STUDENT PERFORMANCE PREDICTION USING DATA MINING TECHNIQUES

Durgesh Ugale[1], Jeet Pawar[2], Sachin Yadav[3], Dr. Chandrashekhar Raut[4]

[1]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Airoli, Maharashtra, India*
[2]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Airoli, Maharashtra, India*
[3]*Student, Dept. of Computer Engineering, Datta Meghe College of Engineering, Airoli, Maharashtra, India*
[4]*Ph.D Professor, Dept. of Computer Engineering, Datta Meghe College of Engineering, Airoli, Maharashtra, India*

---***---

Abstract - *The success of an academic institution can be measured in terms of quality of education provides to its students. In the education system, highest level of quality is achieved by exploring the data relating to redirection about students performance. These days the lack of existing system to analyse and judge the students performance and progress isn"t being addressed. There are 2 reasons why this is often happening. First, the present system is not accurate to predict students" performance. Second, because of shortage of consideration of some vital factor those are affecting students" performance. Predicting students" performance is more challenging task as a result of large amount of information in academic database. This proposed system can help to predict students" performance more accurately. For these suitable data mining approach will be applied. In this approach, preprocessing step will be applied to raw dataset so that the mining algorithm will be applied properly. The prediction about students" performance can help him/her to enhance the performance.*

Key Words: Education, student, performance, data mining, pre-processing, database, prediction

1. INTRODUCTION

Improving student's academic performance is not an easy task for the academic community of higher learning. The academic performance of engineering and science students during their first year at university is a turning point in their educational path and usually encroaches on their General Point Average (GPA) in a decisive manner. The students evaluation factors like class quizzes mid and final exam assignment lab -work are studied. It is recommended that all these correlated information should be conveyed to the class teacher before the conduction of final exam. This study will help the teachers to reduce the drop out ratio to a significant level and improve the performance of students. In this paper, we present a hybrid procedure based on Decision Tree of Data mining method and Data Clustering that enables academicians to predict student"s GPA (SGPA, CGPA) and based on that instructor can take necessary step to improve student academic performance.

Graded Point Average (gpa) is a commonly used indicator of academic performance. Many universities set a minimum gpa that should be maintained. Therefore, gpa still remains the most common factor used by the academic planners to evaluate progression in an academic environment. Many factors could act as barriers to student attaining and maintaining a high gpa that reflects their overall academic performance, during their tenure in university. These factors could be targeted by the faculty members in developing strategies to improve student learning and improve their academic performance by way of monitoring the progression of their performance. With the help of clustering algorithm and decision tree of data mining technique it is possible to discover the key characteristics for future prediction. Data clustering is a process of extracting previously unknown, valid, positional useful and hidden patterns from large data sets.

The amount of data stored in educational databases is increasing rapidly. Clustering technique is most widely used technique for future prediction. The main goal of clustering is to partition students into homogeneous groups according to their characteristics and abilities. These applications can help both instructor and student to enhance the education quality. This study makes use of cluster analysis to segment students into groups according to their characteristics. Decision tree analysis is a popular data mining technique that can be used to explain different variables like attendance ratio and grade ratio. Clustering is one of the basic techniques often used in analysing data sets. This study makes use of cluster analysis to segment students in to groups according to their characteristics and use decision tree for making meaningful decision for the student's.

2. PREVIOUS WORK

Data mining (sometimes known as knowledge or information discovery) is the method of analysing information from totally different views and summarizing it into useful information. Information that may be used to increase revenue, cuts costs, or both data mining software system is one of the varieties of analytical tools for analysing information. It permits users to analyse the information identity. Technically, data mining/data processing is the process of finding correlations or patterns among dozens of fields in massive relational databases. Following are the survey papers being studied:

- Paris et. al.(1), compared data mining methods accuracy to classifying students in order to predicting category grade of a student. These predictions are more helpful for identifying the weak students and helping administration to take remedial measures at initial stages to produce excellent graduates which will graduate at least with the upper second category [1]

• Rathee and Mathur applied ID3, C4.5 and CART decision tree algorithm on educational information for predicting a student performance in the examination. All the algorithms are applied on the internal assessment information of student to predict their academic performance in the final examination. The efficiency of various decision tree algorithms will be analysed based on their accuracy and time taken to derive the tree. The prediction obtained from the system has helped the class teacher to identify the weak students and improve their performance. C4.5 is the best algorithm among all the three because it provides higher accuracy and efficiency than the other algorithms [3].

