TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

Performance Analysis of a Student during a Learning Management System using Classification Algorithms

Dr Srinivasu Badugu
Dept of CSE Stanley College of Engineering and Technology for Women srinivasucse@gmail.com

Article Info
Volume 82
Page Number: 7658 - 7665
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 04 February 2020*

Abstract:
The use of e-learning systems has grown exponentially over the past decade, as a result of which enormous volumes of data are generated, this data must be used effectively to improve society in terms of improving education quality, ease of access to education, prompt response from teachers, etc. Educational data mining helps to solve the challenges of e-learning systems in a big way. In view of this, the primary objective of this paper is to analyze and compare the performance of classification algorithms such as K-nearest neighbor, Naive Bayes classification of data from the KalBoard 360 learning management system in evaluating participant performance.

Keywords: *Educational Data Mining, Data analysis, student data analysis, Learning Management Systems, K-Nearest Neighbor Classifier, Naïve Bayes Classifier, KalBoard 360.*

I. INTRODUCTION

The way we live, work, socialize, play, and learn has been influenced by computer-based technologies. Recent advances in various fields have resulted in huge quantities of data being collected and typically the data is stored in various formats such as documents, files, photographs, audio, and videos. The information collected is used in decision-making processes; but, as the amount of data is immense, it is complex and challenging to manage / analyze data. Use data to make better decisions requires a reasonable way to extract knowledge from vast databases.

To discover interesting and useful information, it is important to use large amounts of data[1]. Data mining is a powerful analytical tool that offers essential information and knowledge that can help improve processes of decision making. Data Mining allows researchers to make very useful discoveries from data, and these discoveries are particularly important in companies as the key decisions are made on the basis of these discoveries. Data mining and analytics techniques have developed field by field over the past decades[2].

"For many different fields (e.g. computer science, education, psychology, psychometrics, analytics, smart tutoring programs, e-learning, adaptive hypermedia, etc.)"[3] Educational Data Mining [2,3] (EDM) has emerged in recent years as a research area for researchers worldwide to evaluate huge amounts of data sets to solve educational research problems. EDM thinks about with developing techniques to research numerous sorts of knowledge in academic settings and to raised perceive learners and their learning environments through these ways. Fig.1 depicts the education data mining process.[3]

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

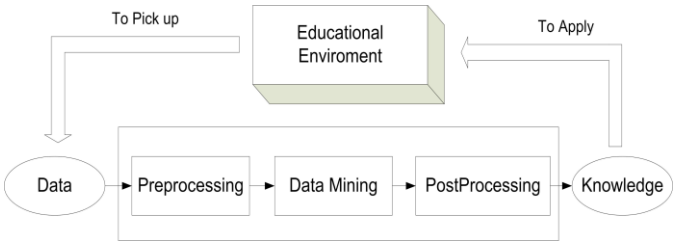

Fig 1 - Educational Data Mining Process diagram

"The EDM[3] approach converts unprocessed information from academic systems into helpful data which will be utilized by academic software package developers, teachers, tutorial researchers, etc"[3].

- Pre-processing: so as to show it into an appropriate format for mining, knowledge from the academic setting should be pre-processed first[3]. a number of the most pre-processing tasks are: data cleaning, attribute identification, knowledge transformation attribute integration, etc.[3]
- "Data mining: this can be the main stage that offers the full system its name. throughout this technique, on antecedently preprocessed data, Some examples of data mining techniques are data mining techniques are applied. visualisation, regression, filtering, clustering, association rule mining, serial pattern mining, text mining, etc."[3].
- Post-processing: this can be the ultimate step within which the tests or model obtained are understood and wont to build decisions regarding the educational environment[3, 4].

The conception of a learning management system(LMS) is software application that is used to organize, execute and assess training processes. an LMS provides instructors with a method of making and distributing content, following the engagement of learners and assessing results[5].

Learning Management System provides interactive features such as group conversations, video conferencing and forums for debate, etc. Types include Moodle, WebCT and Sakai[5].

