

Mining Educational Data to Predict Student's academic Performance using Ensemble Methods

*Elaf Abu Amrieh¹, Thair Hamtini² and Ibrahim Aljarah³

^{1,2,3} Computer Information Systems Department

^{1,2,3} The University of Jordan

¹ilef.kram@hotmail.com, ²thamtini@ju.edu.jo, ³i.aljarah@ju.edu.jo

Abstract

Educational data mining has received considerable attention in the last few years. Many data mining techniques are proposed to extract the hidden knowledge from educational data. The extracted knowledge helps the institutions to improve their teaching methods and learning process. All these improvements lead to enhance the performance of the students and the overall educational outputs. In this paper, we propose a new student's performance prediction model based on data mining techniques with new data attributes/features, which are called student's behavioral features. These type of features are related to the learner's interactivity with the e-learning management system. The performance of student's predictive model is evaluated by set of classifiers, namely; Artificial Neural Network, Naïve Bayesian and Decision tree. In addition, we applied ensemble methods to improve the performance of these classifiers. We used Bagging, Boosting and Random Forest (RF), which are the common ensemble methods used in the literature. The obtained results reveal that there is a strong relationship between learner's behaviors and their academic achievement. The accuracy of the proposed model using behavioral features achieved up to 22.1% improvement comparing to the results when removing such features and it achieved up to 25.8% accuracy improvement using ensemble methods. By testing the model using newcomer students, the achieved accuracy is more than 80%. This result proves the reliability of the proposed model.

Keywords: Student academic performance, Educational Data Mining, E-learning, Ensemble, knowledge discovery, ANN Model

1. Introduction

Recently there is an increasing research interest in educational data mining (EDM). EDM is an emerging field that uses data-mining (DM) techniques to analyze and extract the hidden knowledge from educational data context [1]. EDM includes different groups of users, these users utilize the knowledge discovered by EDM according to their own vision and objectives of using DM [2]. For example, the hidden knowledge can help the educators to improve teaching techniques, to understand learners, to improve learning process and it could be used by learner to improve their learning activities [3]. It also helps the administrator taking the right decisions to produce high quality outcomes [4]. The educational data can be collected from different sources such as web-based education, educational repositories and traditional surveys. EDM can use different DM techniques, each technique can be used for certain educational problem. As Example, to predict an educational model the most popular technique is classification. There are several algorithms under classification such as Decision tree, Neural Networks and Bayesian networks [5].

This paper introduces a students' performance model with a new category of features, which called behavioral features. The educational dataset is collected from learning management system (LMS) called Kalboard 360 [6]. This model used some data mining

techniques to evaluate the impact of student's behavioral features on student academic performance. Furthermore, we try to understand the nature of this kind of features by expanding data collection and preprocessing steps. The data collection process is accomplished using a learner activity tracker tool, which is called experience API (xAPI). The collected features are classified into three categories: demographic features, academic background features and behavioral features. The behavioral features are a new feature category that is related to the learner experience during educational process.

To the best of our knowledge, this is the first work that employs this type of features/attributes. After that, we use three of the most common data mining methods in this area to construct the academic performance model: Artificial Neural Network (ANN) [30], Decision Tree [28], and Naïve Bayes [32]. Then, we applied ensemble methods to improve the performance of such classifiers. The ensembles used to improve the performance of student's prediction model are Bagging, Boosting and Random Forest (RF). The remainder of this paper is organized as follows: Section 2 presents the related work in the area of educational data mining algorithms. In Section 3, presents the data collection and preprocessing. In Section 4 our methodology in predicting students' performance. The experimental evaluation and results are shown in Section 5, and Section 6 presents our conclusions.

2. Related Work

Predicting student's performance is an important task in web-based educational environments. To build a predictive model, there are several DM techniques used, which are classification, regression and clustering. The most popular technique to predict students' performance is classification. There are several methods under classification such as Decision Tree (DT), Artificial Neural Networks (ANN) and Naive Bayes (NB).

Decision tree is a set of conditions arranged in a hierarchical frame. Most of researchers used this technique due to their simplicity, in which it can be transformed into a set of classification rules. Some of the famed DT algorithms are C4.5 [28] and CART. Romero *et al* in [29] used DT algorithm to predict students' final marks based on their usage data in the Moodle system. Moodle is one of the frequently used Learning Content Management Systems (LCMS). The author has collected real data from seven Moodle courses with Cordoba University to classify students into two groups: passed and fail. The objective of this research is to classify students with equal final marks into different groups based on the activities carried out in a web-based course.

Neural network is another popular technique that has been used in educational data mining. A neural network is a biological inspired intelligent technique that consists of connected elements called neurons that work together to produce an output function [30]. Arsad *et al.* in [31] used ANN model to predict the academic performance of bachelor degree engineering students. The study takes Grade Point (GP) of fundamental subjects scored by the students as inputs without considering their demographic background, while it takes Cumulative Grade Point Average (CGPA) as output. Neural Network (NN) trains engineering Degree students GP to get the targeted output. This research showed that fundamental subjects have a strong influence in the final CGPA upon graduation.

The authors in [32] used Bayesian networks to predict the CGPA based on applicant background at the time of admission. Nowadays, educational institutions need a method to evaluate the qualified applicants graduating from various institutions. This research presents a novel approach that integrate a case-based component with the prediction model. The case-based component retrieves the past student most similar to the applicant being evaluated. The challenge is to define similarity of cases (applicants) in a way that is consistent with the prediction model. This technique can be applied at any institution that has a good database of student and applicant information.

In summary, various researches have been investigated to solve the educational problems using data mining techniques. However, very few researches shed light on student's behavior during learning process and its impact on the student's academic success. This research will focus on the impact of student interaction with the e-learning system. Furthermore, the extracted knowledge will help schools to enhance student's academic success and help administrators in improve learning systems.