• Kortemeyer and Punch applied data mining classifiers as a means of comparing and analyzing students' use and performance who have taken a technical course via the web. The results show that combining multiple classifiers leads to a significant accuracy improvement in a given data set. Prediction performance of combining classifiers is often better than a single classifier because the decision is relying on the combined output of several models[3]

3. STEPS OF DATA MINING

Data mining is the method of discovering numerous models, derived values and summaries from a given collection of information. It's necessary that the problem of discovering or estimating dependencies from information or discovering new information is simply one part of the overall experimental procedure utilized by engineers, scientists and others who apply standard steps to draw conclusions from information. The overall method of finding and decoding patterns and models from information involves the recurrent application of the subsequent steps [6]:

1. Understand the application domain, the relevant previous knowledge and the goals of the end-user (formulate the hypothesis).
2. **Data Collection**: Determining how to find and extract the right data for modeling. First, we need to identify the different data sources are available. Data may be scattered in different spreadsheets, files, and hard-copy (paper) lists
3. **Data integration**: Integration of multiple data cubes, databases or files. A big part of the integration activity is to build a data map, which expresses how each data element in each data set must be prepared to express it in a common format and record structure.
4. **Data selection**: First of all the data are collected and integrated from all the various sources, and we select only the data which useful for data mining. Only relevant information is selected.
5. Pre-processing: The Major Tasks in Data Preprocessing are: Cleaning, Transformation and Reduction.
   - **Data cleaning**: Additionally known as data cleansing. It deals with errors detection and removing from information so as to improve the quality of information. Information cleaning sometimes includes fill in missing values and identify or remove outliers.
   - **Data Transformation:** Data transformation operations are additional procedures of data pre-processing that would contribute toward the success of the mining process and improve data-mining results. Some of Data transformation techniques are Normalization, Differences and ratios and Smoothing.
   - **Data Reduction**: For large datasets there‟s an increased probability that an intermediate, data reduction step should be performed before applying data mining techniques. While massive datasets have potential for higher mining results, there‟s no guarantee that they‟ll produce better knowledge than small datasets. Data Reduction obtains a reduced dataset representation that‟s much smaller in volume, however produces constant analytical results.
6. **Building the model**: in this step we elect and implement the appropriate data mining task (ex. association rules, serial pattern discovery, classification, regression, clustering, etc.), the data mining technique and also the data processing algorithm(s) to create the model.
7. **Interpretation of the discovered knowledge (model /patterns):** The interpretation n of the detected pattern or model reveals whether or not the patterns are interesting. This step is additionally known as Model Validation/ Verification and uses it to represent the result in an appropriate approach so it may be examined completely
8. **Decisions / Use of Discovered Knowledge:** It helps to make use of the knowledge gained to take better decisions [7]

4. THE PROPOSED APPROACH

The aim of this project is to improve the current trends in the higher education systems and to find out which factors might help in creating successful students. It is really necessary to find successful students as it motivates higher education systems to know them well and one way to know this is by using valid management and processing of the student's database.

❖ **Classification**
Classification algorithm is a data mining technique that helps us to map data into predefined category. It is a supervised learning technique which needs categorized training data so it can creating rules for categorizing test data into pre-arranged category. [2] Its a 2 phase process. The first phase as the learning phase, where the classification rules are generated and training data is analysed. The second phase as the classification phase, where test data is classified

into predefined groups according to the generated rules. Since classification algorithms requires predefined classes based on values of information component, we had created an component "performance" for all students, for which they may have a value of either "Good" or "Bad".

4. TOOLS AND METHODOLOGY

A. Models

We are using four kinds of classification models so as to learn the predictive function which is required. The models are used for experimental analysis. They are selected on the basis of their frequent usage in the existing literature. The list of methods are as follows:

1) Decision Tree

A decision tree is a tree in which each branch node will represent a choice between several alternatives and each leaf node will represent a decision. A decision tree is commonly used for obtaining information so as to fulfil the purpose of decision making. Decision tree starts from a root node which is there for users to take actions. From root node users split each and every node recursively into different nodes according to decision tree learning algorithm. The final result is a decision tree where each branch represents a possible context of the decision and its outcome.