II. LITERATURE SURVEY

Amireh et al.[6] used the WEKA tool for classification algorithms on the "LMS called kalboard 360" dataset17]. WEKA tool is safe and non-proprietary. This is the first time work combines the conduct of students with their academic success. Naive Bayes Classifier (NB), Decision Tree (DT) J48 and Artificial Neural Networks (ANN) are the classification algorithms implemented. The most popular apps were picked using the Filter Based user selection technique. Models are educated by feeding behavioral characteristics and excluding behavioral characteristics. It has been seen that the enhancement of accuracy when using behavioral features is: ANN gets twenty fifth percent to twenty ninth percent improvement, NB gets twenty second percent improvement, and DT gets six percent to seven percent improvement.

Roy et al.[7] also used WEKA software for classification algorithms, but the dataset used is from the database of UCI. Decision tree J48, Naive Bayes Classifier and Multi-Layer Perceptron (MLP) are the classification algorithms implemented. Among these algorithms J48 had highest accuracy i.e., 73.9% and MLP had the lowest accuracy i.e., 51.4% . The researcher also considers the qualities affecting the students ' success. Weekend and working day alcohol consumption, Parent's health, are the key factors influencing final grades. Romantic relationships also have a small impact on academic performance.

Juwita et al. [8] has used a latest version of dataset from LMS Kalboard 360. The authors used the WEKA tool and enforced Naive Bayes Classifier with 10-fold cross-validation, achieving associate accuracy of 67.7%. Amra et al.[9] focused as major classification algorithms on KNN and Naive Bayes to propose a student performance prediction model supported three analysis parameters (precision, accuracy, and recall). Naive Bayes had a better accuracy and WEKA tool was used for implementation. The used dataset is from Ministry of Gaza. The size of the dataset is 2000x8 but only

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

500 records were used for classification. RapidMiner tool was used for implementation.

The researchers in [10] proposed a framework that would classify students on Moodle course based on their behaviour. The design is checked at KSA's Majmaah University on 35 students enrolled in an online course on data structure. Moodle's versatility helps the teacher to monitor the conduct of the students during the course and in the quiz. This study provides guidance to educators to structure the contents of the course in a suitable manner that suits the learning style of the students in order to achieve the best learning process results. In [4] the authors lists several traditional data mining techniques that have been widely applied in the educational environment.

III. PROPOSED SYSTEM

We use Kalboard 360 LMS database with 480 records in the proposed system. Implemented classification algorithms are: K-nearest neighbor classifier (KNN) and Naive Bayes classifier (NBC)[9]. The algorithms in python programming language are implemented from scratch without any pre-built packages or sophisticated software being used. We want the KNN mode because the size of our dataset is small.

We apply the algorithms on the original dataset in the proposed system as well as the pre-processed dataset to get a clear idea of how the algorithms behave. We divide the data in the 70/30 ratio, i.e. 70% of the model learning instances and the remaining 30% of the model validation or testing instances, so we ultimately analyze the performance and compare the results. We also use the selection technique for correlation features to further improve NBC model efficiency. The model's performance is measured for reliability, accuracy, recall, and F-score. Eventually, to see which classifier works well, the tests are tabulated, contrasted and visualized.

Architecture of the Proposed System

The architecture of the proposed system is given as:

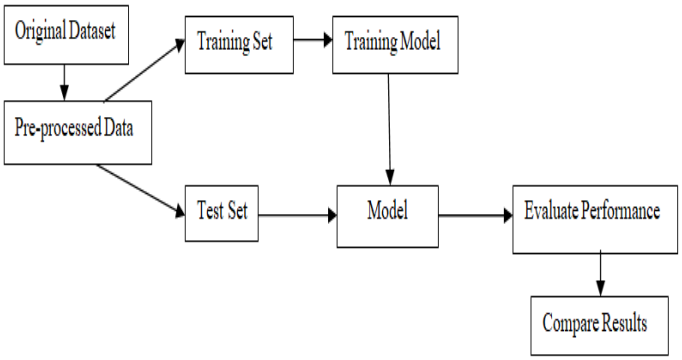

Fig 2 The proposed system's architecture

Module Description

The various modules are available

i) Data Collection ii) Pre-processing iii) Training iv) Testing v) Validation

Data Collection:

