International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Prediction of Student Performance System using Machine Learning Techniques

J. Preethi, S. Maheswari

Abstract: ***Educational organizations are unique and play the utmost significant role in the development of any country. In the Educational database, due to the enormous volume of data for predicting student's achievement becomes more complicated. To upgrade a student's performance and triumph is more efficient in a practical way using Educational Data Mining Techniques. Data Mining Techniques could deliver favor and brunt to educators and academic institutions. The student's data ((i.e.) Name,10th %,12th cut off, CGPA, No of arrears, etc.) are gathered. Then, the datasets are imported into the Anaconda Navigator. Then, analysis and classification based on attributes of the students and the schemes are performed. Then using the prediction algorithm Naïve Bayes what are all the features the particular student is eligible for are predicted as placed. The student's input that has disparate data about their past and present academics report and then apply the Naïve Bayes algorithm using Anaconda Navigator to search the student's achievement for placement. A proposed methodology based on a classification approach to finding an improved estimation method for predicting the placement for students. This project can find the association for academic achievement of each particular student and their placement achievement in campus selection.***

Keywords**: *Classification. Naïve Bayes. Placement. Prediction***

I. INTRODUCTION

After finishing the course, the prediction of students where they can be placed in campus will guide to boost the development of students for appropriate achievement. This project can find the association for the academic progress of each particular student and their placement achievement in campus selection.

Nowadays, most of the students who completed their higher education will join the universities for attaining a secured job. Educational institutions contain a massive number of student's documents. Therefore, finding the impression and nature of this enormous amount of student's data is not a challenging task. Higher education is classified into professional skills and non-professional skills.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
Preethi J*, Computer Science and Engineering, National Engineering College, Tamil Nadu, India. E-mail: 182805@nec.edu.in
Dr. S. Maheswari, Computer Science and Engineering, National Engineering College, Tamil Nadu, India. E-mail: maheswaricse@nec.edu.in

Professional skills grant expert information to students so that the students can take their place in their carrier field. Non-Professional skills allow technology-oriented and focus on developing the managerial skills of the students.

There are several types of jobs, like Project Developer, Designer, Tester, Manager, etc. All these jobs desire some vital information to get placed on campus. So, recruiters consider intelligence, capability, and concern the candidate suited for the right job.

Here, the prediction system makes its placement tasks simple as a result of the given input data.

The opening portals access the pupil professionally and advise them for placement appropriate to their achievement. Different factors, including the capabilities of students in academic details and extracurricular, are captured into consideration. After finishing the course, the prediction of students where they can be placed on campus will guide the development of students for appropriate achievement.

It promotes the teacher's attention towards the progress of each pupil separately and gives renowned fame to the institution in the field of education and extracurricular activities.

Classification of abundant records on a large scale can be easily engaged with a cluster of pre-classified parameters, which enhances the use of Data Mining techniques in a decision-based tree algorithm. Learning and classification are the two crucial schemas in defining a classification procedure.

In the concept of machine learning, the classifier algorithm plays a vital role in training the dataset. If the condition is feasible, the appropriate dataset is progressed to be simulated. Secondly, the required parameters are trained and classified using the classifier training method. Finally, these parameters are encoded into an approach called the classifier approach is Naïve Bayes Algorithm accentuates the performance metrics of the dataset with high accuracy and efficiency. It is well preferred to classify and fragment the trained dataset.

A. Naïve Bayes Algorithm

A well-organized approach to build classifiers with less complexity is Naïve Bayes Algorithm. It is a generic approach to predict instances of rare case problems that are represented by vector notations. One main important key factor to be considered in this model is that some specific attributes are independent of the quality of the variable in a class. For instance, a fruit may be considered as orange if it is orange in color, round, and about 10 cm in diameter.

A Naive Bayesian algorithm contemplates each of these attributes to donate unconditionally to the probability that this fruit is an orange, regardless of any possible correlations between the color, roundness, and diameter features.

