IDEA SUBMISSION

Mining Educational Data to Predict Student's academic Performance using Ensemble Methods
Vellore institute of technology (Chennai)

Team Members –

Deepshikha Tiwari – 19MCA1013

Akanksha Agnihotri – 19MCA1040

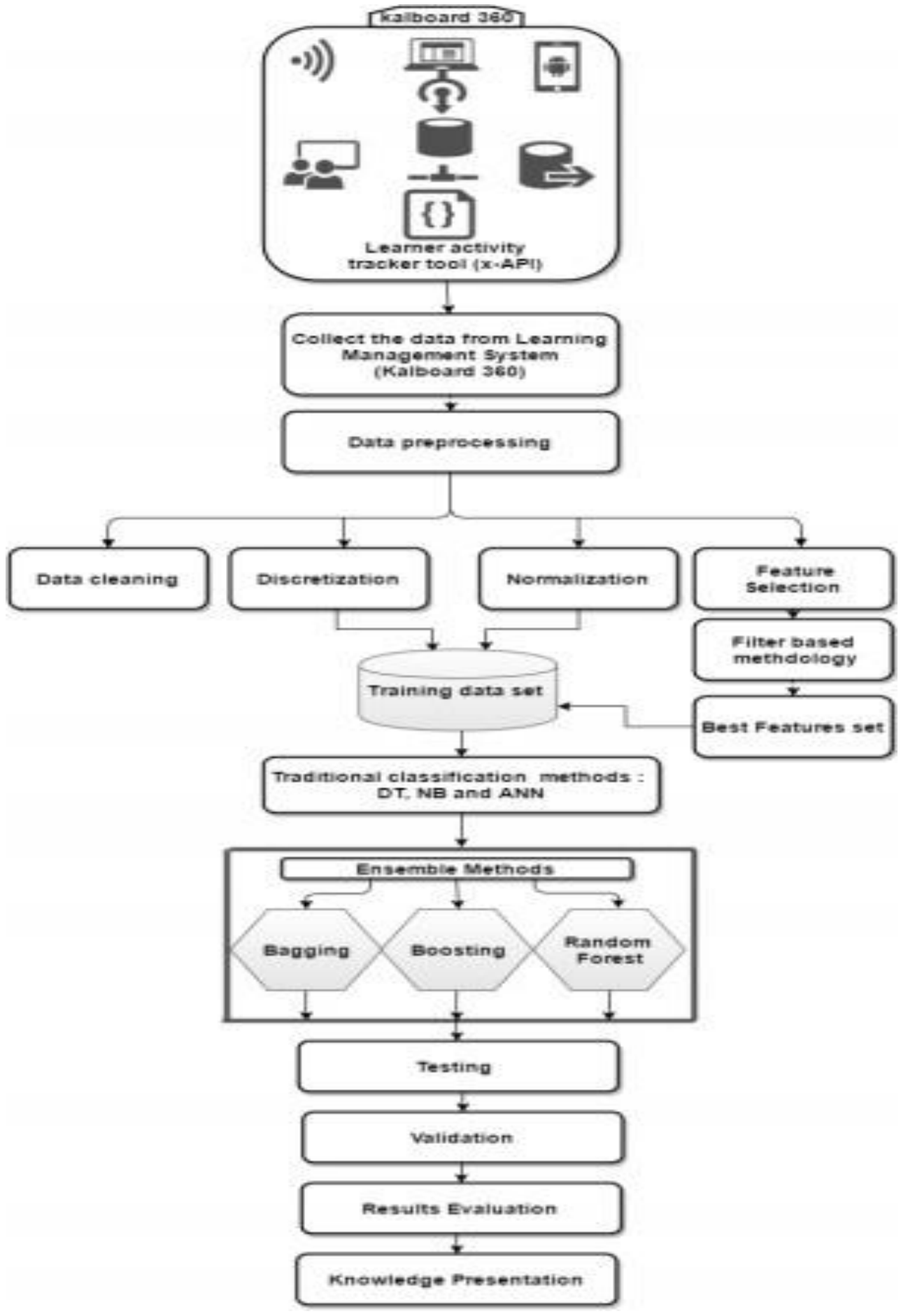
Abstract

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 types 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 the 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.

Introduction

This paper introduces a student's 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.This model used some data mining Online Version Only. 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 leaner 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), Decision Tree, and Naïve Bayes. 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).

Model used with justification


kalboard 360
Learner activity tracker tool (x-API)
Collect the data from Learning Management System (Kalboard 360)
Data preprocessing
Data cleaning
Discretization
Normalization
Feature Selection
Filter based methdology
Training data set
Best Features set
Traditional classification methods : DT, NB and ANN
Ensemble Methods
Bagging
Boosting
Random Forest
Testing
Validation
Results Evaluation
Knowledge Presentation

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 manages and simplifies online learning. The main purpose of the LMS is to manage learners, monitor student participation, keeping track of their progress across the system. 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).

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, Biological confirms that there are differences in the aptitudes of students that depend on gender. Meet in found that most of female students have a positive learning style in compare to male students. The authors in 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. For the family background feature, different studies have shown that there is a positive relationship between the parent's education and student's performance. This relation is particularly valid when the learner is being followed up by their mother. The authors in 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. Previous research 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.

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.

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.

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, and removes redundant and irrelevant data. 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.

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.

Result analysis

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 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. 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).

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 types 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 is 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 student's 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.

Implementation roles within team

Deepshikha Tiwari – Data collection and preprocessing, Feature analysis, Result analysis.

Akanksha Agnihotri – Data visualization, Data cleaning, Feature selection.