The information can be downloaded from the website of Kaggle. This is a collection of academic data collected from the Kalboard 360 learning management system (LMS). Kalboard 360 is an LMS multi-agent designed to facilitate training by using state-of - the-art software. The database is made up of 480 records of students and 16 features. There are 305 males and 175 females in the dataset. The characteristics are grouped into three main categories[17]:

1. Demographic features such as gender and nationality.
2. Academic background features such as educational stage, grade Level.
3. Behavioural features such as raised hand on class, opening resources, answering survey by parents, and school satisfaction.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

Table 1: Class Attribute Description

CLASS	DESCRIPTION	VALUE	COUNT
L	Low Level	Interval: contains a value between 0 and 69	127
M	Medium Level	Interval: contains a value between 70 and 89	211
H	High Level	Interval: contains a value between 90 and 100	143

Preprocessing:

Preprocessing is defined as "a data mining technique that involves transforming raw data into an understandable format" [11].

Since our data have specific (nominal, continuous) variables. Many algorithms of classification may not produce great results. We conduct preprocessing to recognize the noise, irrelevant information and remove the redundancy. To transform the data from categorical type to numeric type, we carried out data transformation on our results. On the attributes we applied Z-score standardization.

Z-score normalization: Z-score normalization is a data standardization technique that eliminates the outer problem. It always assumes that the data are normally distributed. We used standardization of Z-score because it fits well with outliers and manages them perfectly.

The standardization equation for Z-score is as follows:

$$Z-score=\frac{X-\mu}{\sigma} E.q...1$$

In the formula, terms are given by-

X- Value of the data record

μ- Mean of the sample

σ- Standard deviation of the sample

Feature Selection:

Selection of features may be a technique that focuses on reducing the quantity of attributes that seem within the patterns, reducing the spatial property of the feature area, eliminating redundant, irrelevant or noisy information and therefore the understandability of the mining results.[12, 13]. Moreover, in the choice of features, there have been different techniques; we have used the correlation-based function selection (CFS) methodology. CFS assesses the value of a subset of attributes by considering the individual predictive ability of each feature and the degree of redundancy between them. The correlation coefficients are used to estimate the association between the attribute and rank sub-sets, as well as the feature inter-correlations[12]. We consider the features which are highly correlated to the category attribute. We set the comparison of the minimum threshold as >= 0.3 and <= -0.3. The features selected are tabulated as follows:

Table 2 Selected Features and Their Correlation

SNO	Attribute Selected	Correlation Value
1	Relation	0.41
2	Raisedhands	0.62
3	VisITedResources	0.64
4	AnnouncementsView	0.5
5	Discussion	0.3
6	ParentAnsweringSurvey	0.41
7	ParentschoolSatisfaction	0.36
8	StudentAbsenceDays	-0.64

values

Training Module

We train the models in the training module and make the models in the data set know the patterns. In addition, the training unit is split into two sub-modules:

1) KNN Classifier

2) Naïve Bayes classifier

Training Sub-Module for KNN Classifier:

We implement the KNN classification algorithm in two different ways:

i) KNN without k-fold cross-validation:

we use 80/20 ratio for KNN classification, respectively, to train and check the dataset. We divide the dataset into two folders, the training

folder comprises 80% of the total records and the remaining 20% of the test data. Splitting is achieved through a code written in the language of python programming.

ii) KNN with 10 fold cross validation:

first we divided the entire dataset into 10 separate files each with 10% of the total number of records, i.e. 48 records per document, in 10 fold cross validation. This is implemented by another python-written script. We feed 9 of the 10 files into the framework for learning, which also has up to 432 records. The document left over is used to evaluate.

Training Sub-Module for Naïve Bayes Classifier:

Often implemented in two ways is the Naïve classifier:

i) Naïve Bayes without Feature Selection[9,14]:

We use the 80/20 proportion for NB Classifier to train and evaluate the dataset respectively. We split the entire dataset into two sub-sets, the training document contains 80% of the overall records and also the remaining 20% of the test file Splitting is achieved through some kind of code written in the language of python programming.

ii) Naïve Bayes with Feature Selection:

Apply Z-score normalization on the source data, the original dataset is converted. This transformed data set was used to train the algorithm, using the same ratio above.