As Naïve Bayes is a filtering skilled prediction algorithm, it can depict the hinder probability of numerous classes to the aimed variable. Independent rule with a high positive rate in spam filtering and sentimental analysis is a new emerging classification algorithm in real-world entities. Accurate result analysis for datasets with unknown and incomplete information can be predicted using a collaborative filtering approach, which is a substructure of Naïve Bayes.

II. MATERIALS AND METHODS

A. Jupyter Notebook

Jupyter Notebook is open-source software that enables us to create, enhance, share, and modify codes. Various preprocessing stages, like data cleaning, training, and statistical analysis, can be executed. In this application, each cell is executed separately, which is written in python language. As individual implementation is carried out, the end-user test or run the source code from scratch.

B. Naïve Bayes Algorithm

Naive Bayes algorithm is a supervised learning and statistical technique dependent on the Bayes theorem. Naïve Bayes classifier is an accurate, fast, and reliable algorithm. Due to high efficiency and speed, it is simple to apply the Naïve Bayes classifier on large datasets. As it involves the concept of Bayes theorem, it portraits the relationship between statistical quantities.

III. PROPOSED SYSTEM

The proposed system is used to predict student performance based on prediction algorithm, and knowledge discovery is suggested here to mine rules from the dataset of Systems of Learning Management. In this process, contain five methods are Dataset collection, preprocessing, feature extraction, classification, and performance result and, at first, collecting the student data in the dataset collection phase.

After that, perform the preprocessing to remove the unwanted data also unified the DB. In Feature extraction, the standard deviation and mean for each feature is to be calculated.

In classification, the module calculates the probability of positive and negative classes by using naive Bayes. Finally, build the confusion matrix for calculating accuracy, Precision, Recall, F-score.

A. Dataset Collection

Collecting dataset from two different sources. First, one source is obtained from the student, which dataset contains student personal and educational information, extra skills activity information. This data is unlabeled. The second source is a rule-based logic that is collected from the company.

Table I: Example for Data Collection

Name	Department	10th	12th	CGPA	Arrears
Mallika	ME CSE	92	89	9.2	0
Mallika	ME CSE	92	89	9.2	0
Mano	ME CSE	86.4	192	9.05	0
Devi	CSE	94	82	7.8	0
Shilpa	CSE	87	185	8.9	-
Swetha	CSE	87	-	8.6	0
Pavithra	CSE	94	180	8.9	0
Pavithra	CSE	94	180	8.9	0

B. Data Preprocessing

In the preprocessing stage, remove the unwanted feature from the dataset and missing data filling. After that, apply the unified DB module to the dataset. This module can remove repeated data. Finally, use the fuzzy rule to the Unified DB, which can predict the class. In the preprocessing stage, produce the training data.

Table II: Example for Preprocessing

Name	Department	10th	12th	CGPA	Arrears
Mallika	ME CSE	92	89	9.2	0
Mano	ME CSE	86.4	192	9.05	0
Devi	CSE	94	82	7.8	0
Shilpa	CSE	87	185	8.9	0
Swetha	CSE	87	75	8.6	0
Pavithra	CSE	94	180	8.9	0

C. Feature Extraction

The standard deviation and mean is to calculate for each feature. It contains two types of classes that are placed and not placed. Mean, and Standard Deviation calculation formula is

X=input (1)
Y= Actual Predict [Placed or Not Placed]
Class=Classes in Y
Mean [class][feature]=sum of all feature[class]/ Length [Mean[class][feature]]
Standard Deviation[class][feature]=sum of all(feature[class]-Mean[class][feature])^2 / Length [SD [class][feature]]-1

Table III: Example for Feature Extraction

10th	12th cut off	CGPA	Arrears
92	89	9.2	0
86.4	192	9.05	0
94	82	7.8	0
87	185	8.9	0
87	75	8.6	0
94	180	8.9	0

D. Classification

The classifier depicts the Naive supposition of independence among the trained data sets. Without considering the presence or negligence of a parameter being and related to quality.

Consider S to be the unknown input data

$$S = (s_1, s_2, \ldots s_n) \quad (2)$$

$$P(m/s) = P(s/m)\,P(m)/P(s) \quad (3)$$

Where $P(m/s)$ is the probability in the posterior of the defined class m, for this defined class $P(m)$ is preceding to the probability value. The predictor probability $P(s/m)$ is defined as the possible occurrence to the classified result, and $P(s)$ denotes the numeric predictor probability.