2) Naive Bayes

Naive Bayes algorithm is actually based on the probability theory, i.e. the Bayesian theorem [3] and is a simple classification method. It is named as naive because it solves problems based on two critical assumptions: it assumes that there are zero hidden components that will affect the process of analysing and it supposes that the prognostic components are conditionally independent with similar classification. This classifier provides an efficient algorithm for data classification and it represents the promising approach to the discovery of knowledge.

3) Support Vector Machine

Support Vector Machine is used for classification which is also a supervised learning method. There are three research papers that have used Support Vector Machine algorithm as their technique to analyse student's performance to review it thoroughly. Hamalainen et al. (2006) had chosen Support Vector Machine as their analysing method because it suited well in small datasets. [4] Sembiring et al. (2011) demonstrates that Support Vector Machine algorithm has a good ability of performing generalization and is actually found faster than other algorithms. [5] At the same time, the study done by Gray et al (2014) explained that Support Vector Machine algorithm acquires the highest analysing accuracy in identifying student's performance (Failing Risk). [6]

4) K-Nearest Neighbors

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm

B. Data Description

- Data source link: *http://archive.ics.uci.edu/ml/datasets/student+performance*
- Data format: Integer
- Size: 396 rows X 33 columns
- Number of Instances: 396
- Number of Attributes: 33

This data is of student's achievement in secondary education of Portuguese school. The data attributes include student grades, demographic, social and school related features) and it was collected by using questionnaires and school reports. Dataset are provided regarding the performance in subject: Mathematics. The target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade, while G1 and G2 correspond to the 1st and 2nd period grades.

During the data pre-processing set we found out that data present in our dataset was clean, as a result we did not had to perform the data cleaning methods.

In our dataset we had 33 attributes and as result we had to reduce some of the attributes which were not so important, to get better accuracy and low-cost tree. In organizations these kind of strategies is performed to reduce the data, so we also decided to do the same.

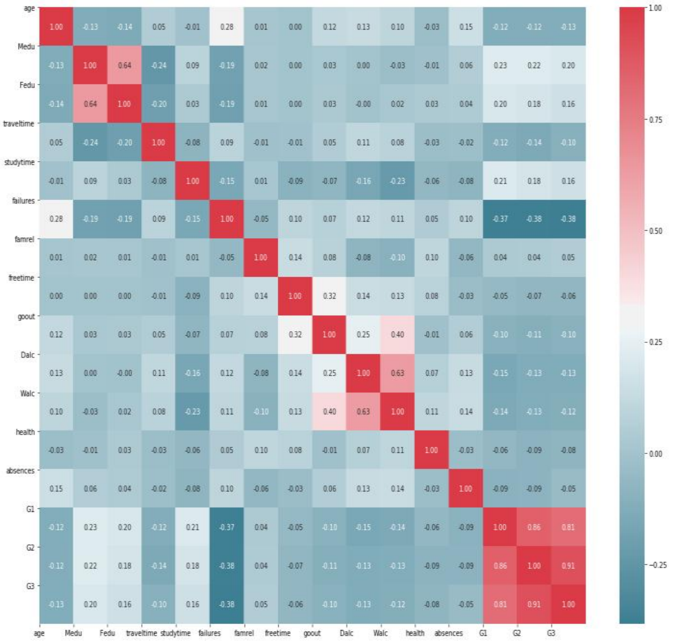
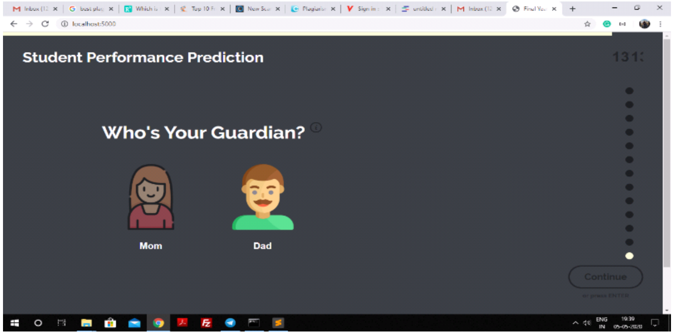


