# STUDENT PERFORMANCE PREDICTION USING DIFFERENT CLASSIFICATION ALGORITHMS

Kolluru Venkata Nagendra[1], K.Sreenivas[2], P.Radhika[3]
[1,2,3]Assistant Professor, Department of CSE, Geethanjali Institute of Science & Technology,
Nellore, Andhra Pradesh, India

**Abstract**

**In this paper final year B.Tech students results are analyzed using Data Mining Techniques. These results are taken from a collage in JNTUA region in Andhra Pradesh in INDIA. The primary goal of this research is to predict the student performance in the last semester external exams. Support Vector Machines, Decision Tree and Gradient Boosting algorithms are used for classifying the performance of students and accuracy of the three algorithms. The result of this study reveals that overall accuracy of the tested classifiers is above 75%. In addition classification accuracy for the different classes reveals that the predictions are not good for distinction and fairly good for the first class. The Gradient Boosting algorithm produces highest classification accuracy for the Distinction.**

**Depends upon the attributes the prediction rate changes. The influence of selected attributes may effect on classification process.**

**Keywords: Support Vector Machines, Decision Tree, Boosting, Classifiers, Comparative Analysis, Predicting Student Performance.**

## I.INTRODUCTION

Now a day's Educational data mining is an increasing research area in Data Mining. The challenging issue in the real world is the prediction of the student performance. One of the primary requirements in this process is that high quality and relevant data has to be provided to the educational leaders at the right time. Traditionally educational institutions are collecting large volumes of data related to students, faculty members, the organization and management of the educational process, and other managerial issues. However, the extent to which the available and collected data is being used is not so significant. In general, the data is used for producing simple queries and traditional reports that are not highly significant in contributing to the decisions making process in the institutions. Moreover, the volume and complexity of the data is often very huge that it becomes difficult to the management of the educational institutions to handle the data and hence remains unused. The potentiality of the available volume of data can be exploited only if it transformed into useful information and in turn is used to generate knowledge to support decision making.

Data mining is the process of discovering meaningful patterns in large quantities of data. Considering the potential application of data mining in educational sector, Educational Data Mining (EDM) was started as a new stream in the data mining research field2. EDM concerns with new methods and techniques by inquiring into eccentric type of data from educational settings to understand students learning ability. The aim of classification is to predict the future output based on the available data. Hence, educational institute is looking to predict the future output of their enrolled students based on their available previous and current students' data, which make classification one of the techniques better suited for educational analysis. Most of the previous studies focus on the use of classification for predictions based on enrollment data, Performance of students in certain course, grade inflation, anticipated percentage of failing students, and assist in grading system. Up to our knowledge, there are no studies that use classification to predict a student final outcome based on his/her grades in a program study plan. Analyzing all the courses that are required in the study plan will identify the list of courses that have a huge impact on final results.

## II. CLASSIFICATION ALGORITHMS

Decision trees, neural networks, k-nearest Neighbor, Naive Bayes and support vector machines are used in Educational Data mining. Using these methods many kind of knowledge can be discovered such as association rules, classification, clustering, and pruning the data. Some of the Classification algorithms mentioned here for the proposed work have provided a better understand in educational resources.

### 2.1. Decision Tree Classifier

Decision tree classifiers are one of the popular and powerful tools for classification. Generally, decision tree classifiers have a tree-like structure which starts from root attributes, and ends with leaf nodes. It also has several branches consisting of different attributes, the leaf node on each branch representing a class or a kind of class distribution. Decision tree algorithms describe the relationship among attributes, and the relative importance of attributes. The advantages of decision trees are that they represent rules which could easily be understood and interpreted by users, do not require complex data preparation, and perform well for numerical and categorical variables.

### 2.2. Support Vector Machines

Support Vector Machines (SVM) with linear or nonlinear kernels have become one of the most promising learning algorithms for classification as well as for regression which are two fundamental tasks in data mining via the use of kernel mapping, Variants of SVMs have successfully incorporated effective and flexible nonlinear models Kernel-based techniques (like support vector machines, kernel principal component analysis, Bayes point machines, and Gaussian processes) represent a major development in machine learning algorithms. SVM (support vector machines)is a group of supervised learning techniques or methods, which is used to do for classification or regression. SVM (support vector machines) represents an extension to nonlinear models of the generalized portrait algorithm. The basic idea of SVM (support Vector Machines) is to map the original data X into a feature space F with high dimensionality through a non-linear mapping function and construct an optimal hyper-plane in new space. SVM can be applied to both classification and regression. In the case of classification, an optimal hyper-plane is found that separates the data into two classes. Whereas in the case of regression a hyper-plane is to be constructed or developed that lies close or near to as many points as possible.