E. Performance Metric

- Confusion Matrix creation by using predicted and Actual Data
- Calculate True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).
- Calculate Precision.
- Determine the recall value.
- Simulate the accuracy of the dataset.

Considering the truly positive data points plotted as true positive, and truly negative data points as true negative. (TP+TN)/(TP+FP+FN+TN) gives the accuracy of the dataset. The precision ratio is provided by the true positive cases to the sum of true positive and false positive claims (TP/(TP+FP)). The capability of the estimator or the sensitivity (SN) identifies applicable instances, and harmonic mean is used to analyze the F1 score, which is a combination of recall and precision ratio.

F. Accuracy of the Dataset

Accuracy of the given dataset defines the sum of all corrected True Positive and True Negative predictions to the total count of the given dataset. The minimum and maximum skill of accuracy are given as 0.0 and 1.0, respectively.

Accuracy=((True Positive + True Negative)/(True Positive + False Positive + False Negative + True Negative)) (4)

G. Precision ratio of the Dataset

The precision ratio is generally labeled to all the true positive predictions to the sum of true positive and false positive predictions. It is also defined as the positive predictive value (PPV) of the dataset. Similar to accuracy, the minimum and maximum value for the precision ratio are 0.0 and 1.0.

Precision = ((True Positive)/
(True Positive + False Positive)) (5)

H. Recall (Sensitivity)

The recall is also generally labeled to all true positive predictions to the sum of true and false negative predictions. It is also denoted by the true positive rate (TPR) of the dataset. For maximum sensible data, the value approached is 1.0, and for the worst sensible data, the value gained is 0.0.

Recall = ((True Positive)/
(True Positive + False Negative) (6)