3. Data Collection and Preprocessing

The increase of internet using in education has produced a new context known as web-based education or learning management system (LMS). The LMS is a digital framework that manage and simplify online learning [7]. The main purpose of the LMS is to manage learners, monitor student participation, keeping track of their progress across the system [8]. The LMS allocates and manages learning resources such as registration, classroom and the online learning delivery. In this paper, the educational data set is collected from learning management system (LMS) called Kalboard 360 Kalboard [6]. Kalboard 360 is a multi-agent LMS, which has been designed to facilitate learning through the use of leading-edge technology. Such system provides users with a synchronous access to educational resources from any device with Internet connection. In addition to involve parents and school management in the learning experience. This makes it a truly extensive process, which connects and properly engages all parties. The data is collected using a learner activity tracker tool, which called experience API (xAPI) [9]. The xAPI is a component of the Training and Learning Architecture (TLA) that enables to monitor learning progress and learner's actions like reading an article or watching a training video. The Experience API helps the learning activity providers to determine the learner, activity and objects that describe a learning experience.

The goal of X-API in this research is to monitor student behavior through the educational process for evaluating the features that may have an impact on student's academic performance. The educational data set that used in the previous work [10] contains only 150 student's records with 11 features. In the current paper that data set extends into 500 students with 16 features. The features are classified into three main categories: (1) Demographic features such as gender and nationality. (2) Academic background features such as educational Stage, grade Level and section. (3) Behavioral features, such as raised hand on class, visited resources, parent Answering Survey and Parent School Satisfaction. This feature cover learner and parent progress on LMS. Table 1 shows the dataset's attributes/features and their description. Table 1 was used in the previous research [10], by reviewing the table we can notice a new feature category which is a behavioral feature. These features present the learner and the parent participation in the learning process.

Table 1. Student Features and their Description

Features Category	Feature	Description
Demographical Features	Nationality	Student nationality
	Gender	The gender of the student (female or male)
	Place of Birth	Place of birth for the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Parent responsible for student	Student's parent as (father or mum)
Academic Background Features	Educational Stages (school levels)	Stage student belongs such as (primary, middle and high school levels)
	Grade Levels	Grade student belongs as (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11, G-12)
	Section ID	Classroom student belongs as (A, B, C)
	Semester	School year semester as (First or second)
	Topic	Course topic as (Math, English, IT, Arabic, Science, Quran)
	Student Absence Days	Student absence days (Above-7, Under-7)
Parents Participation on learning process	Parent Answering Survey	Parent is answering the surveys that provided from school or not.
	Parent School Satisfaction	This feature obtains the Degree of parent satisfaction from school as follow (Good, Bad)
Behavioral Features	Discussion groups	Student Behavior during interaction with Kalboard 360 e-learning system.
	Visited resources	
	Raised hand on class	
	Viewing announcements	

After the data collection task, we apply some pre-processing mechanisms to improve the quality of the data set. Data pre-processing is considered an important step in the knowledge discovery process, which includes data cleaning, feature selection, data reduction and data transformation.

3.1. Feature Analysis

There are many features affecting the student performance. This section will use the previous works to identify the important features in predicting students' performance. For the gender differences feature, biologicals confirm that there are differences in the aptitudes of students that depend on gender [11]. Meit in [12] found that most of female students have a positive learning style in compare to male students. The authors in [13] prove that female students are more satisfied than male students with e-learning systems. Other researches address that male students have a positive perception of e-learning compared to female students [14]. For the family background feature, different studies have shown that there is a positive relationship between the parent's education and student's performance [16]. This relation is particularly valid when the learner is being followed up by their mother. The authors in [17] observed that mothers have a more influence on their children academic achievements. Third school attendance feature, school attendance is an important feature in educational success [18]. Previous research [19] has shown a direct relation between good attendance and student achievement. These researches prove the positive relation between such features: gender, family background and school attendance students' performance. This research will shed a light on new category of features, called behavioral features. This feature related to the learner engagement with educational system. Student engagement is one of the main researches in educational psychology field. Student engagement was defined by Gunuc and Kuzu [20] as "the quality and quantity of students' psychological, cognitive, emotional and behavioral reactions to the learning process as well as to in-class/out-of-class academic and social activities to achieve successful learning outcomes". Kuk [21] refers to the student's engagement by the spent time in classroom. According to Stovall [22], student engagement includes not only the spent time on tasks but also their desire to participate in some activities. There are various researches that light on student's engagement and behavior. All of these researches confirm the positive relationship between students' behavior and student's academic achievement.

3.2. Data Preprocessing

This section will intensively talk about the data preprocessing. Data preprocessing is the step before applying data mining algorithm, it transforms the original data into a suitable shape to be used by a particular mining algorithm. Data preprocessing includes different tasks as data cleaning, feature selection and data transformation [23].

3.2.1. Data Visualization

Data visualization is an important preprocessing task, which used graphical representation to simplify and understand complex data. Visualization techniques have been recently used to visualize online learning aspects. Instructors can utilize the graphical representations to understand their learners better and become aware of what is occurring in the distance classes. This research visualizes the current data set using Weka tool. As shown in Figure1, the data set is visualized based on gender feature into 305 males and 175 females.

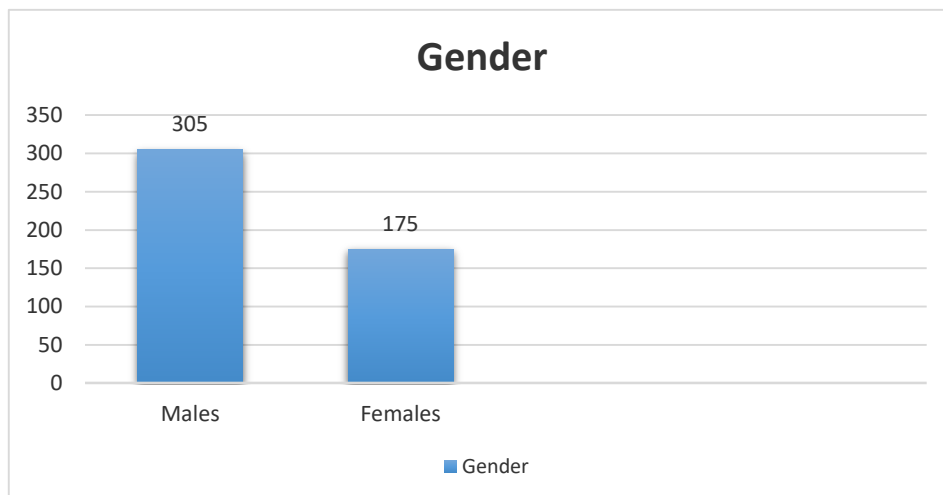


Figure 1. Gender Feature Visualization

As shown in Figure2, students come from different origins such as 179 students are from Kuwait, 172 students are from Jordan, 28 students from Palestine, 22 students are from Iraq, 17 students from Lebanon, 12 students from Tunis, 11 students from Saudi Arabia, 9 students from Egypt, 7 students from Syria, 6 students from USA, Iran and Libya, 4 students from Morocco and one student from Venezuela.