### 2.3. Boosting

The concept of boosting applies to the area of predictive data mining, to generate multiple models or classifiers, and to derive weights to combine the predictions from those models into a single prediction or predicted classification. A simple algorithm for boosting works like this: Start by applying some method to the learning data, where each observation is assigned an equal weight. Compute the predicted classifications, and apply weights to the observations in the learning sample that are inversely proportional to the accuracy of the classification. In other words, assign greater weight to those observations that were difficult to classify, and lower weights to those that were easy to classify. In the context of C&RT for example, different misclassification costs (for the different classes) can be applied, inversely proportional to the accuracy of prediction in each class. Then apply the classifier again to the weighted data, and continue with the next iteration.

Boosting will generate a sequence of classifiers, where each consecutive classifier in the sequence is an "expert" in classifying observations that were not well classified by those preceding it. During deployment (for prediction or classification of new cases), the predictions from the different classifiers can then be combined to derive a single best prediction or classification.

## III.DATASET FOR PROPOSED WORK

A student’s dataset was created based on the characteristics of the students along with their performance in the class and university examinations. The dataset was used to evaluate the performance of various classification algorithms in predicting the performance of the students in the final exams. The data mining classification algorithms that are compared in the study includes Decision Tree algorithm, Support Vector Machine and Boosting.

The dataset used in the study consists of primary data generated from the student's admission data available with the college database. In addition, certain aspects of the dataset are collected by administering a structured questionnaire to the concerned students. The target variable or the output variable is Student Final Semester Marks which is usually available in the numeric form in terms of percentage. Hence categorical target variable was constructed based on the original numeric parameter (percentage score). The target variable has four distinct values as Distinction (Score is greater than 75%), First Class (Score lies between 60 to 74%), Second Class (Score lies between 50 and 59%), Third Class (Score lies between 35 and 49%), Fail (Score less than 35%).

The attributes referring to the students' schooling characteristics include Students Grade in High School and Students Grade in Senior Secondary School. The attributes describing other college features include thebranch of study of the students, place of stay, previous semester mark, class test performance, seminar performance, assignment, general proficiency, class attendance and performance in the laboratory work. The study is limited to student data for different branches of a college in JNTUA region.The detailed description of the dataset are provided in Table 1. The domain values for some of the variables were defined for the present investigation as follows:

- GENDER: Gender of the students. It is split into two classes values: Male and Female
- BRANCH: Students branch obtained. Branch is split into five Classes: CSE, ECE, EEE, CE and ME.
- ST_CAT: Students category obtained. Here Category is split into six classes: BC-Backward class, OBC- other backward class, OC-Open category, SC- Schedule Castes, ST-Schedule Tribal's.
- SSC: Students Grade in High School.(10th Class). Here grade is divided into Seven class values: O=95-100%, A=80%-89%,B=70%-79,C= 60%-69%,D=50%-59%,E=35%-49%,FAIL <35%.
- DS-Day scholar
- HS: Hostler -Students stay in hostel.
- LSM – Last Semester marks of Students obtained in different branches.
- Mid Term Exams: In each semester two internal tests are conducted and average of two tests are used. CTG is split into three classes: Poor - < 40%, Average - >40% and <60%, Good - >60%.
- AE: Assignment Exams
- PE: Present Attendance
- EE: External Exams

**Table 1**: **Description of the attributes used for Classification**

| **Variables** | **Description** | **Possible Values** |
|---|---|---|
| Gender | Students Sex | {Male, Female} |
| Branch | Students Branch | {CSE,ECE,EEE,CE,ME} |
| St_Cat | Students category | BC, OBC, OC, SC &ST |
| SSC | Students grade in High School | {O – 90% -100%,<br>A – 80% - 89%,<br>B – 70% - 79%,<br>C – 60% - 69%,<br>D – 50% - 59%,<br>E – 35% - 49%,<br>FAIL - <35%} |
| Inter | Intermediate % | {O – 90% -100%,<br>A – 80% - 89%,<br>B – 70% - 79%,<br>C – 60% - 69%,<br>D – 50% - 59%, |

