Model To Analyze And Predict The Academic Performance Of Students Using Data Mining Techniques

Athullya Bhuvanendran, Dr. Sminu Izudheen

1PG Scholar, Rajagiri School of Engineering and Technology
Email: athullya8@gmail.com
2Associate Professor, Rajagiri School of Engineering and Technology
Email: sminu_i@rajagiritech.ac.in

Abstract
One of the most discussed challenges in this dynamic era is to improve the quality of educational process so as to complement and enhance student’s performance and monitor their overall growth in halls of knowledge such as universities. Educational Data Mining simply refers to digging out the crucial information from the large educational database which is formed by the learning activities of students in institutions. This work presents the output of applying a mining approach to model the academic performance. A data mining model was developed to analyze the academic data, the model uses four classification techniques such as Naive Bayes, Multilayer Perceptron, ID3 and Random Tree in order to acquire a wide knowledge of the performance of students during the upcoming semesters. Historical academic records of students were used as training data to train the model. This proposed model helps to predict the student’s future learning outcomes using data sets of students till the current scenario to predict next semester performance and guide them for their improvement in academics. The accuracy of the work varies from 50% to 80%, with highest accuracy being attained by Random Forest followed by Multilayer Perceptron and with the least accuracy of 70% by ID3. The final remark obtained for the student could be used as a guideline for mentors to help the students improvise their future performance.

Keywords: Educational Data Mining, Classification, Multi-layer Perceptron, Random Forest, ID3, Naive Bayes.

1.Introduction
In this developing era, tremendous population hike leads to the formation of new Educational Institutions and settings. Educational institutions are becoming more competitive in the terms of quality education, infrastructural setups and numerous factors as the number of institutions are growing rapidly day by day. To be in a dynamic position, these institutions are focusing more on improving various aspects and one of the most inevitable factors among them is quality learning. For providing a dynamic work environment and quality proceedings, the institutions need to know about their core positives which are explicitly seen and negatives which are beneath the bars. To excel and to be competitive, the institutions should mark their own pillars of foundation and implement a technique to bring it out for progress. In past few years, Educational Data Mining has put on a global recognition within the research areas as it has been noticed as a vital need for the academic institutions to enhance their quality of education. Higher education institutions need to enhance their quality of education, as it is an alarming situation for them to extract a substantial amount of hidden knowledge from their repositories. The technique behind the extraction of the hidden knowledge from repositories of data is known as Knowledge Discovery Process[1]. It extracts the knowledge from available dataset and creates a knowledge base for the benefit of the educational institution.

Higher education of students has a proportional impact on the work profile provided to the industry and hence it unswervingly affects the overall economy of the country. In academic institutions the cumulative
quality of the education provided is being judged by the overall success rate of the students and till what extent the institution is capable of sustaining its students from moving out of the institution. A student's academic performance is a combination of various factors like social, personal, and psychological and various other environmental factors. Prediction of academic performance in any area whether school level or graduate level is always beneficial for the mentors as it will give the management a timely check on their growth graph and take essential steps to coach the students to improve his/her performance. Thus by doing this the overall quality of education could be enhanced and the institution along with the mentors can provide the students with best of the possible help, which in turn in a long run will help the economy build a strong foundation for its economic growth[2].

Various mining methods have been applied to predict and analyze the academic performance of the scholars based on the factors like socio-demographic conditions and previous academic records. Classification is one of the widely used mining techniques which classify an input data to a predefined class or category based on the model being trained earlier. It is a predictive method which classifies based on training set and uses the constructed pattern to classify a new data from testing set. Classification maps the data into predefined sets or groups of classes. It is often referred to as supervised learning as the final output classes are determined prior to the data examination. Patterns that are identified by data mining methods from learning centers data can be used to improvise the decision making process in terms of determining the students who all are at risk, identifying the student who are going to drop out, enhancing the students success rate and learning outcomes[3].

2. Organization and Summary

In the first section, some detailed information are provided to familiarize the concept of educational data mining including the different aspects in which it proves its significance. The second section gives a summary on the related works including different mining techniques used to analyze the academic performance of students. This section also focuses on various outcomes obtained by the works of various authors which has produced a successful result. The third section consists of the methodology of the proposed work. In this work four learning algorithms, naïve Bayes, ID3, Multilayer Perceptron, and Random forest were used to build classification models to predict the academic status. Based on the chosen classifier a model is built to predict the student’s range of percentage he might belong to by learning his previous data. The fourth section discusses the results obtained while implementing the model on the dataset of students. The fifth section concludes the overall findings with the remark of the work done. The final section mentions about the future scope of the work.