Fig Correlation Heatmap

5. RESULTS

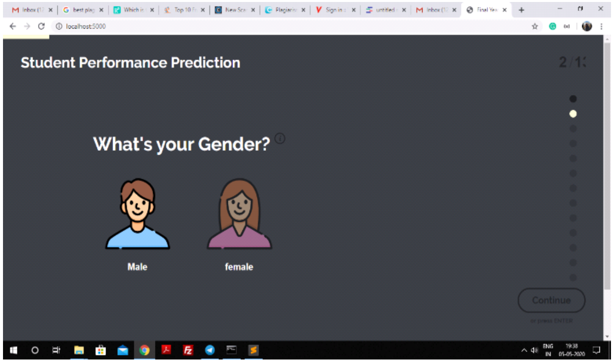
We have implemented our algorithms with the help of Python. We have made use of in-built python libraries and packages to implement our classification algorithms We have made use of the following libraries and packages:

1. Numpy
2. Pandas
3. Scikit-learn
4. Matplotlib
5. Flask



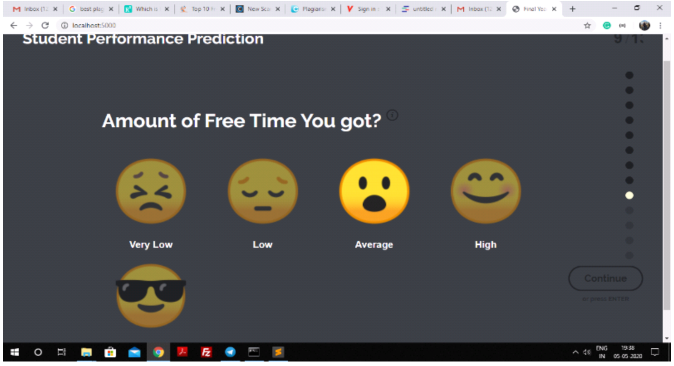
Student Performance Prediction

Who's Your Guardian?



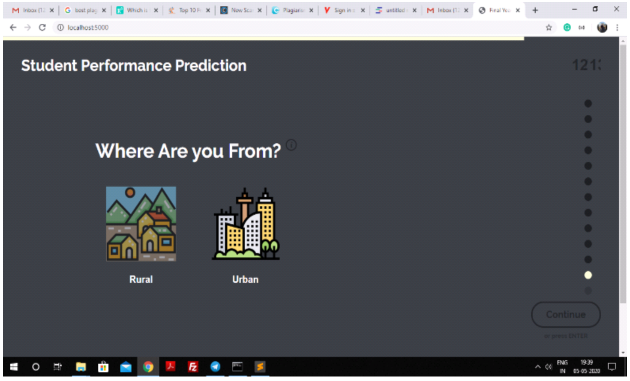
Student Performance Prediction

What's your Gender?



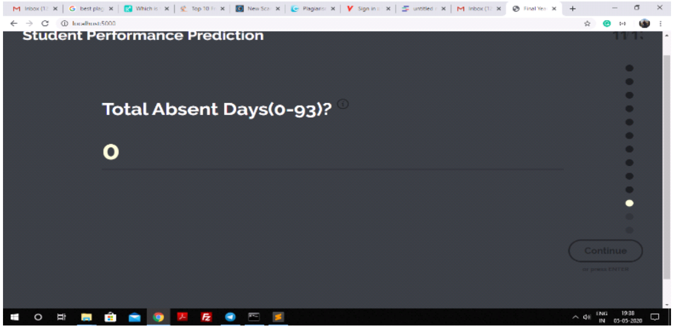
Student Performance Prediction

Amount of Free Time You got?



