.8, 2005, pp.1226-1238.

[16] Punlumjeak w. and N. Rachburee, "A Comparative Study of Feature Selection Techniques for Classify Student Performance", *IEEE 7th International Conference on Information Technology and Electrical Engineering*, 2015, pp.425-429.