3. Recent Works

The application of the overall analytics in the educational context has increased in the last decade. Ferguson presents in [4] three reasons that why educational analytics has come into play: first, the data is increasing day by day and the overall data from various sources has increased remarkably from educational institutions, course or learning management systems. The second is the usage of e-learning: as it has a enormous strengths in its versatility of data collection, this also has possible disadvantages like it was difficult to know whether the students interest was consistent, the lack of motivation and the difficulties faced by the mentors to gain the mood of students and their interest towards the content. Finally, the political threats: economies are having a better knowledge of the importance of education in
their overall development and authorities have a keen interest in improvising it to offer better opportunities which will in turn enhance the academic results.

Brijesh Kumar Baradwaj and Saurabh Pal [5] made a model for classification of data for higher education. The work used decision trees for classification. The work extracted the knowledge which describes student’s performance in the end semester examination. It would provide a brief idea about the students who are in a verge of dropping out and for the mentors for helping the weak students at an early stage. Variables like Attendance of students during lectures, Class test marks, Seminar marks and individual and group Assignment marks were collected from the students historic database, so as to predict the end of the semester performance.

Ernesto Pathros[6] considered socio demographic and academic variables for the model implementation and classified the output data as low, medium and high. The dataset contained data of first semester students at a School of Engineering. Students who passed zero or up to two courses were classified as low, who passed three or four courses were classified as middle, and who passed all ve courses were classified as high. Naive Bayes classifier and the Rapidminer software were used for implementation and obtained accuracy of 60%. Regarding the methods of EDM, Baker proposed in[7] classification as follows: Prediction of data, Clustering of data into subgroups and Relationship mining for defining relationship between variables and further concluded that classification gave better prediction with substantial accuracy.

As the dropout rate of students reaching a higher rate day by day AnjanaPradeep[8] suggested a model using classification techniques to identify the weak students who are likely to perform poorly in their academics. WEKA was used to evaluate the variables for predicting academic failure. The data set is a combination of 67 attributes of 150 students of engineering graduate degree course of the academic year 2014-18 from Mahatma Gandhi University, Kerala, India. Classification techniques like decision tree and induction rules have been applied to the dataset. Several experiments were performed to get the highest accuracy for the model. In the first experiment, 8 classification algorithms were implemented with the all 67 attributes. In a second experiment, best attributes that is 13 were selected out of 67 attributes and that were used for the second experiment. In a third experiment, the model was implemented with refined dataset for precise accuracy.

In [9] R. Sumitha and Vinod Kumar gave a model which predicts the learning outcome of students using the historic dataset of previous students. Data set containing data of about 300 students were collected by three weeks. The dataset has been divided into training set with 250 data and test data with 50 records. Thereby after implementing the several classification algorithms, J48 algorithm gave the better prediction accuracy when compared to others.

Romero and Ventura suggested in [10] a whole new range of predictions and analysis of various possibilities of educational data. The work not only focuses about the academic records but also gave a wide range of areas for research like providing feedback system, behavior detection of students, academic performance prediction, grouping students according to their characteristics, scheduling and planning, development of a concept map and constructing the course content. However, Educational Data Mining being an application of data mining, its roles are the same i.e. Classification, Association Rules Analysis and Clustering. SuhemParack[11] used apriori algorithm to profile the students. Weka tool was used to mine the data. Apriori algorithm was used to get the best association rules which will in turn help in profiling the students data. Later k-means algorithm was also used to group the data categorically. The
students who were profiled were classified as good, satisfactory and poor by the means of clustering. The classification algorithm naive bayes and decision tree were observed as a better algorithm to classify the data. Decision tree uses a tree like structure to classify the data where each node is a decision maker while Naive Bayes provide a probabilistic relation between variables to classify the data set.

Pattaramon Vuttipittayamongkol[12] in his work focused on investigating predicting factors of student’s academic performance at a Thailand University. Author proposed three identified factors: activity hours of students, English scores obtained, and number of students admitted in a particular course. The influence of English scores was highly significant in predicting student’s performance. Not only the scores the proportional distribution of activities and scores could be used to attain a better quality of education and could be used as a guideline for admission process.