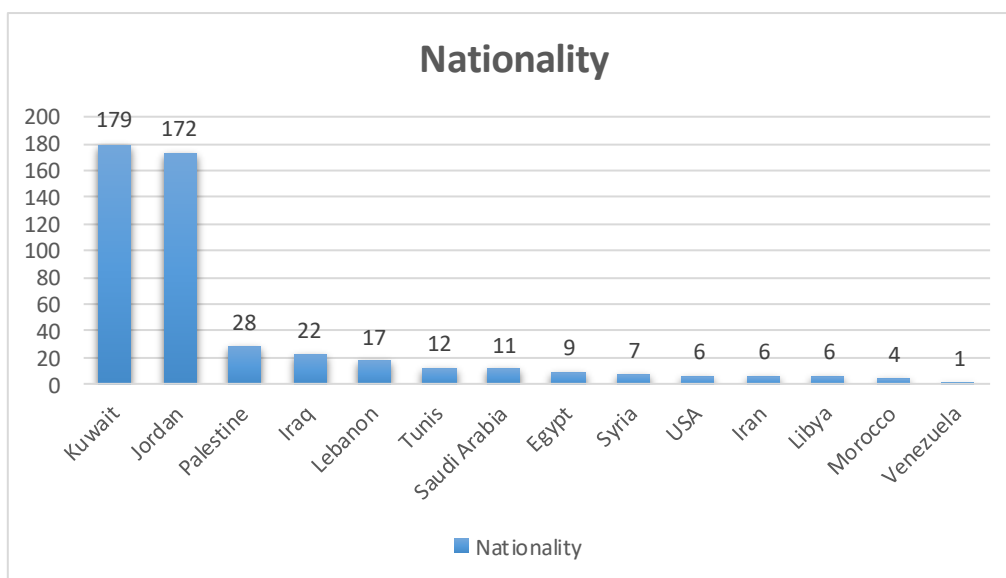


Figure 2. Gender Feature Visualization

According to the diversity of nationalities, we can conclude a hidden impact of such diversity on student's performance. As shown in Figure3, students are partitioned into three educational stages as follow: 199 students in the lower level, 248 students in the Middle level, and 33 students in the High level. Students are divided into three sections as follow: 283 students from Section A, 167 students from Section B and 30 students from Section C.

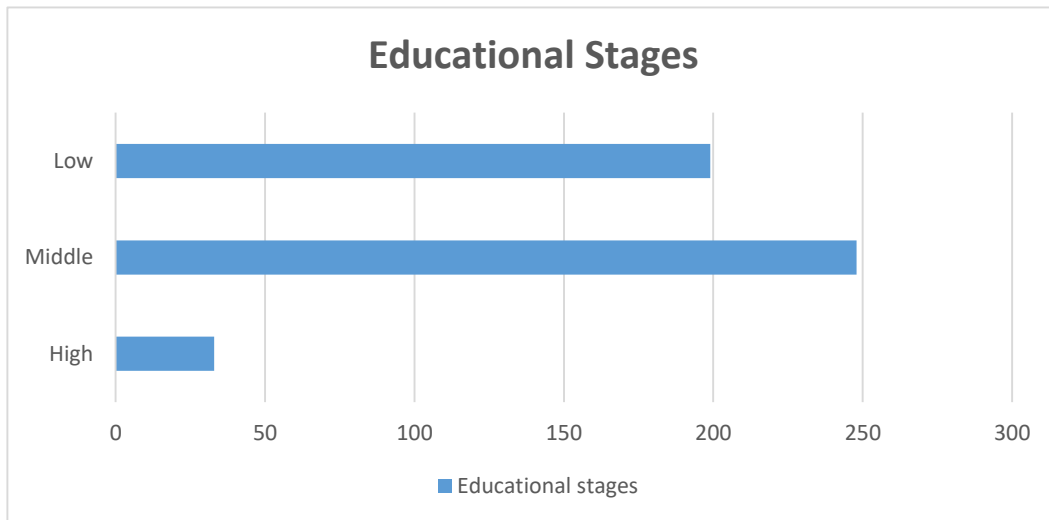


Figure 3. Educational Stages Visualization

The student's data collected through two educational semesters: First and second, in which 245 students record collected during the first semester and 235 student's record collected during the second semester. Students through these different semesters take different topics as shown in Figure4, There are 95 students take IT topic, 65 students take French topic, 59 students take Arabic topic, 51 students take science topic, 45 students take English topic, 30 students take Biology, 25 students take Spanish, 24 students take both chemistry and Geology topics, 22 students take Quran topic, 21 students take math topic, 19 students take History topic. Each student in the data set is followed up by a different parent as follow: 283 students are followed by their fathers and 197 students are followed by their moms.