Student Performance Prediction

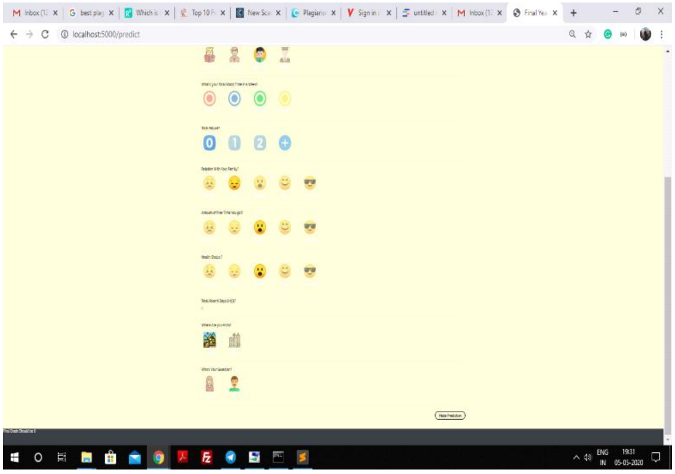
Where Are you From?

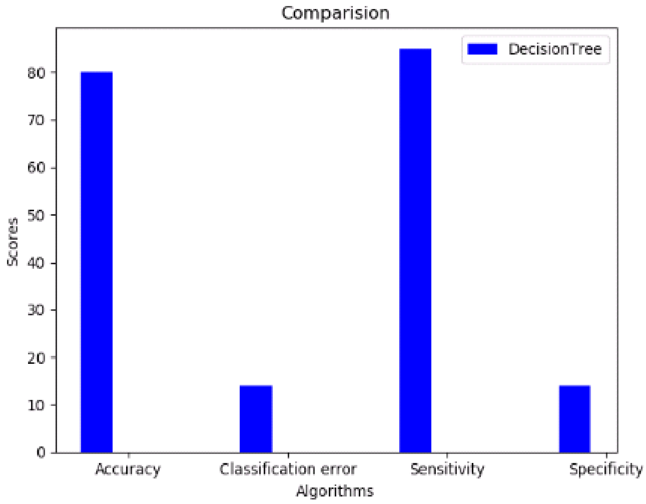


Student Performance Prediction

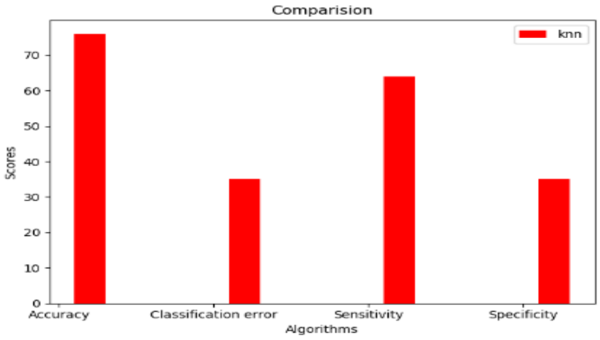
Total Absent Days(0-93)?