Camilo Lopez and Elizabeth presented the results of applying a data mining approach to know the fall in academic status of students at the University of Colombia, the model used two classification algorithm that is Naive Bayes and Decision tree and obtained a substantial accuracy when further data was added and Naive Bayes gave a better result when compared with other classifiers[13,14].

4. Proposed Work

In present day’s educational system, a student’s performance is analyzed by the internal assessment, co-curricular activities end semester examination results. The internal assessment is carried out by the teacher based upon student’s performance in educational activities such as subject internals, seminar, attendance, number of backlogs and lab work. The end semester examination is one that is scored by the student in semester examination. Each student has to score a bare minimum for the internal as well as the final end examination. In this work the previous semester data of students is trained to predict his upcoming semester performance by taking 5 major attributes internal marks, external marks, previous semester marks, backlog, attendance. The overall flow diagram of the proposed system is shown in figure1.

![High level data flow diagram](image)

**Figure 1:** High level data flow diagram

4.1 Data Gathering

Data may be gathered from many different and hetero-geneous data sources. This stage comprises of gathering all available information of students. Set of factors that affects a student’s performance is first identified and collected from various sources of data available. This is then
integrated into a single data set. The data set used in this work was obtained from Rajagiri School Of Engineering And Technology, Ernakulam (Kerala). Students data of Computer Science Department and Electronics and Communication department (session 2013 to 2017) were taken for the experiment. Initial size of the data was 110. Later on the data set was increased to 230. Most students were between the ages of 18 and 19, as this is the year when most of the students experience a new environment and infrastructure of study.

4.2 Data Segregation

In this step only those required attributes were selected which were useful for the mining process. A few already existing variables were selected. While some of the information for the variables were extracted from the database. To be determined variables like End Semester Percentage and the other selected attributes from the database are given in the Table I.

<table>
<thead>
<tr>
<th>Seq no</th>
<th>Name of Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Internal Marks</td>
</tr>
<tr>
<td>2</td>
<td>External Marks</td>
</tr>
<tr>
<td>3</td>
<td>Semester Percentage</td>
</tr>
<tr>
<td>4</td>
<td>Backlogs</td>
</tr>
<tr>
<td>5</td>
<td>Attendance</td>
</tr>
<tr>
<td>6</td>
<td>End Semester Percentage</td>
</tr>
</tbody>
</table>

TABLE 1: List Of Attributes

4.3 Data Pre Processing

Pre-processing of data is considered as a very important task in this work as we need quality and reliability of available information which directly affects the results attained. Before applying the data mining algorithms it is essential to carry out some pre-processing tasks like cleaning of data, integration and transformation. Pre-processing task includes finding the incorrect or missing data. Erroneous data or ambiguous data may be corrected or removed, whereas missing data must be supplied. Pre-processing also includes removal of noise or outliers and segregating only necessary information and discarding irrelevant data. Data Transformation is the process where conversion of the data into a standard format for further implementation is done. Some specific pre-processing tasks were applied to the data set so that data mining algorithms can be applied correctly. Firstly, students without 100% complete information were removed and then all the available data were integrated into a single dataset. Attributes that does not affect the classification results are discarded. For a compact view of data, continuous values were transformed into discrete variables. For example, the numerical values of marks or scores were converted into categorical values as per in table II,III and IV.

4.4 Classification

This section describes the experiments and data mining techniques used for obtaining the prediction models of student’s academic status at the end of the semester. Several experiments
were performed in order to try to obtain the highest classification accuracy. In the experiment, four classification algorithms using all available information were executed.

4.4.1 ID3 Classifier

It builds the decision trees from training data using the concept of information entropy. The training data is a set of various samples which are already classified. At each node of the tree formation, ID3 chooses the best attribute which splits the dataset effectively. The splitting criterion is the information gain that is the difference in entropy calculated. The attribute which scored the highest information gain value is chosen as the root node. The ID3 algorithm is again applied on the smaller subsets obtained after first iteration to classify each element in the dataset.

4.4.2 Naïve Bayes Classifier

It is based on conditional probabilities. It uses Bayes Theorem, an equation which calculates a probability by taking the count of the frequency of values and the combination of those values in the training[15]. It computes the probability of the event to happen given the probability of another event has already occurred. If B represents the dependent event and A represents the event that has already occurred, the theorem can be stated as follows.

