International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT20653

Prediction and Analysis of Student Performance in Secondary Education Based on Data Mining and Machine Learning Techniques

Meenal Joshi[1], Shiv Kumar[2]

[1]M.Tech Scholar, Computer Science and Engineering, Mewar University, Gangrar, Chittorgarh, Rajasthan, India
[2]Computer Science and Engineering, Mewar University, Gangrar, Chittorgarh, Rajasthan, India

Article Info
Volume 6, Issue 5
Page Number: 294-301
Publication Issue :
September-October-2020

Article History
Accepted : 15 Sep 2020
Published : 23 Oct 2020

ABSTRACT

According to modern era education is the key to achieve success in the future; it develops a human personality, thoughts, and social skills. The purpose of this research work is to focus on educational data mining (EDM) through machine learning algorithms. EDM means to discover hidden knowledge and pattern about student's performance. Machine learning can be useful to predict the learning outcomes of students. From last few years, several tools have been used to judge the student's performance from different points of view like the student's level, objectives, techniques, algorithms, and different methods. In this paper, predicting and analyzing student performance in secondary school is conducted using data mining techniques and machine learning algorithms such as Naive Bayes, Decision Tree algorithm J48, and Logistic Regression. For this the collection of dataset from "Secondary School" and then filtration is applying on desired values using WEKA, tool.

Keywords : Naive Bayes, J48, Logistic Regression, Classification, Prediction, WEKA

I. INTRODUCTION

"Educational Data Mining actually refers to the methodology designed for analysis of the data from the particular learning environment to better understand students and assess the student learning performance." (International Educational Data Mining Society, 2011).

Nowadays, increasing awareness for Artificial Intelligence stimulate the development of data mining and analytics in the student domain (Jesse Tetsuya, 2019). This research paper analyzes the correctness of classification techniques for predicting the student's performance. The performance of every individual student is obtained by this method and it reduces the time for evaluation. It is more helpful for education institutions and training centers. According to various aspects, including methods (classification, clustering, association, etc.) and performance metrics (accuracy, mean absolute error, etc.). The important thing in this research work is that choosing the most suitable method for predicting students overall performance with the help of historical data.

A. Data Mining

Data mining are procedure of extracting essential knowledge or information from the huge amount of data. The main cause for which DM algorithms are used is that it collects relevant information which provides us better results. Nowadays, data mining plays an vital role in education area. Some steps and functionalities of data mining are Data Cleaning and Integration, selection, Preprocessing, Transformation, Mining, Pattern Evaluation, and Presentation.

Techniques of Data Mining:

Classification: It is the method to categorize the data in different groups. The various classifications algorithms are:

- J48 Algorithm
- Naive Bayesian Algorithm
- Random Forest

Regression: Regression is the supervised learning statistical method to identifying the correlations among variables. Basically it works on two types of variables - independent variable and dependent variable. Algorithms are:

- Linear Regression
- Logistic Regression

Clustering: Clustering is a unsupervised learning method of data mining in which the data is grouped into similar items together. Clustering shows the differences and similarities between the data. Some clustering algorithms are:

- K means,
- K-modes
- Hierarchical clustering

Association Rule Learning: This mining method is based on rules for finding relationship between data items. It helps to mine frequent patterns that occur in the data. Algorithm of association rules are:

- Apriori

Prediction: Prediction is a very powerful technique of data mining that represents future possibilities based on current or historical data.

- Naïve Bayes
- Logistic Regression

Machine Learning

Machine learning is foremost technique of Artificial Intelligence. It is the study of computer algorithms which provides machine ability to learn automatically and improve from experience without being explicitly programmed. In machine learning usually the dataset split into two subsets :

- **Training dataset (80% part of the original data):** This dataset is used to build up a model.
- **Testing dataset (20% part of the original data):** Testing phase is used to assess or evaluate the final model performance.