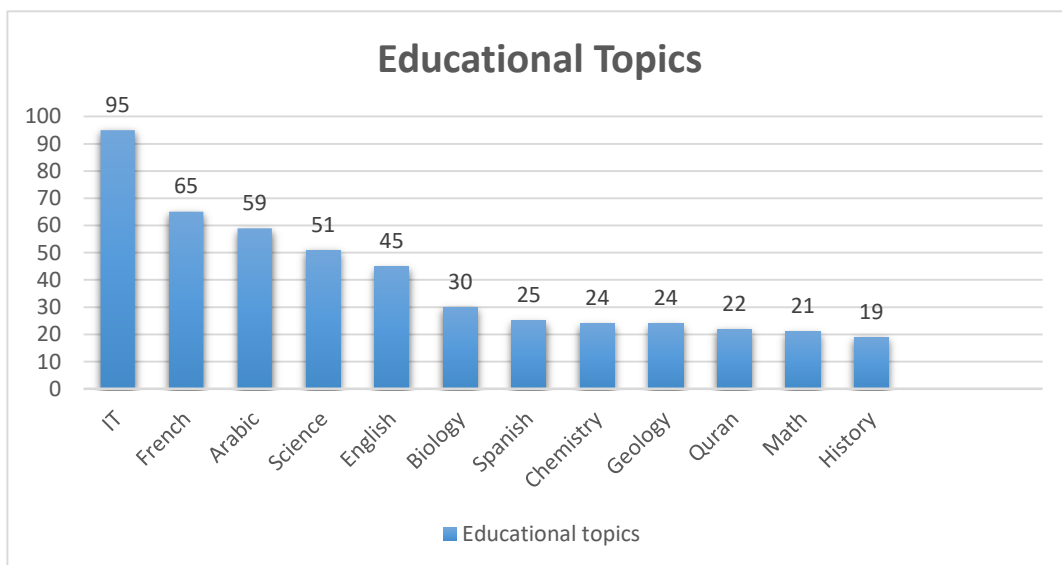


Figure 4. Educational Topics Visualization

The data set includes also the school attendance feature, as shown in Figure5, the students are visualized into two categories based on their absence days: 191 students exceed 7 absence days and 289 students their absence days under 7.

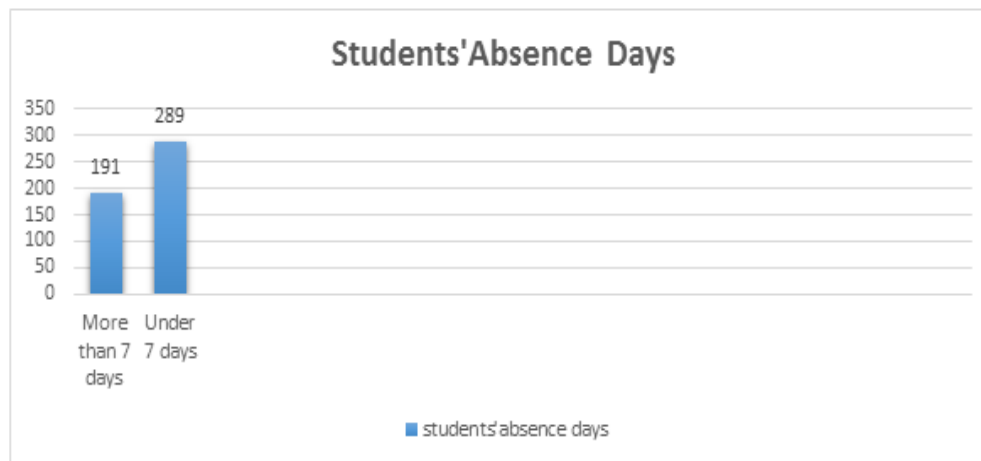


Figure 5. Students' Absence Days' Feature Visualization

This research uses the “student absence days” feature to show the influence of such feature on student’s performance. This research also utilizes new category of features; this feature is parent participation in the educational process. Parent participation feature has two sub features: Parent Answering Survey and Parent School Satisfaction. There are 270 of the parents answered survey and 210 are not, 292 of the parents are satisfied from the school and 188 are not. Data preprocessing used in this research to study the nature of students’ performance features, and to get the influence ratio of features by defining the percentage value of each feature. The influence ratio of features will be defined accurately using feature selection process.

3.2.2. Data Cleaning

Data cleaning is one of the main preprocessing tasks, is applied on this data set to remove irrelevant items and missing values. The data set contains 20 missing values in various features from 500 records, the records with missing values are removed from the data set, and the data set after cleaning becomes 480 records.

3.2.3. Feature Selection

Feature selection is a fundamental task in data preprocessing area. The objective of feature selection process is to select an appropriate subset of features which can efficiently describe the input data, reduces the dimensionality of feature space, removes redundant and irrelevant data [24]. This process can play an important role in improving the data quality therefore the performance of the learning algorithm. Feature selection methods are categorized into wrapper-based and filter-based methods. Filter method is searching for the minimum set of relevant features while ignoring the rest. It uses variable ranking techniques to rank the features where the highly ranked features are selected and applied to the learning algorithm. Different feature ranking techniques have been proposed for feature evaluations such as information gain and gain ratio.

In this research, we applied filter-method using information gain based selection algorithm to evaluate the feature ranks, checking which features are most important to build students’ performance model. Figure6, shows the feature ranks after filter-based evaluation. During feature selection, each feature assigned a rank value according to their influence on data classification. The highly ranked features are selected while others are excluded.

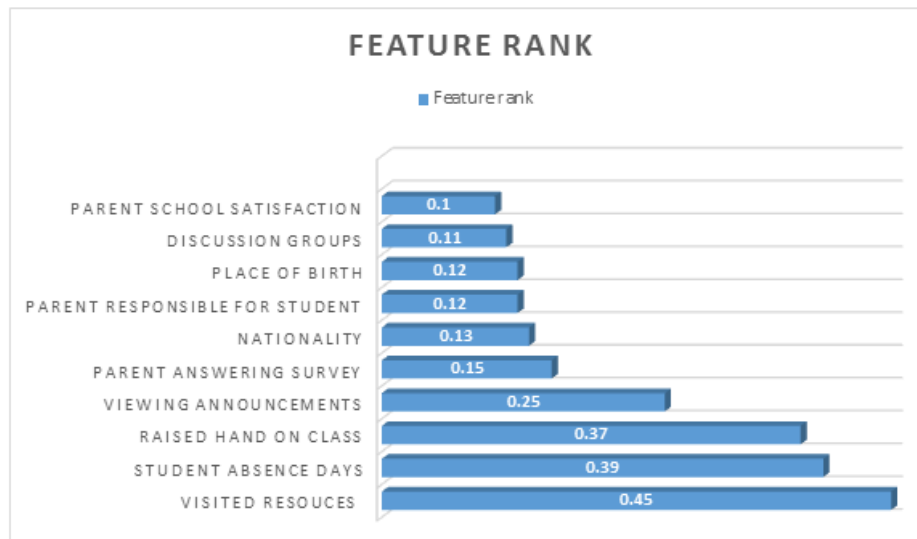


Figure 6. Filter-Based Feature Selection Evaluation

As shown in Figure6, visited resources feature got the higher rank, then followed by student absence days, raised the hand on classroom, parent answering survey, nationality, parent responsible for student, place of birth, discussion groups and parent school satisfaction features. As we can see the appropriate subset of features consist of ten features while other ones are excluded. In summary, the features that are related to student and parent progress during the usage of LMS got the highest ranks, which means the learner behavior during the educational process have an impact on their academic success.