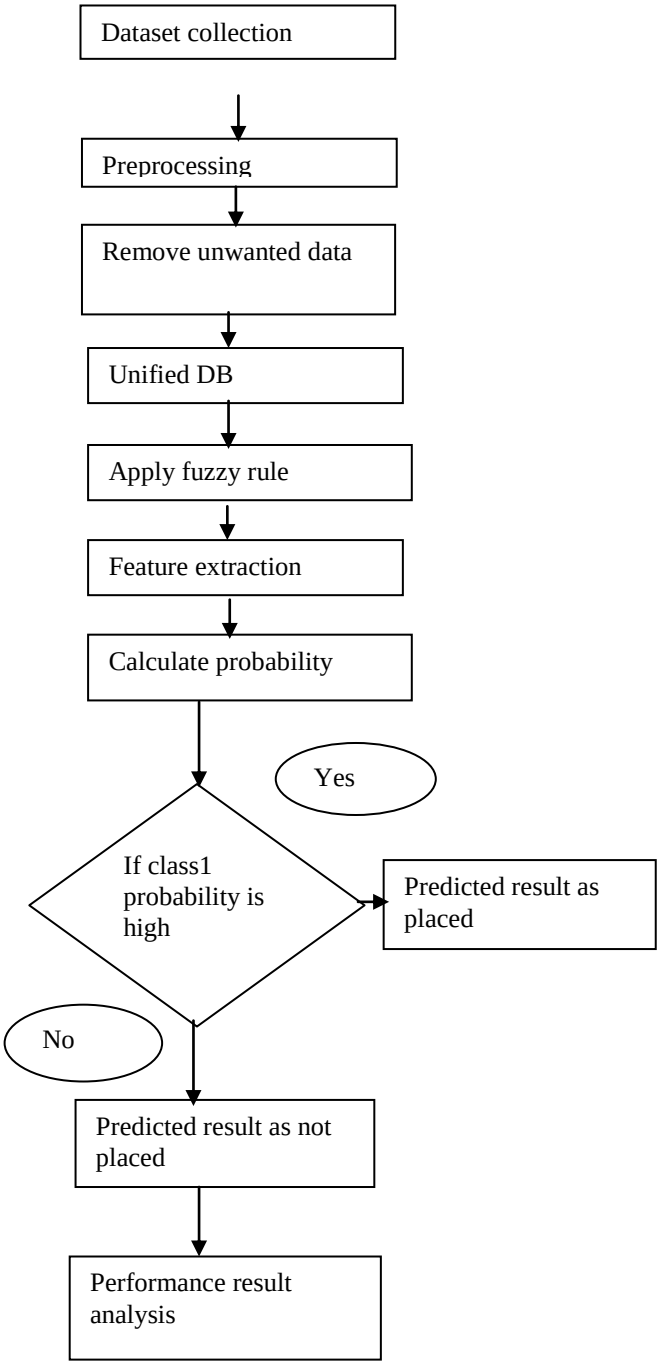

Fig 1. Work Flow Diagram

I. F Measure (F1 Score)

F1 Score is a combined measure of precision and sensitivity. It is defined as two times the product of sensitivity and precision value to the total sum of sensitivity and precision value.

F1Score=2*(Recall*Precision)/
(Recall+ Precision) (7)

IV. IMPLEMENTATION

A. Data Collection

The dataset is collected from a student, which contains student personal and educational information and extra skills activities information.

Table IV. Sample Collected Data

Name	10th	12th	CGPA	Year	Arrear	Department
Mallika R	92	89	9.2	2	0	ME CSE
Mallika R	92	89	9.2	2	0	ME CSE
S. Manoruthra	86.4	192.25	9.05	2	0	ME CSE
Prakashini. S	90	164.75	-	2	0	ME CSE
Malini S L	-	182.75	9.26	4	0	CSE
K Pavithra	94	180	8.9	4	0	CSE
K Pavithra	94	180	8.9	4	0	CSE
Preethi J	86	110	8.5	4	0	CSE
Angelin Swetha	87	-	8.6	4	0	-
Deepa Lakshmi	79.8	135	8.6	4	0	CSE
Lakshmi Priya E	96.4	177.25	-	4	0	CSE
M. Madhan Kumar	93	147	6.52	4	9	CSE
Shyamala Bharathi	98	-	8.35	4	0	CSE
Shilpa Sankar	87	185.25	8.93	4	-	CSE
B Devi Sujatha	94	-	7.89	4	0	CSE
V Valar Mathi	97	176.75	8.1	4	0	CSE
Abirami B	-	192.75	9.76	4	0	CSE
Abirami B	-	-	-	-	-	CSE
Srinandhini M	98	182.75	9.1	4	0	CSE
V Subhashini	97	181.5	8.7	4	0	CSE
Sneha C	95	186.5	8.72	4	0	CSE
Selva Pradeepa R	97	177.5	8.8	4	0	CSE

The dataset is collected from the company. It is a rule-based logic that is received from the company. The company dataset contains 10^{th} %, 12^{th} cut off, CGPA, Communication skill, Person skill, Arrears.

Table V. Company Dataset

Company	10th	12th	CGPA	Communication Skill	Person Skill	Arrears
Hitech	90	180	8.9	1	1	0
CTS	78	170	7.9	1	1	2
TCS	78	156	6.9	0	0	1
Infosys	85	180	8.0	1	1	1

B. Preprocessing

In the preprocessing stage, produce the training data from the collected student's dataset. Remove the unwanted feature from the dataset and missing data filling.