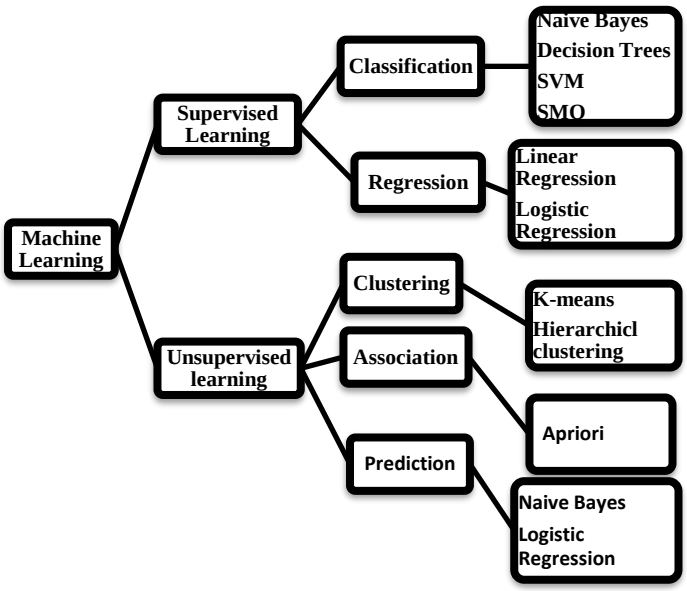

Fig.1 Taxonomy of machine learning algorithms

II. LITERATURE SURVEY

Manojit Debnath & et.al, 2010 coducted a case study in agartala municipal council area for surveying the government aided schools students .It is based on primary data of annual exam which is related to student performance. Regression method used to testing the different parameters which affect the students performance in academics [1].

Bharadwaj & Pal in 2011 conducted a paper for applied ID3 algorithm for classification on student's dataset to analyze Student's performance for identifying the failure rate and students who need special attention and allow the teacher to provide appropriate advising [2].

S.K Yadav & et.al in 2012 applied ID3, CART and C4.5 decision trees algorithms to predict the performance of engineering students in final examination to identify the total number of students result showing likely pass or fail [3].

V.Ramesh & et al. in 2013 conducted a research study on statistical and data mining approach. Purpose of this paper is to discover the factors affect the performance of students in final examinations. The survey cum experimental methodology was taken on to generate database and it was
build from primary and secondary source [4].

Irfan Ajmal Khan & et.al in the 2014 research build a model to predict the success rate of scholarships. Some decision tree algorithms are used for comparisons such as J48, C4.5 and ID3. Dataset of students were classified for the scholarship which is evaluated by "IF-THEN" Rules and scholarship calculator [5].

(Haris Agic & et al, 2014) describes CRISP-DM modeling for dealing with specific problem related to education field. SMOTE function is used to escape unbalanced distribution of the class variables. Four algorithms investigated in this paper C4.5, Multilayer Perceptron, Random Forest, Naïve Bayes. These algorithms is used for create a classification model. Expert presents low dimensional complexity for KDD analysis. [6].

Parneet Kaur & et.al in the ICRTC-2015 conducted a research study on "Classification and prediction-based data mining algorithms to predict slow learners in the education sector" analyzed 152 student's performance and also slow learners within them with the help Weka tool. He also stated that various new factors can be introduced to bring improvement in the student's performance in their learning as well as retention capabilities among them [7].

Amirah Mohamed Shahiria & et.al in 2015 presented a paper "A Review on Predicting Student's Performance using Data Mining Techniques" to provide an overview on the data mining techniques basically on Decision Tree, Neural network, Naive Bayes, K-Nearest Neighbor, and Support Vector Machine to predict students' performance [8].

Ahmed Ashraf & et.al in July-sept 2017 presented a paper to improve secondary school student's e-learning system by adopting DM models. Applying Linear regression, SVM, Decision tree, M5- rules. Mathematics is used for prediction [9].

Sagardeep Roy & et.al in October 2017 presented a paper to review the use of Learning Analytics and Educational Data Mining for examine the performance of students using statistical data mining techniques and machine learning algorithms [10].

(Nitin Umesh, & et.al, 2018) used 1735 instances with 37 attributes from B.Tech second year. Applying different classification algorithms (J48, Naïve Bayes, FL etc.) and LMT to identify the most relevant attribute and removed the less relevant attribute. Fuzzy logic used for prediction of student performance. [11].

Prayuk Chaisanit & et al, in June 2019 conducted a study to explore the relations between emotional skills of the students and previous academic outcome for students performance prediction on basis of two data mining techniques - classification and clustering [12].