4. Methodology

In this paper, we introduce a student's performance model using ensemble methods. Ensemble methods is a learning approach that combines multiple models to solve a problem. In contrast to traditional learning approaches which train data by one learning model, ensemble methods try to train data using a set of models, then combine them to take a vote on their results. The predictions made by ensembles are usually more accurate than predictions made by a single model. The aim of such approach is to provide an accurate evaluation for the features that may have an impact on student's academic success. Figure 7 shows the main steps in the proposed methodology.

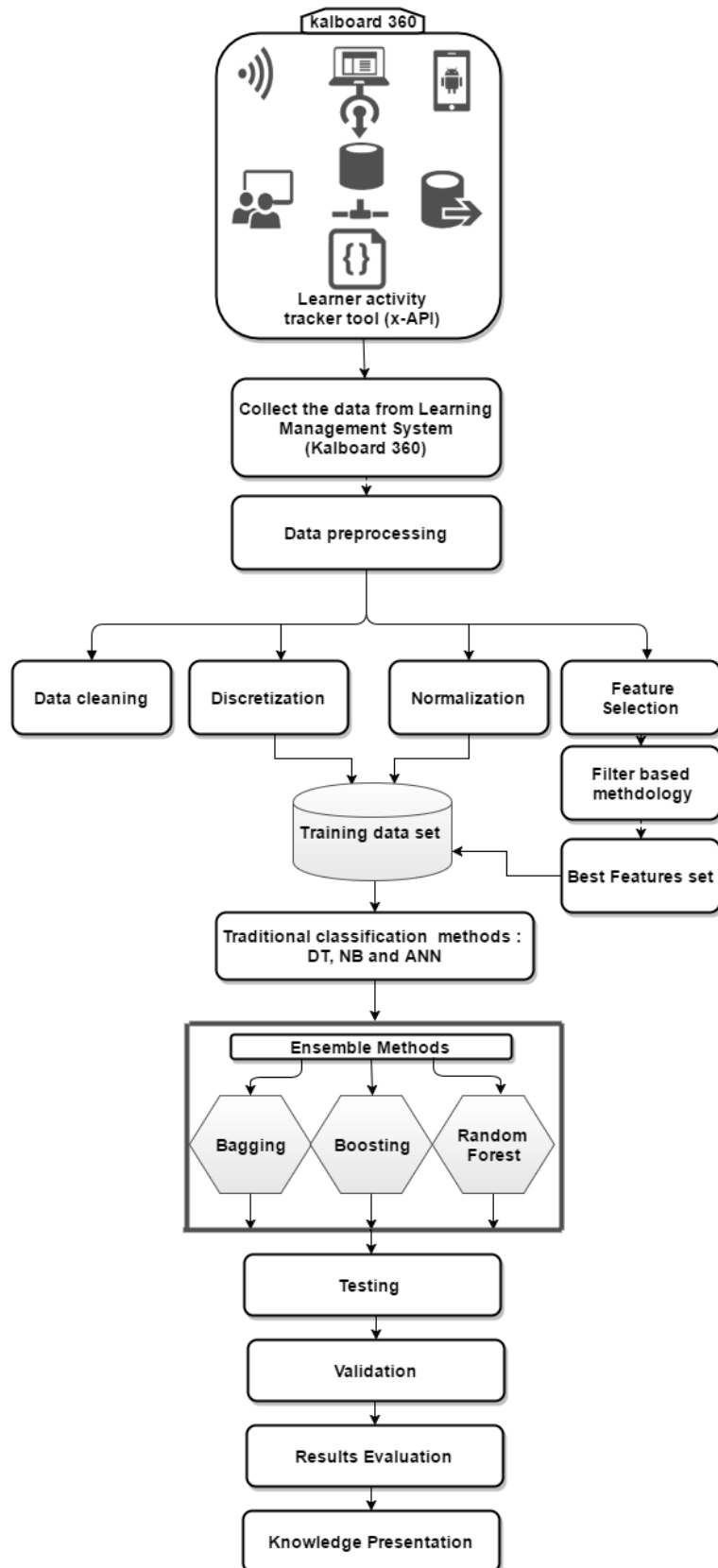


Figure 7. Student's Performance Prediction Model Research Steps

This methodology starts by collecting data from Kalboard 360 (LMS) system using experience API (xAPI) as mentioned in Section 3. This step is followed by data preprocessing step, which concerns with transforming the collected data into a suitable format. After that, we use discretization mechanism to transform the students' performance from numerical values into nominal values, which represents the class labels of the classification problem. To accomplish this step, we divide the data set into three nominal intervals (High Level, Medium Level and Low Level) based on student's total grade/mark such as: Low Level interval includes values from 0 to 69, Middle Level interval includes values from 70 to 89 and High Level interval includes values from 90-100. The data set after discretization consists of 127 students with Low Level, 211 students with Middle Level and 142 students with High Level. Then, we use normalization to scale the attributes values into a small range [0.0 to 1.0]. This process can speed up the learning process by preventing attributes with large ranges from outweighing attributes with smaller ranges. After that, feature selection process is applied to choose the best feature set with higher ranks. As shown in Figure7, we applied filter-based technique for feature selection.

In this paper, ensemble methods are applied to provide an accurate evaluation for the features that may have an impact on the performance/grade level of the students, and to improve the performance of student's prediction model. Ensemble methods are categorized into dependent and independent methods. In a dependent method, the output of a learner is used in the creation of the next learner. Boosting is an example of dependent methods. In an independent method, each learner performs independently and their outputs are combined through a voting process. Bagging and random forest are example of independent methods. These methods resample the original data into samples of data, then each sample will be trained by a different classifier. The classifiers used in student's prediction model are Decision Trees (DT), Neural Networks (NN) and Naïve Bayesian (NB). Individual classifiers results are then combined through a voting process, the class chosen by most number of classifiers is the ensemble decision.