Table VI. Sample Preprocessed Data

Name	10th	12th	CGPA	Year	Arrear	Department
Mallika R	92	89	9.2	2	0	ME CSE
Mallika R	92	89	9.2	2	0	ME CSE
S. Manoruthra	86.4	192.25	9.05	2	0	ME CSE
Prakashini S	90	164.75	8.5	2	0	ME CSE
Malini S L	93.5	182.75	9.26	4	0	CSE
K Pavithra	94	180	8.9	4	0	CSE
K Pavithra	94	180	8.9	4	0	CSE
Preethi J	86	110	8.5	4	0	CSE
Angelin Swetha	87	75	8.6	4	0	CSE
Deepa Lakshmi	79.8	135	8.6	4	0	CSE
Lakshmi Priya E	96.4	177.25	7.83	4	0	CSE
M. Madhan Kumar	93	147	6.52	4	9	CSE
Shyamala Bharathi	98	139	8.35	4	0	CSE
Shilpa Sankar	87	185.25	8.93	4	0	CSE
B Devi Sujatha	94	82	7.89	4	0	CSE
V Valar Mathi	97	176.75	8.1	4	0	CSE
Abirami B	99.40	192.75	9.76	4	0	CSE
Abirami B	99.40	192.75	9.76	4	0	CSE
Srinandhini M	98	182.75	9.1	4	0	CSE
V Subhashini	97	181.5	8.7	4	0	CSE
Sneha C	95	186.5	8.72	4	0	CSE
Selva Pradeepa R	97	177.5	8.8	4	0	CSE

C. Remove Unwanted Data

Remove unwanted features from the dataset and missing data filling. It can remove the repeated data.

Table VII. Sample Student data

Name	10th	12th	CGPA	Person Skill	Arrears
Mallika R	92	89	9.2	0	0
S. Manoruthra	86.4	192.25	9.05	2	0
Prakashini S	90	164.75	8.5	0	0
Malini S L	93.5	182.75	9.26	3	0
K Pavithra	94	180	8.9	5	0
Preethi J	86	110	8.5	1	0
Angelin Swetha	87	75	8.6	5	0
Deepa Lakshmi	79.8	135	8.6	2	0
Lakshmi Priya	96.4	177.25	7.83	3	0
Madhan Kumar	93	147	6.52	2	9
ShyamalaBharathi	98	139	8.35	4	0
Shilpa Sankar	87	185.25	8.93	2	0
B Devi Sujatha	94	82	7.89	3	0
V Valarmathi	97	176.75	8.1	1	0
Abirami B	99.40	192.75	9.76	2	0
M Sri Nandhini	98	182.75	9.1	2	0
V. Subhashini	97	181.5	8.7	2	0
Sneha C	95	186.5	8.72	3	0
Selva Pradeepa	97	177.5	8.8	2	0

D. Unified Database

The unified database contains 10^{th} mark, 12^{th} cut-off, CGPA, No of outside participation, Arrears, and Eligible count for companies.

Table VIII. Predicted Attributes

10th	12th	CGPA	Skillset	Arrears	Companies
92	89	9.2	0	0	2
86.4	192.25	9.05	2	0	2
90	164.75	8.5	0	0	1
93.5	182.75	9.26	3	0	0
94	180	8.9	5	0	0
86	110	8.5	1	0	1
87	75	8.6	5	0	2
79.8	135	8.6	2	2	2
96.4	177.25	7.83	3	2	1
93	147	6.52	2	0	1
98	139	8.35	4	0	1
87	185.25	8.93	2	0	2
94	82	7.89	3	4	2
97	176.75	8.1	1	0	1
99.4	192.75	9.76	2	0	0
98	182.75	9.1	2	0	0
97	181.5	8.7	2	0	1
95	186.5	8.72	3	0	1
97	177.5	8.8	2	0	1

E. Apply Fuzzy Rule

Finally, apply the fuzzy rule to the Unified DB, which can predict the class. The preprocessing data contains predicted attributes, i.e., 10th %, 12th cut off, CGPA, No. of outside participation, and Arrear.

Table IX. Pre Processed Data

Name	10th	12th	CGPA	Person Skill	Arrears	Predicted
Mallika. R	92	89	9.2	0	0	Not Placed
Manoruthra	86.4	192.25	9.05	2	0	Placed
Prakashini S	90	164.75	8.5	0	0	Placed
Malini S L	93.5	182.75	9.26	3	0	Placed
K Pavithra	94	180	8.9	5	0	Placed
Preethi J	86	110	8.5	1	0	Placed
Angelin Swetha	87	75	8.6	5	0	Not Placed
Deepa Lakshmi C	79.8	135	8.6	2	0	Not Placed
Lakshmi Priya E	96.4	177.25	7.83	3	0	Not Placed

F. Feature Extraction

The standard deviation and class mean for each feature is to calculate. It contains two types of classes they are placed and not placed. The predicted attributes of 10th, 12th, CGPA, Communication skill, Person skill, and Arrears.