III. PROBLEM STATEMENT

Failure of students may be due to some factors such as lack of teaching skills, shortage of useful tools, lab support etc. that influence the student performance. There are three components are required for prediction: Attributes affect the student Performance, Data mining techniques and tools. Nowadays, government; local or private bodies are governing secondary schools. Lack of equipments with adequate IT infrastructure, namely adequate number of computers, projectors, power backups and internet connectivity secondary school is the major problem of education system.

IV. THE MAIN OBJECTIVES

- Raw data collection from the student database of secondary school.
- Extract meaningful information from the selected data through data mining methodologies.
- Analyze and classify the student data to divide them into different groups.
- Classification over the dataset through Navie bayes and J48.
- Perform prediction of result using Logistic Regression algorithm.
- Implementation of algorithms will be done by using WEKA tool.

Required Software (Weka)

Weka is abbreviated as Waikato Environment For Knowledge Analysis is an open source data mining and machine learning software that can be accessed through a graphical user interface (GUI) or Java API. Weka tool supports large number of algorithms and very large datasets. It contains several tools such as data pre-processing, clustering classification, regression, association, visualization and prediction. Weka widely used in teaching and research areas. In this study Weka toolkit 3.8.4 is used for generating classification and prediction.

V. PROPOSED SYSTEM

Proposed model is based on data mining and machine learning algorithms. It describes the flowchart that provides the baseline for data collecting and data processing from secondary school.

Analyse and classify the major factors that influence the student performance with the help of most suitable data mining algorithms J48 decision tree algorithm and naive bayes .

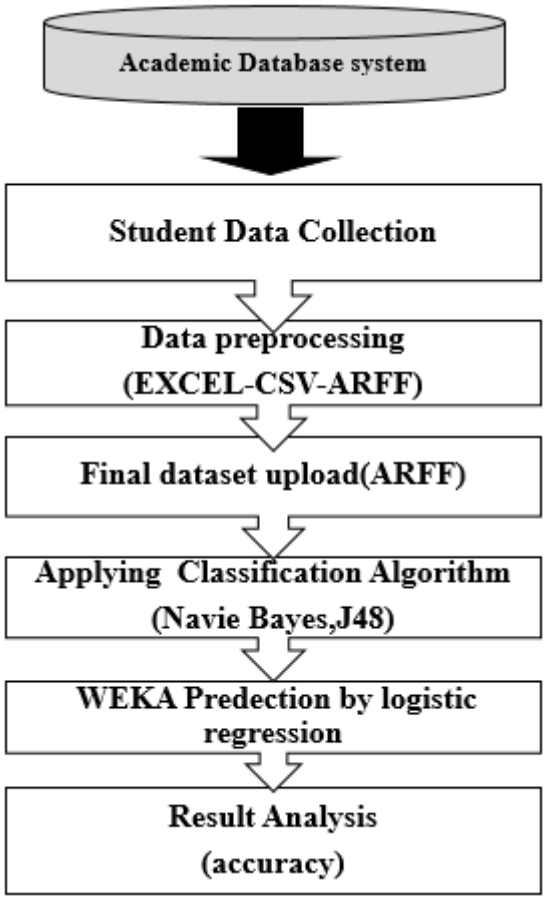

Fig .2 Methodology of Proposed System

A. Dataset

A dataset of 150 students is collected from secondary school Udaipur.

Table.1 Student Data collection

Parameters	Description	Value
Gender	Gender	{M,F}
Nationality	Nationality	{India,Australia,Jordan ,USA}
PlaceofBirth	PlaceofBirth	{Jaipur,Patna,chittorgarh,Udaipur}
StageID	StageID	{lowerlevel,Middleschool,Highschool}
GradeID	GradeID	{G-02,G-04,G-05,G-06,G-07,G-08,G-09,G-10,G-11,G-12}
SectionID	SectionID	{A,B,C}
Age	Age	{numeric: 15 to 201}
Behavior	Student behavior	{Good,Average,Poor}
Hometown	House location	{Rural ,Urban}
Atd	Attendance	{Above-75,Below-75}
Medium	Medium	{English, Hindi}
Unit test	Unit test marks	{I>60%,II>45%&<60%, III>36%&<45%,Fail<36%}
Half yearly	Halfyearly marks	{I>60%,II>45%&<60%, III>36%&<45%,Fail<36%}
Parentsansweringsurvey	Parentsansweringsurvey	{Yes, No}
Parentsschoolsatisfaction	Parentsschoolsatisfaction	{Good, Bad}