Boosting belongs to a family of algorithms that are capable of converting weak learners to strong learners. The general boosting procedure is simple, it trains a set of learners sequentially and combine them for prediction, then focus more on the errors of the previous learner by editing the weights of the weak learner. A specific limitation of boosting that is used only to solve binary classification problems. This limitation is eliminated with the AdaBoost algorithm. AdaBoost is an example of boosting algorithm, which stands for adaptive boost. The main idea behind this algorithm is to pay more attention to patterns that are hard to classify. The amount of attention is measured by a weight that is assigned to every subset in the training set. All the subsets are assigned equal weights. In each iteration, the weights of misclassified instances are increased while the weights of truly classified instances are decreased. Then the AdaBoost ensemble combines the learners to generate a strong learner from weaker classifiers through a voting process [33].

Bagging is an independent ensemble based methods. The aim of this method is to increase the accuracy of unstable classifiers by creating a composite classifier, then combine the outputs of the learned classifiers into a single prediction. The Bagging algorithm is summarized in Figure8, it starts with resampling the original data into different training data sets (D1-Dn) which called bootstraps, each bootstrap sample size is equal to the size of the original training set. All bootstrap samples will be trained using different classifiers (C1-Cm). Individual classifiers results are then combined through majority vote process, the class chosen was by the most number of classifiers is the ensemble decision [33].

In boosting, as contrary to bagging, each classifier is influenced by the performance of the previous classifier. In bagging, each sample of data is chosen with equal probability, while in boosting, instances are chosen with a probability that is proportional to their

weight. Furthermore, bagging works best with high variance models which produce variance generalization behavior with small changes to the training data. Decision trees and neural networks are examples of high variance models.

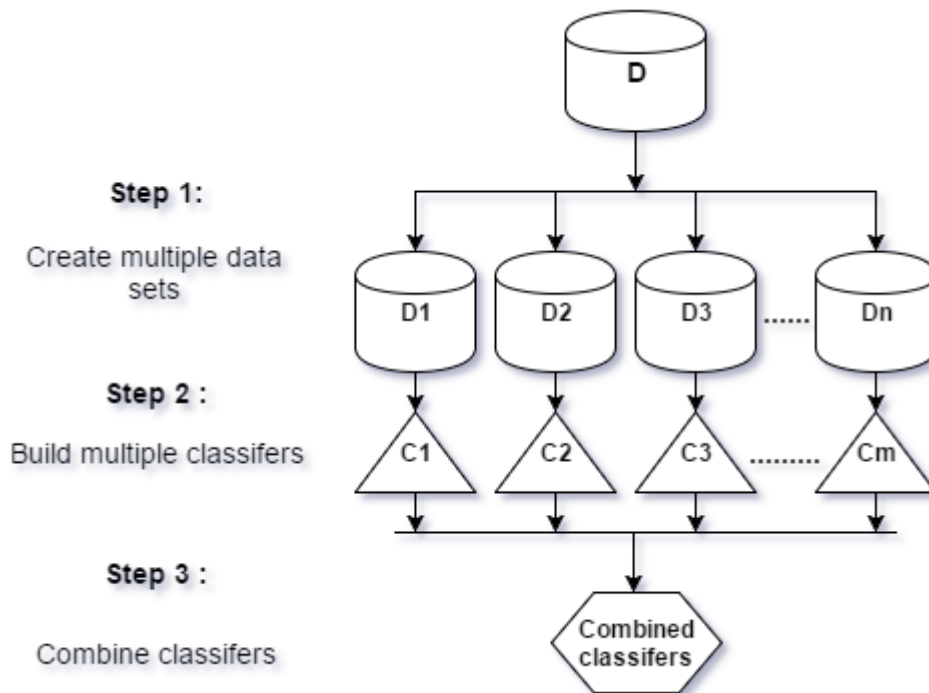


Figure 8. The General Bagging Procedure

Random Forest (RF) is a special modification of bagging where the main difference with bagging is the integration of randomized feature selection. Through the decision tree construction process, RF uses random decision trees to select a random subset of features. Notice that randomness is only performed on the feature selection process, but the choice of a split point on the selected features is performed by bagging. The combination between decision tree and bootstrapping makes RF strong enough to overcome the overfitting problem, and to reduce the correlation between trees which provides an accurate prediction [33].

All the above classification methods are trained using 10-folds cross validation. This technique divides the data set into 10 subsets of equal size, nine of the subsets are used for training, while one is left out and used for testing. The process is iterated for ten times, the final result is estimated as the average error rate on test examples. Once the classification model has been trained, the validation process starts. Validation process is the last phase to build a predictive model, it used to evaluate the performance of the prediction model by running the model over real data.

5. Experiments and Results

5.1. Environment

We ran the experiments on the PC containing 6GB of RAM, 4 Intel cores (2.67GHz each). For our experiments, we used WEKA [25] to evaluate the proposed classification models and comparisons. Furthermore, we used 10-fold cross validation to divide the dataset into training and testing partitions.

5.2. Evaluation Measures

In our experiments, we use four common different measures for the evaluation of the classification quality: Accuracy, Precision, Recall and F-Measure [26, 27]. Measures calculated using Table 2, which shows classification confusion matrix based on the Equations 1, 2, 3 and 4, respectively.

Table 2. Confusion Matrix

		Detected	
		Positive	Negative
Actual	Positive	True positive (TP)	False Negative(FN)
	Negative	False Positive (FP)	True Negative (TN)

Accuracy is the proportion of the total number of predictions where correctly calculated.

Precision is the ratio of the correctly classified cases to the total number of misclassified cases and correctly classified cases. Recall is the ratio of correctly classified cases to the total number of unclassified cases and correctly classified cases. In addition, we used the F-measure to combine the recall and precision which is considered a good indicator of the relationship between them [27].

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

$$F_c = 2 \frac{\text{Precision}_c * \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \quad (4)$$

5.3. Evaluation Results

5.3.1. Evaluation Results Using Traditional DM Techniques

There are many features directly or indirectly affecting the effectiveness of student performance model. In this section, we will evaluate the impact of behavioral features on student's academic performance using different classification techniques such as (DT, ANN and NB). After applying the classification techniques on the data set, the results are distinct based on different data mining measurements. Table 3, shows the classification results using several classification algorithms (ANN, NB and DT). Each classifier introduces two classification results: (1) classification results with student's behavioral features (BF) and (2) classification results without behavioral features (WBF).