X=10th,12th cut off, CGPA, Communication skill, Person skill, Arrears

Y=TCS, Hytech, Not placed

Mean(μ)=90+164.75+9.26+0+1+1/6 =44.335

Standard Deviation(σ)=$\sqrt{\sum |xi - \mu|^2}$/N=62.6

G. Calculate Probability

The classification shows that the predicted attributes and find accuracy, precision, recall, and F1 Score.

Calculation for single student:

$$P(m/s) = P(s/m)\,P(m)/P(s)$$

Input: 10th=<90, 12th=<120, CGPA=<6, Communication skill=0, Person skill=0, arrear=2

P(<120| Not placed) = P (Not placed | <120)* P(<120)/ P(Not placed)

P(Not placed|<120)=12/35=0.342

P(<120)=35/163=0.214

P(Not placed)=35/163=0.214

P(<120|Not placed)=0.342*0.214/0.214
=0.342

V. RESULT AND DISCUSSION

The standard deviation (σ) and mean (μ) is determined for the features extracted using the classifier algorithm. Various performance metrics for the predicted attributes are simulated using Naïve Bayes Algorithm.

Table X. Performance Analysis

Algorithm	TP	TN	FP	FN	Accuracy
Naïve Bayes	40	34	17	9	0.74

Accuracy = (40+34)/ (40+34+17+9) = 0.74
Precision =40/ (40+17) = 0.70
Recall=40/ (40+9) = 0.81

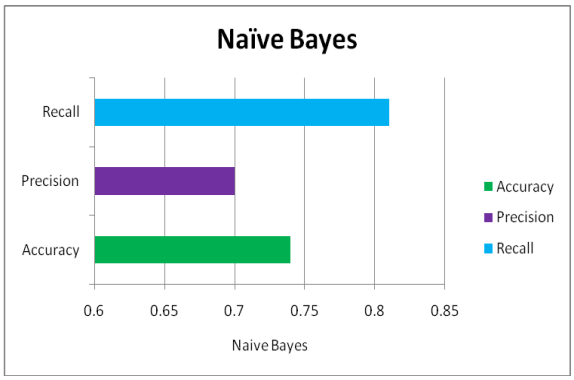

Fig. 1. Prediction Accuracy for Naïve Bayes

The visualized form of accuracy, recall, precision for Naïve Bayes Classifier. Accuracy for Naïve Bayes classifier is 0.74. The precision and recall value for the Naïve Bayes classifier is 0.70 and 0.81, respectively.

F Measure= 2*(0.81*0.70)/(0.81+0.70)= 0.75

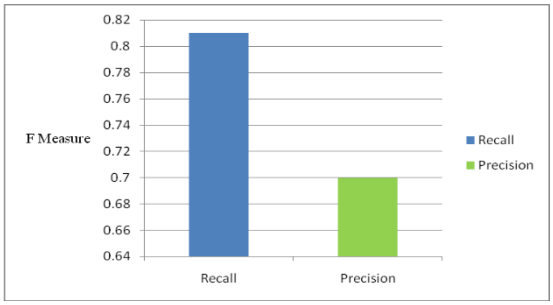

Fig. 2. A Visualized form of F Measure using Recall and Precision

A recall is more significant than precision because the Naïve Bayes classifier gives less false results. Using recall and precision, F Measure is calculated.

VI. CONCLUSION

The main reason to opt decision trees in classification algorithm is that which is easy to predict an interpret compared to the other traditional classification models. As they produced classification rules which are less tedious to implement and analyze, they become more popular in Data Mining techniques. Frequent visualization of decision tree classifier experiments is proceeded to pick out the best response classifier for student's carrier prediction in their outcome.

In this proposed Naïve Bayes technique, the predicted result enables educational officials to recognize and approach the students with high risk in their academic year. Hence, the experimental result depicts that the proposed approach tends to increase the student's behavior in academic and co-curricular activities.