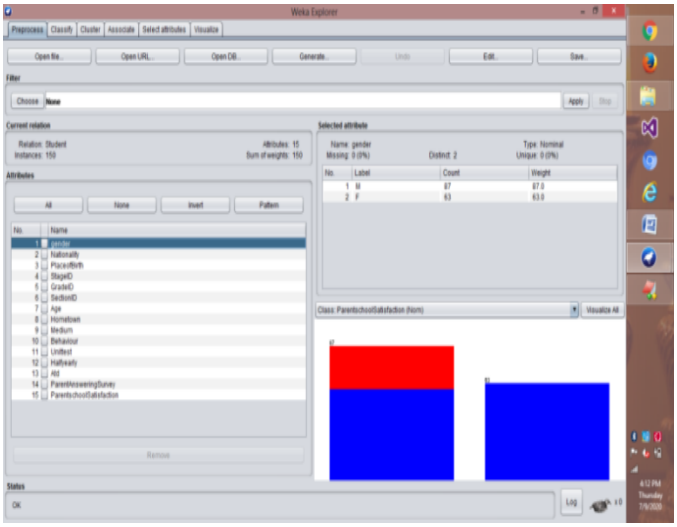

Fig. 3 Data set file upload (preprocessing)

VI. CLASSIFICATION

- **J48**

J48 is an enhanced form of C4.5 and ID3 algorithm which is introduced by Ross Quinlan. It is used to generate decision tree for classification of data so, it is called statistical classifier. J48 handles training data with missing value of attribute.