Table 3. Classification Method Results with Behavioral Features (BF) and Results without Behavioral Features (WBF)

Evaluation Measure	DT (J48)		ANN		NB	
Behavioral features existence	BF	WBF	BF	WBF	BF	WBF
Accuracy	75.8	55.6	79.1	57.0	67.7	46.4
Recall	75.8	55.6	79.2	57.1	67.7	46.5
Precision	76.0	56.0	79.1	57.2	67.5	46.8
F-Measure	75.9	55.7	79.1	57.1	67.1	46.4

As shown in Table 3, we can notice that the ANN model outperforms other data mining techniques. ANN model achieved 79.1 accuracy with BF and 57.0 without behavioral features. The 79.1 accuracy means that 380 of 480 students are correctly classified to the right class labels (High, Medium and Low) and 100 students are incorrectly classified.

For the recall measure, the results are 79.2 with BF and 57.1 without behavioral features. The 79.2 recall means that 380 students are correctly classified to the total number of unclassified and correctly classified cases.

For the precision measure, the results are 79.1 with BF and 57.2 without behavioral features. The 79.1 precision means 380 of 480 students are correctly classified and 100 students are misclassified.

For the F-Measure, the results are 79.1 with BF and 57.1 without behavioral features. The experimental results prove the strong effect of learner behavior on student's academic achievement. We can get more accurate results by training the data set with ensemble methods.

Evaluation Results Using Ensemble Methods

In this section, we applied ensemble methods to improve the evaluation results of traditional DM methods. Table 3, presents the results of the traditional classifiers and the results of traditional classifiers using ensemble methods (Bagging, Boosting and RF).

As shown in in the Table 3, we can see good results using ensemble methods with traditional classifiers (ANN, NB and DT). Each ensemble trains the three classifiers, then combine the results through a majority voting process to achieve the best prediction performance of student's model. Boosting method outperform other ensembles methods, in which the accuracy of DT using boosting is improved from 75.8 to 77.7, which means that the number of correctly classified students are increased from 363 to 373 of 480. Recall results are increased from 75.8 to 77.7, which means that 373 students are correctly classified to the total number of unclassified and correctly classified cases. Precision results are also increased from 76.0 to 77.8, which means 373 of 480 students are correctly classified and 107 students are misclassified.

Table 4. Classification Method Results Using Ensemble Methods

Evaluation Measure	Traditional classification methods			Bagging			Boosting			Random Forest
Classifiers type	DT	ANN	NB	DT	ANN	NB	DT	ANN	NB	DT
Accuracy	75.8	79.1	67.7	75.6	78.9	67.2	77.7	79.1	72.2	75.6
Recall	75.8	79.2	67.7	75.6	79.0	67.3	77.7	79.2	72.3	75.6
Precision	76.0	79.1	67.5	75.7	78.9	67.1	77.8	79.1	72.4	75.6
F-Measure	75.9	79.1	67.1	75.6	78.9	66.7	77.7	79.1	71.8	75.5

Boosting also achieved a noticeable improvement with NB model, in which the accuracy of NB using boosting increased from 67.7 to 72.2, which means the number of correctly classified students increased from 324 to 346 of 480 students. Recall results increased from 67.7 to 72.3, which means that 347 students are correctly classified to the total number of unclassified and correctly classified cases. Precision results are also increased from 67.5 to 72.4, which means 347 of 480 students are correctly classified. ANN model performance using boosting method is not differed much from ANN model results without boosting. Once the classification model has been trained using 10-folds cross validation, the validation process starts. Validation is an important phase in building predictive models, it determines how realistic the predictive models are. In this research, the model is trained using 500 students and the model is validated using 25 newcomer students. In validation, the data set contains unknown labels to evaluate the reliability of the trained model. Table 5, shows the evaluation results using several classification methods (ANN, NB and DT) through testing process and validation process.

Table 5. Classification Methods Results through Testing and Validation

Evaluation Measure	Testing results			Validation results		
Classifiers type	DT	ANN	NB	DT	ANN	NB
Accuracy	75.8	79.1	67.7	82.2	80.0	80.0
Recall	75.8	79.2	67.7	82.2	80.0	80.0
Precision	76.0	79.1	67.5	85.0	84.7	83.8
F-Measure	75.9	79.1	67.1	81.8	79.2	80.2

As shown in Table 5, we can notice that the evaluation measure results increased for the three prediction models through validation process. The three prediction models achieved accuracy more than 80%, which means that 20 of 25 new students are correctly classified to the right class labels (High, Medium and Low) and 5 students are incorrectly classified. The results of the validation process prove the reliability of the proposed model.

6. Conclusion

Academic achievement is being a big concern for academic institutions all over the world. The wide use of LMS generates large amounts of data about teaching and learning interactions. This data contains hidden knowledge that could be used to enhance the academic achievement of students. In this paper, we propose a new student's performance prediction model based on data mining techniques with new data attributes/features, which called student's behavioral features. These type of features are related to the learner interactivity with learning management system. The performance of student's predictive model is evaluated by set of classifiers, namely; Artificial Neural Network, Naïve Bayesian and Decision tree. In addition, we applied ensemble methods to improve the performance of these classifiers. We used Bagging, Boosting and Random Forest (RF), which are the common ensemble methods that used in the literature. The obtained results reveal that there is a strong relationship between learner's behaviors and their academic achievement. The accuracy of student's predictive model using behavioral features achieved up to 22.1% improvement comparing to the results when removing such features, and it achieved up to 25.8% accuracy improvement using ensemble methods. The visited resources feature is the most effective behavioral feature on students' performance model. In our future work, we will focus more on analyzing this kind of

feature. After completing the training process, the predictive model is tested using unlabeled newcomer students, the achieved accuracy is more than 80%. This result proves how realistic the predictive model is. Lastly, this model can help educators to understand learners, identify weak learners, to improve learning process and trimming down academic failure rates. It also can help the administrators to improve the learning system outcomes.