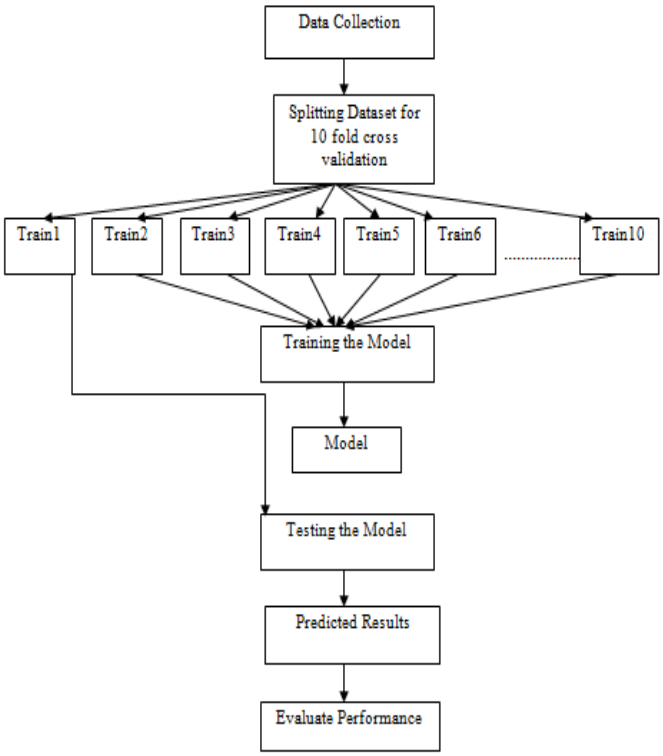

Fig 3 Cross-fold validation work-flow for KNN

Testing Module:

The model makes predictions on the test records in the test module. Secondly, the evaluation module is split into two sub-modules:

1) Testing Sub-Module For KNN Classifier

2) Testing Sub- Module For Naïve Bayes Classifier

Testing Sub-Module For KNN Classifier:

I) KNN without k-fold cross validation:

The test dataset has 20% of the actual dataset records i.e. 96. The 96 records has been split into five test files. There have been 20 records in each of the 4 test files and 16 records in the fifth test file Each sample file is checked one after the other and kept separate for both the output file.

ii)KNN with 10-fold cross-validation:

first we partition the whole dataset into 10 separate files each with 10 percent of the total number of records, i.e. 48 records each file, in 10 fold cross-validation. This will be implemented by another python-written script. We feeding 9 of the 10 files into the system for learning, which has equal to 432 records.

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

For evaluation, we test the model with 48 records containing a leftover single test file. For another use, the output file was saved. For different combinations of training files and test records, the above process will continue.

3.2.4.2 Testing Sub- Module For Naïve Bayes Classifier:

i) Naïve Bayes model without Feature Selection:

The test dataset has 20% records of the original dataset i.e., 96 records. We test model using test dataset and the Predictions of output have been stored in the output file

ii) Naïve Bayes model with Feature attribute Selection:

The test dataset includes 20% of the revised dataset records i.e., 96 records. We use the model to evaluate the test data and store the calculated performance predictions.

Validation Module:

We verify the results contained in output files in the validation module The validation process would be the same for KNN and Naïve Bayes models. The output file results are evaluated to test if the classifier has correctly predicted the students ' actual class and this is achieved through a python language written program. We also construct the confusion matrix without any of the predefined packages being used.

For the assessment of classification performance, we use five specific different measures: reliability, error rate, accuracy, recall, and F-measure. Accuracy is the percentage which is correctly determined for the total number of predictions. Precision is the proportion of cases correctly classified to the total number of cases misclassified and cases properly classified. Recall is the proportion of correctly categorized cases to the total number of cases that have not been identified and correctly classified. We also used the F-measure to combine the recall with the reliability that is considered a good predictor of their relationship[11].

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad E.q, \quad 2$$

$$Error-rate = \frac{FP+FN}{TP+TN+FP+FN} \quad E.q, \quad 3$$

$$Precision = \frac{TP}{TP+FP} \quad E.q, \quad 4$$

$$Recall = \frac{TP}{TP+FN} \quad E.q \quad 5$$

$$F-measure = 2\left(\frac{Precision * Recall}{Precision+Recall}\right) \quad E.q, \quad 6$$

IV. Implementation and Result Analysis:

Python 3.6 can be downloaded from [15 , 16] to implement the project and referenced [15, 16] for python programming and implemented the classification algorithms from scratch without using any predefined packages.