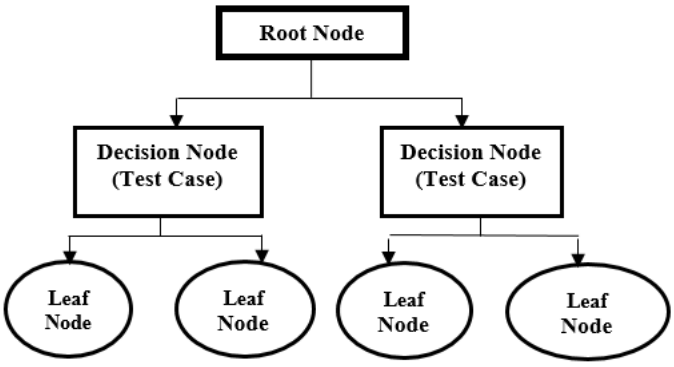

Fig. 4 Graphical Representation of Decision Tree

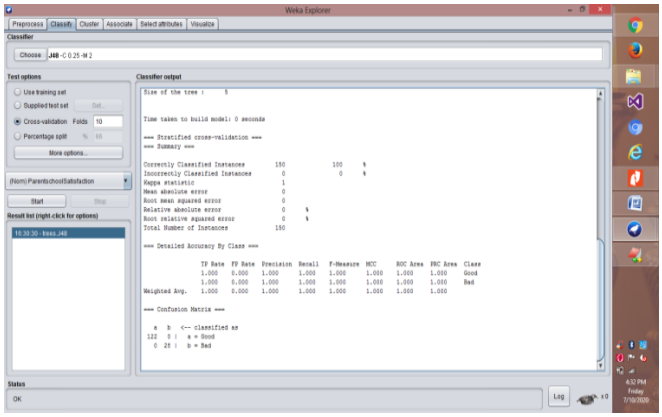

Fig.5 J48 Classification

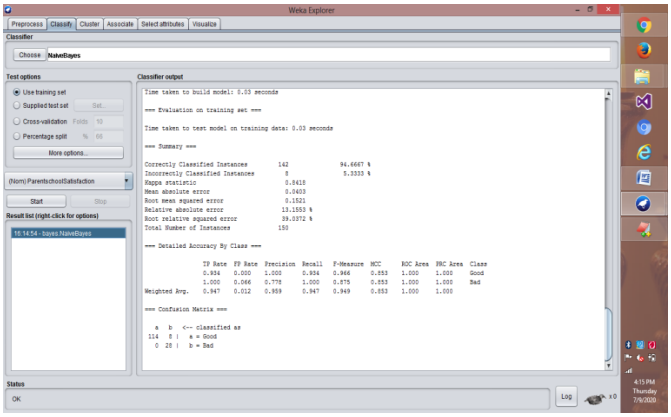

Fig.7 Navie Bayes Classification

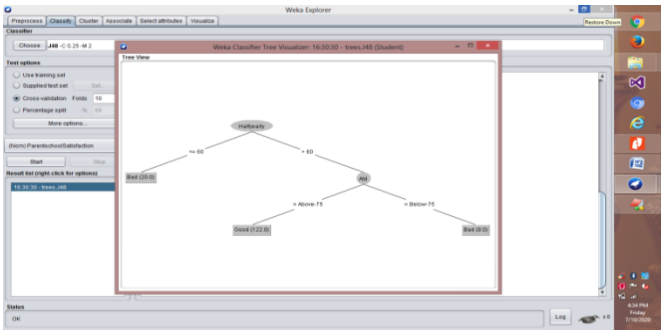

Fig.6 Visualize tree in J48

- **Naive Bayes**

Naive Bayes comes under the supervised learning algorithm. It is a simple probability classification method. This algorithm is based on Bayes theorem following equation:

Likelihood prior probability of Class

$$P(C|A) = \frac{P(A|C)*P(C)}{P(A)}$$

Posterior Probability of Class Predictior prior probablity

Here, C represents the class eg . student , Weather etc. A represents the attributes calculated individually.

VII. WEKA PREDICTION

A. Logistic Regression

This algorithm predicts the probability of target variable or outcome for binary classification. It is used when dependent variable is dichotomous; the nature of target or dependent variable is bifurcate, which means there would be only two possible classes. In simple words, the dependent variable is binary in nature having data coded as either 1 (Positive) or 0 (Negative). The goal of logistic regression is to find the best fitting model for independent and dependent variable relationship.

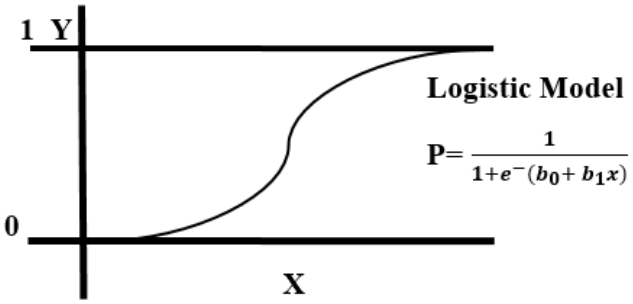

Fig.8 Logistic Regression

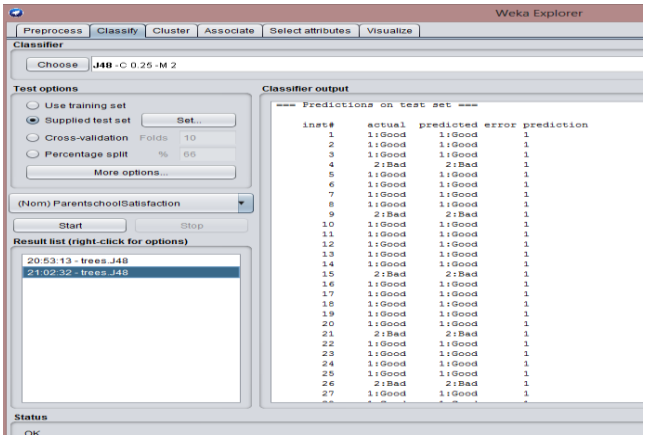

Fig.9 Prediction by Logistic Regression

VIII. RESULT & ANALYSIS

A. Accuracy

Accuracy can be evaluated by using following formula:

$$\textbf{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

B. Precision

Precision is also known as Positive Predictive Value (PPV)

Precision calculated as follows:

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

C. Recall

It is also called True Positive Rate (TPR) or sensitivity. The formula of Recall is given below:

$$\textbf{Recall} = \frac{TP}{TP + FN}$$

Table. 2 Accuracy percentage

Algorithm used	Calculation	Accuracy
Naive Bayes	$\frac{115 + 28}{115 + 28 + 7 + 0}$	**95%**
J48	$\frac{122 + 28}{122 + 28 + 0 + 0}$	**100%**
Logistic Regression	$\frac{(121 + 28)}{121 + 28 + 1 + 0}$	**99%**