| | | |
|---|---|---|
| | | E – 35% - 49%,<br>FAIL - <35% } |
| DS | Day scholar- Living Location of Student | {Village, Taluk, Rural, Town, District} |
| HS | Hostler- Student stay in hostel or not | {Yes, No} |
| LSM | Last Semester Mark | {First > 60%<br>Second >45 &<60%<br>Third >36 &<45%<br>Fail < 36%} |
| Mid Term | Internal Exams Grade | {Poor, Average, Good} |
| SEM_P | Seminar Performance | {Poor , Average, Good} |
| AE | Assignment Exam | {Yes, No} |
| PATT | Present Attendance | {Poor , Average, Good} |
| EE | ExternalExams | Distinction >75%<br>First Class >60 and <75<br>Second Class >50 and <59,<br>Third Class >35 and <49<br>Fail <35 |

## IV.RESULTS ANALYSIS

The main objective of the study is to explore if it is possible to predict the performance of the student (output) based on the various explanatory (input) variables which are retained in the model. The classification model was built using several different algorithms and each of them using different classification techniques. R-Programming is used to find the results.

### 4.1 Decision Tree Classifier Results

In the present study, Decision Tree classification algorithm was implemented on the data and the results of the classification is analyzed. It is observed that the results reveal that the True Positive Rate is high for three of the classes – Third (100 %), First (84-98 %). The TP rate is low for the class - Distinction (50 %), while it is very low for the class– Second (40-66 %), Fail (11-16 %). The Precision is high for the First class (67-76 %), Second class (72-85 %), medium for the Distinction (54 %) and low for the class Fail (29-33 %) classes.

### 4.2 SVM Results

The present study implements Support Vector Machines on the dataset and the results are correctly classified. It reveals that the True Positive Rate is high for most of the classes: First, second and Third. TP rate is very low for the class Fail (16.7 %). The precision is also high for the classes - First, Second and Third. It can be verified that SVM correctly classifies about 73.38 % for the 10-fold cross-validation testing and 74.23 % for the percentage split testing. The results shows that the True Positive Rate is high for the classes - First and Third. TP rate is very low for the class Distinction (11.1 %). The precision is also high for the classes - Third.

### 4.3 Boosting Results

In R programming, Boosting is applied and it correctly classifies about 68.32 % for the 10-fold cross-validation testing and 62.92% for the percentage split testing. The results show that the True Positive Rate is high for the Third and First class (73-87%). TP rate is very low for the classes Distinction and Fail. The precision is found to be high for the classes –First and Third and very low for the classes – Distinction. TP rate is Zero for the class Fail.

## V.PERFORMANCE COMPARISON BETWEEN THE APPLIED CLASSIFIERS

The results for the performance of the selected classification algorithms (TP rate, percentage split test option) are summarized and presented. The results of the classification reveals that the Boosting classifiers performs very well in comparison with other classifiers Decision Tree, Support Vector Machines. The overall accuracy of all the tested classifiers is well above 60%. Decision Tree and Support Vector Machines registered accuracy higher than 69%. But the best classifier for this data set is Boosting; the accuracy is very high up to 75%. In addition, further detailed analysis of the classification accuracy for the different classes reveals that the predictions are worst for the distinction class and fairly good for the other classifiers. The classification accuracy is very good for first class. The Boosting produces highest classification accuracy for the Distinction.

## VI. CONCLUSION

The accuracy of Data Mining algorithms depends on the attributes of the student data. The student performance prediction may vary from 65 to 75%. The classifiers prediction is different on depending up on different classes and also branches. The significant influence of classification on data attributes is first and second. The performance of student external exams is also depends on living status of the student. If the student lives in hostel the results are high otherwise the results are lie between second and third class. The Boosting algorithm gives more accurate results on student performance compared to other classifiers Decision Tree, Support Vector Machines.

## VII. REFERENCES

[1]. C.Anuradha and T Velmurugan, A Comparative Analysis on the evalution of classification algorithms in the prediction of student performance, IJSt,Vol8(15),july-2015.

[2]. Shanmuga PK. Improving the student’s performance using Educational data mining. International Journal of Advanced Networking and Application. 2013; 4(4):1680–5.