Table 2 Comparison of results

Evaluation Criteria	KNN classifier	KNN with cross validation	NBC	NBC with Z-score & feature selection
Accuracy	71	84.8	54.16	64.58
Error Rate	29	15.2	48.83	35.41
Precision	70.3	83.9	46.48	71.11
Recall	72.4	85.6	50	63.74
F-score	72.1	84.4	48.18	67.2

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

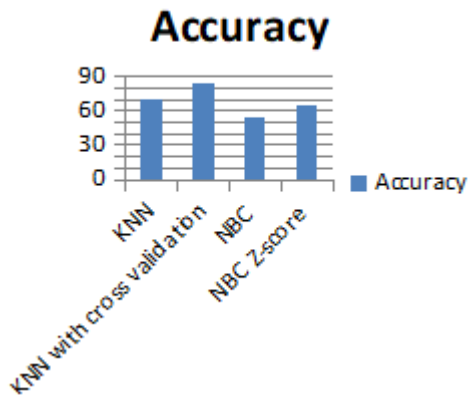

Fig4 Comparison of Accuracy

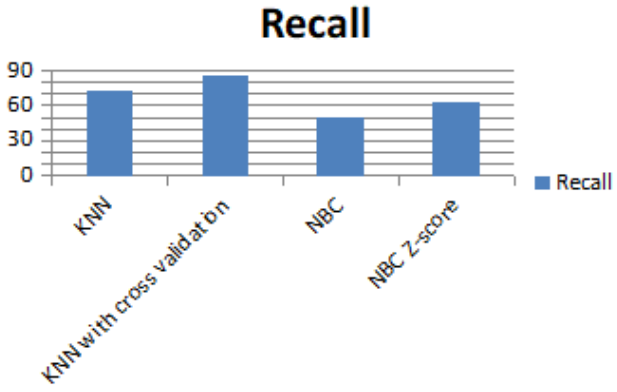

Fig 7 Comparison of Recall

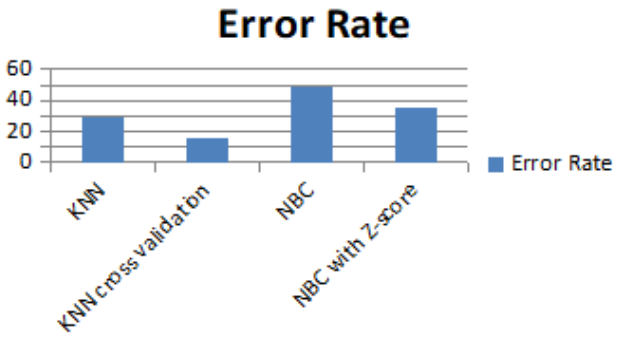

Fig 5 Comparison of Error rate

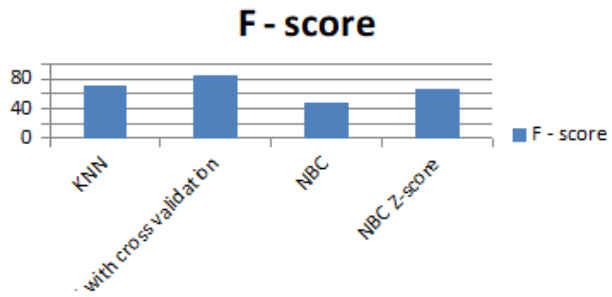

Fig 8 Comparison of F-score

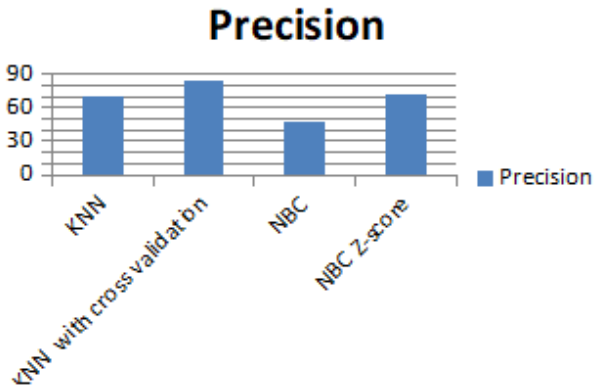

Fig 6 Comparison of Precision

V. Conclusions and Future Scope:

We presented a student performance analysis model on the Learning Management System dataset i.e., Kalboard 360 basically from Lebanon in this paper. Two classification algorithms have been implemented i.e. KNN classifier and Naive Bayes classifier. Using 10 fold cross-validation, the KNN classifier is implemented without cross-validation. Naive Bayes classifier was implemented with a choice of features on the original data and NBC on pre-processed data. Among these four, KNN achieves its highest accuracy with cross-validation, i.e. 84.8%, and Naive Bayes achieves the lowest accuracy, i.e. 54%. Although, NBC's reliability is increased when we use the pre-processed data feature collection and we got 64% accuracy which implies that the student's behavior and parent's participation is a key factor in correctly classifying the students.

January - February 2020
ISSN: 0193 - 4120 Page No. 7658 - 7665

Our Future work includes exploring other classification algorithms on a much more diverse dataset with different data mining techniques.

VI. References

[1] Baker RS(2010 May). "Data mining for education". International encyclopedia of education , 7(3), pp. 112-8.

[2] Baker RS(2014, May-June), "Educational Data Mining" , An Advance for Intelligent Systems in Education," in IEEE Intelligent Systems, 29(3), pp. 78-82,

[3] García, Enrique, Cristóbal Romero, Sebastián Ventura, and Carlos De Castro. (2011). "A collaborative educational association rule mining tool", The Internet and Higher Education, 14(2), pp. 77-88.

[4] C. Romero and S. Ventura(2010 NoV), "Educational Data Mining: A Review of the State of the Art", in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), pp. 601-618.