Table.3 Overall Performance Accuracy Between Classifiers

Overall Performance Accuracy Between Classifiers		
Startified cross validation	Navie Bayes	J48
Total no. of instances	150	150
Correctely classified instances	95.33%	81%
Kappa statics	0.8598	0
MAE	0.0439	0.3067
RMSE	0.1639	0.3899
RAE	14.3137%	100%
RRSE	42.0276%	100%

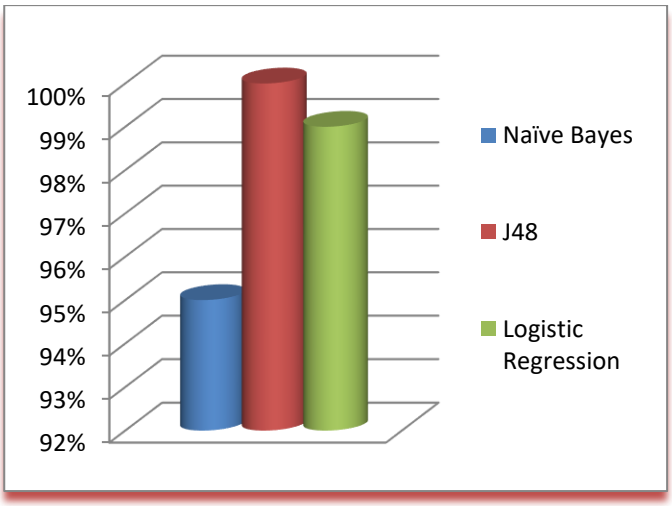

Fig.10 (Table.2) Accuracy Chart

IX. COCLUSION

Educational data mining is completely appropriate method to do analyze and predict the academic performance of students by considering various performance factors. Survey shows that Naive Bayes

classifiers and J48 decision tree classification better performs as compared to other classifiers. Prediction of student's performance is most common research fields in the educational data mining. Logistic Regression is to be most suitable algorithm for binary classification and prediction. This system will provide platform to develop predictive model to analyze student performance using parameters –Attendance, progressive assessment and overall performance. The purpose of this study is to identify at risk students in terms of failure and it will help to reduce the dropout ratio and improve the performance level of the school. So, it is the motivating concept for students who may fail in the final examinations.

X. REFERENCES

[1] Manojit Debnath & et.al, " Factors affecting students academic performance", e-journal of sociology, vol.7, No.2, July 2010.

[2] Bharadwaj & Pal, "Mining Educational Data to Analyze Students' Performance", IJACSA, vol.2, no.6, 2011.

[3] Surjeet Kumar Yadav & et.al "Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification", WCSIT, vol. 2, No.2, 2012.

[4] V.Ramesh & et al, "Predicting student performance: A statistical and data mining approach", Volume 63, No.8, February 2013.

[5] Irfan Ajmal Khan & et.al, "An Application of Educational Data Mining (EDM) Technique for Scholarship Prediction", IJSEIA, vol. 8, No.2, January 2014.

[6] Edin Osmanbegovic, Haris Agic, Mirza Suljic, "Prediction of students success by applying data mining algorithms", JATIT, Vol. 61, No.2 , March 2014.

[7] Parneet Kaur & et.al, "Classification and prediction-based data mining algorithms to predict slow learners in education sector" ICRTC, 2015 .

[8] Amirah Mohamed Shahiria & et.al, "AReview on Predicting Student's Performance using Data Mining Techniques" 2015.

[9] Ahmed Ashraf & et.al, "Improving secondary school student performance using data mining techniques", vol.5, Issue.3, July-Sept.2017

[10] Sagardeep Roy & et.al, "Analyzing performance of students by using data mining techniques a literature survey", 4th IEEE UPCON, October 2017.

[11] Amandeep kaur ,Nitin Umesh and Barjinder Singh title "Machine learning approach to predict student academic performance", IJRASET, Vol.6, Issue.4, April 2018.

[12] Mohammed Afzal Ahamed & et, "Performance of student prediction", IJCSMC, Vol.8, Issue.6, June 2019.

Cite this article as :

Meenal Joshi, Shiv Kumar, "Prediction and Analysis of Student Performance in Secondary Education Based on Data Mining and Machine Learning Techniques", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6 Issue 5, pp. 294-301, September-October 2020. Available at doi : https://doi.org/10.32628/CSEIT20653

Crossref Cited-by

Journal URL : http://ijsrcseit.com/CSEIT20653