REFERENCES

1. Bashir Khan, Malik Sikandar Hayat Khiyal, et al., " Final Grade Prediction of Secondary School Students using Decision Tree," International Journal of Computer Applications (0975-8887) Volume 115- No. 21 April 2015.
2. Heena Sabnani, Mayur More, Prashant Kudale, Prof. Surekha Janrao, "Prediction of Student Enrolment Using Data Mining Techniques," International Research Journal of Engineering and Technology (IRJET) Volume: 05 Issue: 04 | Apr-2018.
3. Mrs. Varsha. P. Desai, "Classification Technique for Predicting Learning Behavior of Student in Higher Education" International Journal of Trend in Scientific research and Development (IJTSRD) oct-2018.
4. Maria Koutina and Katia Lida Kermanidis," Predicting Postgraduate Student's Performance Using Machine Learning Techniques" International Federation for Information Processing (IFIP) 2011.
5. Amirah Mohamed Shahiri, Wahidah Husain, Nuraini Abdul Rashid," A Review on Predicting Student's Performance using Data Mining Techniques" ELSEVIER/(PCS) Procedia Computer Science 72 (2015) 414-422.
6. Manisha Sahane, Sanjay Sirsat, Razaullah Khan, Balaji Aglave, "Prediction of Primary Pupil Enrollment in Government School Using Data Mining Forecasting Technique" International Journal of Advanced Research in Computer Science and Software Engineering 4(9), September – 2014, pp. 656-661.
7. Amandeep Kaur, Nitin Umesh, Barjinder Singh, "Machine Learning Approach to Predict Student Academic Performance" International Journal for Research in Applied Science and Engineering Technology (IJRASET) Volume 6 Issue IV, April 2018.
8. Havan Agrawal and HarshilMavani, "Student Performance Prediction using machine Learning" International Journal of Engineering Research and Technology (IJERT) Vol. 4 Issue 03, March-2015.
9. Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Students" I. J. Modern Education and Computer Science, 2013, 11, 49-56.
10. Ioannis E. Liveris, Tassos A. Mikropoulos, Panagiotis Pintelas, "A Decision Support System for Predicting Student's Performance" Themes in Science and Technology Education, 9(1), 43-57, 2016.
11. Umesh. Kumar. Pandey, S. Pal, "Data Mining: A Prediction of Performer or Under Performer using Classification" International Journal of Computer Science and Information Technology (IJCSIT), Vol. 2(2), pp.686-690, ISSN: 0975-9646, 2011.
12. Alaa Khalaf Hamoud, "Selection of Best Decision Tree Algorithm for Prediction and Classification of Students Action" (AIJRSTE) American International Journal of Research in Science, Technology, Engineering, and Mathematics SSN (Print): 2328-3491, ISSN (Online): 2328-3491, ISSN (Online): 2328-3580, ISSN(CD-ROM): 2328-3629.
13. Raheela Asif, Saman Hina, Saba Izhar Haque, "Predicting Student Academic Performance using Data Mining Methods" (IJCSNS) International Journal of Computer Science and Network Security, Vol. 17 No. 5, May 2017
14. Manmohan Singh, Harish Nagar, Anjali Sant, "Using Data Mining to Predict Primary School Student Performance" IJARIIE-ISSN(O)-2395-4396 Vol-2 Issue-1 2016
15. Dr. Anjali B Raut, Ms. Ankita A Nichat, " Students Performance Prediction using Decision Tree Technique" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 7 (2017), pp. 1735-1741

AUTHORS PROFILE

Preethi J, currently pursuing a Master's Degree in the Department of Computer Science and Engineering at National Engineering College, Tamil Nadu, India. She received her Bachelor's Degree in Computer Science Engineering from Dr. Sivanthi Aditanar College of Engineering, Tamil Nadu, India.

Dr.S.Maheswari is working as an Associate Professor at National Engineering College, Kovilpatti, India. She has received PhD from, Anna University; Chennai. Her area of interest includes Semantic web service selection and machine learning concepts. Her current research focuses on Semantic web service selection, Machine learning and Computer Networks. She has published papers in national/international journals and conferences and a Reviewer of international journals.