Bayes Theorem: Prob(B given A) = Prob(A and B)/Prob(A)

4.4.3 Multi-Layer Perceptron Classifier

It is a feed forward artificial neural network model that maps sets of input data onto a set of outputs. An MLP comprises of multiple layers of input nodes in a directed graph, with each layer connected to the next node. All the nodes except the input nodes the activation function is there. MLP is a supervised learning technique which uses back propagation for training the entire network. MLP is a modification of the standard linear perceptron and can differentiate data that are not separable linearly[16].

4.4.4 Random Forest Classifier

They are a learning method for classification of data, regression of data and for other tasks, that operates by constructing a collection of decision trees at the time of training the dataset and outputting the class which is the mode of classes in classification and mean of the classes in regression[17]. Random Forests are a combination of tree predictors where each tree depends on randomly generated values of vector which are sampled independently with the same or equal distribution for all the trees in the forest. The basic principle is that a group of weak learners can come together to form a strong learner.

The above classifiers were used by the model to train the dataset of students for classifying the test dataset. Initially dataset of 110 students were taken later on additional data of 120 were added for the experiment purpose. Evaluation was carried out with the best five attributes. The 'End Semester Percentage' is the output attribute which was calculated and
was classified as A,B,C,D,E classes where A being the highest value class and E being the least.

<table>
<thead>
<tr>
<th>MARKS OBTAINED</th>
<th>CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>80 Above</td>
<td>A</td>
</tr>
<tr>
<td>71-80%</td>
<td>B</td>
</tr>
<tr>
<td>61-70%</td>
<td>C</td>
</tr>
<tr>
<td>51-60%</td>
<td>D</td>
</tr>
<tr>
<td>Below 50</td>
<td>E</td>
</tr>
</tbody>
</table>

**TABLE II: DATA TRANSFORMATION FOR MARKS OBTAINED**

<table>
<thead>
<tr>
<th>NUMBER OF BACKLOGS</th>
<th>CATEGORY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than equal to 2</td>
<td>Good</td>
</tr>
<tr>
<td>3 to 4</td>
<td>Average</td>
</tr>
<tr>
<td>Above 4</td>
<td>Poor</td>
</tr>
</tbody>
</table>

**TABLE III: DATA TRANSFORMATION FOR NUMBER OF BACKLOGS**

<table>
<thead>
<tr>
<th>Attendance Percentage</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>95 And Above</td>
<td>Excellent</td>
</tr>
<tr>
<td>85-94%</td>
<td>Good</td>
</tr>
<tr>
<td>75-84%</td>
<td>Average</td>
</tr>
<tr>
<td>Below 75%</td>
<td>Poor</td>
</tr>
</tbody>
</table>

**TABLE IV: DATA TRANSFORMATION FOR ATTENDANCE**

4.5 Analysis

In this section the comparative analysis of the four classifiers are done. The data set is divided into 5 subsets with 90% training data and 10% test set,70% training and 30% test set,50% training and 50% test set,30%training and 70% test set,10% training data and 90% test set. This is done to measure the accuracy of the model with varying training set to get a clear idea about the precision of the classification model.

5. Result and Conclusion

From the above discussions, it is clear that the model constructed using various algorithms can be used to predict the results of students once their details are known. If the information like the scores in
various subjects, number of backlogs, attendance rate etc. are known, the students result can be easily predicted using the constructed model using Naive Bayes, MLP, Random forest, ID3. The accuracy of the classifier is measured in terms of correctly classifying the unknown variable. It is based on true positives and false negatives. True positive is when the classifier classified it correctly and false negative is when it classified it incorrectly to any other class.

Accuracy of the classifier = true positives/(true positives + false negatives)

MLP and Random forest algorithm showed maximum accuracy upto 80%. A case study was done with the data of 110 students out of which 100% students have opted for Computer Science and Engineering (CSE). Later for better computation, data set was increased with 120 students data from Electronics and Communication department. Only the best 5 attributes were collected and used for the experiment. The experiment was done using WEKA with a complete dataset of 230 students data. The accuracy of the classifiers is shown in Figure 2 with accuracy on y axis and training dataset on x axis. The training dataset description is given in table V.