[3]. DorinaKabakchieva. Predicting Student Performance by using Data mining Methods for Classification. Bulgarian Academy of Science, Cybernetics and Information Technologies. 2013; 13(1):61–72.

[4]. Longbing C. Data mining and multi-agent integration. Springer Science & Business Media; 2009.

[5]. Tan, Pang-Ning, Steinbach M, Kumar V. Introduction to data mining. Boston: Pearson Addison Wesley; 2006; 1.

[6]. Al-Barrak MA, Al-Razgan M. Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology. 2016 Jul; 6(7):528–33.

[7].Han J, Kamber M. Data Mining: Concepts and Techniques, New Delhi: Morgan Kaufmann Publishers; 2006. ISBN: 978-81-312-0535-8.

[8]. Baradway BK, Pal S. Mining Educational Data to Analyze Students Performance. Int Journal of Advances in Computer Science and Applications. 2011; 2(6):63–9.

[9].Galit S, Patel NR, Bruce PC. Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner. John Wiley & Sons; 2007.

[10].Romero C, Ventura S. Educational data mining: A survey from 1995 to 2005. Expert systems with applications. 2007; 33(1)135–46.

[11].Baker RSJD. Data mining for education. International encyclopedia of education. 2010; 7:112–8.

[12]. Pal AK, Pal S. Analysis and Mining of Educational Data for Predicting the performance of Students. International Journal of Electronics Communication and Computer Engineering. 2013; . 4(5):1560–5.

[13]. Rathee A, Mathur RP. Survey on Decision Tree Classification algorithm for the Evaluation of Student Performance. International Journal of computers & Technology. 2013; 4(2):244–7.

[14]. Aher SB, Lobo LMRJ. Data mining in educational system using Weka. IJCA Proceedings on International Conference on Emerging Technology Trends (ICETT). 2011; 3:20–5.

[15]. Ajith P,, Tejaswi B, Sai MSS. Rule Mining Framework for Students Performance Evaluation. International Journal of Soft Computing and Engineering. 2013; 2(6):201–6.

[16]. Ogunde AO, Ajibade DA. A Data Mining System for Predicting University Students’ Graduation Grades Using ID3 Decision Tree Algorithm. Journal of Computer Science and Information Technology. 2014; 2(1):21–46.

[17]. Trivedi A. Evaluation of Student Classification Based On Decision Tree. Int Journal of Advanced Research in Computer Science and Software Engineering. 2014 Feb; 4(2):111–2.

[18]. Agrawal BD, GuravBharti B. Review on Data Mining Techniques used For Educational System. Int Journal of Emerging Technology and Advanced Engineering. 2014 Nov; 4(11):325–9.

[19]. Suman, Pooja Mittal P. A Comparative Study on Role of Data Mining Techniques in Education: A Review. International Journal of Emerging Trends & Technology in Computer Science. 2014 Jun; 3(3):65–9.

[20]. Dinesh KA, Radhika V, A Survey on Predicting Student Performance. Int Journal of Computer Science and Information Technologies. 2014; 5(5):6147–9.

[21]. Chirumamilla V, BhagyaSruthi T, Velpula S, Sunkara I. A Novel approach to predict Student Placement Chance with Decision Tree Induction. International Journal of Systems and Technologies. 2014; 7(1):78–88.

[22].Ross QJ. C4. 5: programs for machine learning. Elsevier; 2014.

[23]. Stuart R, Norvig P. Artificial Intelligence: A Modern Approach. EUA: Prentice Hall; 2003.

[24].Altman NS. An introduction to kernel and nearest neighbor non-parametric regression. The American Statistician. 1992; 46(3):175–85.

[25]. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann; 2005.

[26]. Cohen WW. Fast effective rule induction. Proceedings of the twelfth international conference on machine learning; 1995. p. 115–23.

[27]. Masethe MA, Masetha HD. Prediction of work integrated Learning placement using Data mining Algorithms. Proceedings of the World Congress on Engineering and Computer Science. 2014 Oct; 1:22–4.

[28]. Sundar PVP. A Comparative Study For Predicting Students Academic Performance using Bayesian Network Classifiers. IOSR Journal of Engineering. 2013 Feb; 3(2):37–42.

[29]. Kulkarni P, Ade R. Prediction of Students Performance based on Incremental Learning. International Journal of Computer Applications. 2014 Aug; 99(14):10–6.