References

- [1] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", *Expert systems with applications*, vol. 33, no. 1, (2007), pp. 135-146.
- [2] M. Hanna, "Data mining in the e-learning domain", *Campus-wide information systems*, vol. 21, no. 1, (2004), pp. 29-34.
- [3] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art. Systems, Man, and Cybernetics", Part C: Applications and Reviews, *IEEE Transactions on*, vol. 40, no. 6, (2010), pp. 601-618.
- [4] M. E. Zorrilla, E. Menasalvas, D. Marin, E. Mora and J. Segovia, "Web usage mining project for improving web-based learning sites", In *Computer Aided Systems Theory–EUROCAST 2005*, Springer Berlin Heidelberg, (2005), pp. 205-210.
- [5] A. M. Shahiri and W. Husain, "A Review on Predicting Student's Performance Using Data Mining Techniques", *Proceeding Computer Science*, vol. 72, (2015), pp. 414-422.
- [6] "Kalboard360-E-learning system", <http://kalboard360.com/> (accessed February 28, 2016).
- [7] G. Kakasevski, M. Mihajlov, S. Arsenovski and S. Chungurski, "Evaluating usability in learning management system Moodle", *Information Technology Interfaces*, 2008. ITI 2008. 30th International Conference on IEEE, (2008), pp. 613-618.
- [8] S. Rapuano and F. Zoino, "A learning management system including laboratory experiments on measurement instrumentation", *Instrumentation and Measurement, IEEE Transactions on*, vol. 55, no. 5, (2006), pp. 1757-1766.
- [9] V. Moisa, "Adaptive Learning Management System", *Journal of Mobile, Embedded and Distributed Systems*, vol. 5, no. 2, (2013), pp. 70-77.
- [10] E. A. Amrieh, T. Hamtini and I. Aljarah, "Preprocessing and analysing educational data set using X-API for improving student's performance", In *Applied Electrical Engineering and Computing Technologies (AEECT)*, 2015 IEEE Jordan Conference on. IEEE, (2015), pp. 1-5.
- [11] S. Putrevu, "Exploring the origins and information processing differences between men and women: Implications for advertisers", *Academy of marketing science review*, vol. 2001, no. 1, (2001).
- [12] S. S. Meit, N. J. Borges, B. A. Cubic and H. R. Seibel, "Personality differences in incoming male and female medical students", *Online Submission*.
- [13] F. G. Gómez, J. Guardiola, O. M. Rodríguez and M. A. M. Alonso, "Gender differences in e-learning satisfaction", *Computers & Education*, vol. 58, no. 1, (2012), pp. 283-290.
- [14] C. S. Ong, and J. Y. Lai, "Gender differences in perceptions and relationships among dominants of e-learning acceptance", *Computers in human behavior*, vol. 22, no. 5, (2006), pp. 816-829.
- [15] C. Romero, S. Ventura, P. G. Espejo and C. Hervás, "Data mining algorithms to classify students", in: *Educational Data Mining*, vol. 2008, (2008).
- [16] J. Ermisch and M. Francesconi, "Family matter: Impacts of family background on educational attainment", *Economical*, vol. 68, (2001), pp. 137-156.
- [17] A. Agus and Z. K. Makhbul, "An empirical study on academic achievement of business students in pursuing higher education: An emphasis on the influence of family backgrounds", In *International Conference on the Challenges of Learning and Teaching in a Brave New World: Issues and Opportunities in Borderless Education*. Hatyai Thailand, (2002).
- [18] S. Rothman, "School absence and student background factors: A multilevel analysis", *International Education Journal*, vol. 2, no. 1, (2001), pp. 59-68.
- [19] J. DeKalb, "Student truancy. (Report No. EDO-EA-99-1). Washington, DC: Office of Educational Research and Improvement", (ERIC Document Reproduction Service No. ED429334), (1999).
- [20] S. Gunuc and A. Kuzu, "Student engagement scale: development, reliability and validity", *Assessment & Evaluation in Higher Education*, vol. 40, no. 4, (2015), pp. 587-610.
- [21] G. D. Kuk, "Assessing what really matters to student learning", *Change*, vol. 33, no. 3, (2001), pp. 10-17.
- [22] I. Stovall, "Engagement and Online Learning. UIS Community of Practice for ELearning. <http://otel.uis.edu/copel/EngagementandOnlineLearning.ppt>, (2003).
- [23] C. Romero, J. R. Romero and S. Ventura, "A survey on pre-processing educational data", In *Educational Data Mining*. Springer International Publishing, (2014), pp. 29-64.

- [24] A. G. Karegowda¹, A. S. Manjunath² and M. A. Jayaram³, “Comparative study of attribute selection using gain ratio and correlation based feature selection”, *International Journal of Information Technology and Knowledge Management*, vol. 2, no. 2, **(2010)**, pp. 271-277.
- [25] R. Arora and S. Suman, “Comparative analysis of classification algorithms on different datasets using WEKA”, *International Journal of Computer Applications*, vol. 54, no. 13, **(2012)**, pp. 21-25.
- [26] D. M. Powers, “Evaluation: from precision, recall and F-measure to ROC”, *informedness, markedness and correlation*, **(2011)**.
- [27] T. Y. Chen, F. C. Kuo and R. Merkel, “On the statistical properties of the f-measure. In *Quality Software, 2004. QSIQ 2004*”, *Proceedings. Fourth International Conference on. IEEE*, **(2004)**, pp. 146-153.
- [28] M. M. Quadri, and N. V. Kalyankar, “Drop out feature of student data for academic performance using decision tree techniques”, *Global Journal of Computer Science and Technology*, vol. 10, no. 2, **(2010)**.
- [29] C. Romero, S. Ventura and E. García, “Data mining in course management systems: Moodle case study and tutorial”, *Computers & Education*, vol. 51, no. 1, **(2008)**, pp. 368-384.
- [30] M. F. Moller, “A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning”, *Neural Networks*, vol. 6, no. 4, (1993), pp. 525-533.
- [31] P. M. Arsad, N. Buniyamin and J. L. A. Manan, “A neural network students' performance prediction model (NNSPPM)”, In *Smart Instrumentation, Measurement and Applications (ICSIMA)*, 2013 *IEEE International Conference on. IEEE*, **(2013)**, pp. 1-5.
- [32] N. T. N. Hien and P. Haddawy, “A decision support system for evaluating international student applications”, In *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual. IEEE*, **(2007)**, pp. F2A-1.
- [33] Z. H. Zhou, “Ensemble methods: foundations and algorithms”, *CRC Press*, **(2012)**.