<table>
<thead>
<tr>
<th>X axis</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10% training data</td>
</tr>
<tr>
<td>1</td>
<td>30% training data</td>
</tr>
<tr>
<td>2</td>
<td>50% training data</td>
</tr>
<tr>
<td>3</td>
<td>70% training data</td>
</tr>
<tr>
<td>4</td>
<td>90% training data</td>
</tr>
</tbody>
</table>

TABLE V: DATASET SIZE DESCRIPTION

In this work, classification techniques were considered on the dataset of students, to predict and analyze their performance. The results of the data mining algorithms for the classification based on the attributes selected reveals that the prediction rates are not uniform among the algorithms, it varies according to the variation in training dataset. The range of prediction varies from (50-80%) across the evaluation phase. By comparative analysis of classification algorithms (such as Naive Bayes, MLP, random forest, ID3) using WEKA it is proven that the attributes chosen from the original dataset have high influence using Random Forest and MLP with an accuracy of 50% minimum when training dataset is very low, ranging up to approximately 80% maximum when the training data set is increased. Among all data mining classifiers Naive Bayes gave a consistent accuracy of approximately 60% throughout. While MLP and Random forest gave its best accuracy of about 80%. ID3 gave the least accuracy that is 70% when the training dataset was maximum. The overall conclusion for the work is as the data in the training dataset is increased the accuracy went high. As there is a limitation with academic records of students the accuracy could reach upto 80% only, while if there were huge training datasets of students of a particular academic period with large number of records the classifiers could outperform the prediction considerably. The comparison graph is shown in figure 2.
6. Conclusion

The academic success of the students of any institution has become an alarming concern for the management. An early detection of students with risk would help the management to uplift them from downfall at the initial stage itself. Employing data mining techniques like classification, clustering could be used to develop a support system to help authorities identify the weak students and take timely measures to curb the extremes. After the detailed study about the classification algorithms applied in this work Random Forest classifier and Multilayer Perceptron outperformed other classifiers that is ID3 and Naive Bayes. As the training data increased the overall prediction rate also increased for the classifiers. Any model with such specification could be useful in the educational system like Universities and Colleges for maintaining the overall quality of the education. By doing this we can know the academic status of the students in advance and can concentrate on students to improve their academic results and overall placements rates. Thereby improve their standards and reputations. As a byproduct the quality of education can be pushed high to maintain a standard benchmark for other competitors.

7. Future Scope

Data from the admissions process are merged with the academic information that is collected from each academic period of a student, however, the reasons of low academic credential occur on a daily basis and waiting until the next academic session ends, could be crucial step. This forces us to think the new and possibly non-traditional ways, for collecting information close to real time are needed. This work focused on the loss of academic status due to low performance, however, the academic performance can also be studied at a whole different level, perhaps at the course level. The classification model could also include the non-academic loss of student status, or a new model could be built to react to this specific situation. Further enhancements and various alternatives could be used to consider a larger number of parameters which will guide us to predict more in-depth information about the student under consideration. Possible implementation of further prediction could be used to recommend a student on the choice of his/her major subject based on factors like interest in a particular field, marks obtained, and personal habits. On the other side we could also predict potentially violent behavior among students. Instead of taking into account of only the academic credentials we could also take other socio-demographic parameters like his/her personal details, parents educational qualification, living habits, his/her past high school information about academics, hobbies etc. The academic prediction not only means to find out his/her progress but also to get a particular progress graph about the students as well as to finding out the overall flaw of the students. The work could be extended in such a way where the co-curricular activity of the student could also be mined to find out the best we have in our institution. This is to be done not only for the benefit of the institution but also for enhancing the personal skills of the students. Once it has been done, the clustering method can be used to group the students into various clusters according to their academic and co-curricular activities. After the cluster formation is done the Educational Institutions can use all the determined results for further processing and for mentoring the students accordingly to enhance their not only academic skill but also inter personal skills.

References


[16] https://en.wikipedia.org/wiki/Multilayer_perceptron

[17] https://docs.opencv.org/2.4/modules/ml/doc/random_trees.html

**Authors Biography**

**AthullyaBhuvanendran** received the Bachelors of Engineering Degree in Computer Science from Rashtrasant Tukadoji Maharaj Nagpur University. She is currently a student at Rajagiri School of Engineering and Technology pursuing her Masters of Technology in Computer Science and Information Systems affiliated to A P J Abdul Kalam Kerala Technological University. Her research interests include applying data mining in educational data and data analytics.

**Dr. SminulZudheen** received her PhD from CUSAT, Kerala. She is currently an associate professor at Rajagiri School of Engineering and Technology in Computer Science and Engineering Department, affiliated to A P J Abdul Kalam Kerala Technological University. She is a member at Indian Society For Technical Education. Her research interests include Bioinformatics, Data Mining, Automata Theory.