[5] Abazi-Bexheti, L., Apostolova-Trpkovska, M., & Kadriu, A. (2014, May). "Learning management systems: Trends and alternatives." In 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (pp. 773-777). IEEE.

[6] Amrieh, E.A., Hamtini, T. and Aljarah, I., (2015, November), "Preprocessing and analyzing educational data set using X-API for improving student's performance". In 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT) (pp. 1-5). IEEE.

[7] Roy, S., & Garg, A. (2017, October)." Predicting academic performance of student using classification techniques". In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON) (pp. 568-572). IEEE.

[8] Maulida, J. D., & Kariyam. (2017, December). "Students academic performance based on behavior", In AIP Conference Proceedings (Vol. 1911, No. 1, p. 020010). AIP Publishing.

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

[10] ABDULLAH, Manal Abdulaziz,(2015, March) ,"Learning Style Classification Based on Student's Behavior in Moodle Learning Management System", Transactions on Machine Learning and Artificial Intelligence, [S.l.], v. 3, n. 1, p. 28, ISSN 2054-7309.

[11] Han, Jiawei, Jian Pei, and Micheline Kamber. (2011), "Data mining: concepts and techniques", Elsevier, 3rd edition., ISBN 978-0-12-381479-1 1.

[12] Asha.G.Karegowda, A. S. Manjunath & M.A.Jayaram,(2010, December), "Comparative study of attribute selection using gain ratio and correlation based feature selection", International Journal of Information Technology and Knowledge Management, 2(2), pp.271-277.

[13] Blum, A. L., & Langley, P. (1997), "Selection of relevant features and examples in machine learning", Artificial Intelligence, 97(1-2), 245-271.

[14] Basnet, R. B., Sung, A. H., & Liu, Q. (2012, June)., "Feature selection for improved phishing detection." In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 252-261. Springer, Berlin, Heidelberg,

[15] Müller, Andreas C., and Sarah Guido.(2016), "Introduction to machine learning with Python: a guide for data scientists.", O'Reilly Media, Inc.

[16] G. van Rossum, (1995) "Python tutorial, Technical Report", CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, May 1995.

[17] "Kalboard360-E-learning system", http://kalboard360.com/ (accessed February 28, 2016)