0

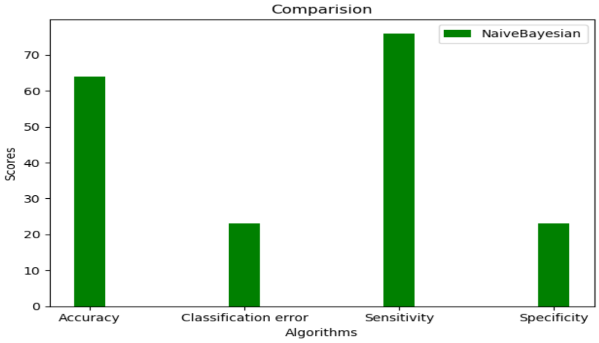




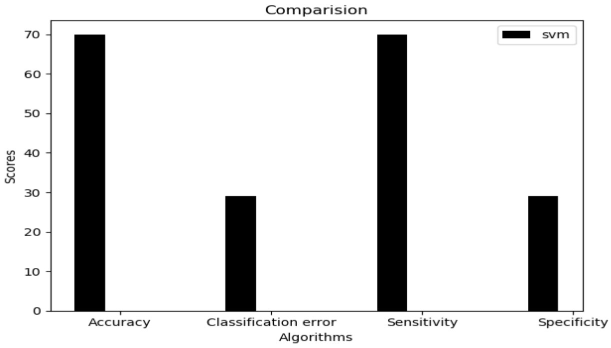
(a) Decision Tree



(b) K-Nearest Neighbors



(c) Naïve Bayesian



(d) Support Vector Machine

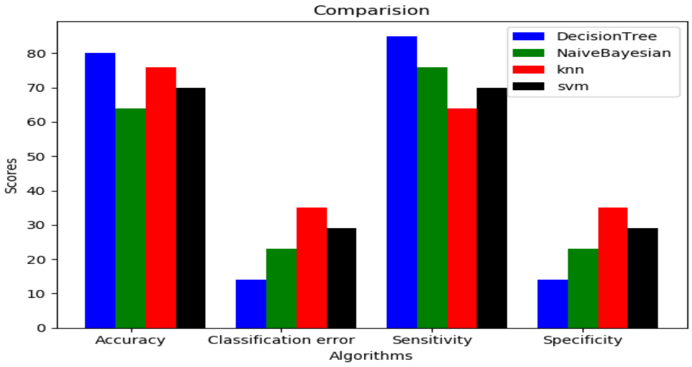


Fig Comparison of Models

Models	Decision Tree	KNN	Naïve Bayesian	SVM
Accuracy of test dataset	0.79	0.72	0.64	0.70
Error rate	0.21	0.27	0.35	0.29
Sensitivity	0.78	0.72	0.64	0.70
Specificity	0.21	0.27	0.35	0.29

Fig Comparison Table

Thus, we Successfully Implemented all the four algorithms and noted the accuracy, performance of each one. After comparing the accuracy we came to conclusion that Decision Tree is the best suitable algorithm for this dataset.

Environment

We run the experiments on the 4 GB RAM PC, with 1.90GHz of Intel i5 Processor. In evaluating this models, we used python Programming. We split the data into two parts, train data set containing 70% of the data and test data set containing the remaining 30%

6. CONCLUSION

In this research, an effort is made to find the impact of our proposed features on student performance prediction with the help of classification models. A feature space is constructed by considering characteristics of family expenditure, family income, personal information and family assets of students. The potential/dominant features selection is unavoidable as it provides us with a subset of features. By using Decision Tree classification algorithm we found our analysis very effective for our proposed features of family expenditure and student personal information categories. It can be easily derived from the results we got that academic information, family details and personal information have very strong impact on the students' performance due to instinctive reasons provided in discussions. The meta-analysis on analysing student's performance has encouraged us to carry out further examination to be applied in our educational institutes. Hence, Educational system can take the help of this model to review the student's performance in a suitable manner.

7. FUTURE WORK

This experiment can be done with more components to get more accurate outputs which will be useful for improving the results of students learning process. Also the experiments can be done by using some other technologies for getting a broader approach and more accurate results. Some different tools can be used while at the same time different factors will be used. Most of the educational institutes mostly find it difficult to provide skilful employee to the society. Many universities/institutes are not in the state that they can provide proper learning environment because of lack of information and lack of proper guidance. To better administer and serve student population, the universities/institutions need better assessment, analysis, and prediction tools. A considerable huge volume of examination is done in field of analysing student performance but all these are detached. So, it is easily understood that a combined approach is needed. Other than academic attributes, there are other components also which are responsible for students overall performance like personal and emotional stability. So proper data mining techniques are used to analysing the existing components and then classifying them in order to provide relevant results or outcomes. Hence if all factors and components are considered for the analysis, it can effectively increase the prediction model accuracy

8. REFERENCES

[1] D. Kabakchieva, "Student performance prediction by using data mining classification algorithms", International Journal of Computer Science and Management Research, vol. 1, 2012.

[2] K. V. J.K. Jothi Kalpana, "Intellectual performance analysis of students by using data mining techniques", International Journal of Innovative Research in Science, Engineering and Technology, vol. 3, 2014.

[3] V. Ramesh, "Predicting student performance: A statistical and data mining approach", International Journal of Computer Applications, vol. 63, no. 8, 2013.

[4] D. A. M. Dr. Abdullah AL-Malaise and M. Alkhozae, "Students performance prediction system using multi agent data mining technique", International Journal of Data Mining and Knowledge Management Process, vol. 4, no. 5, 2014.

[5] P. Kavipriya, "A Review on Predicting Students‟ Academic Performance Earlier, Using Data Mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016.

[6] Shruthi P, Chaitra B P, "Student Performance Prediction in Education Sector Using Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 3, March 2016.

[7] Humera Shaziya, et.al. "Prediction of Students Performance in Semester Exams using a Naïve bayes Classifier", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 4, Issue 10, October 2015.