International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

Students Result Prediction using Different Data Mining Classifiers

Prathyakshini[1], Pratheeksha Hegde N[2], Nikitha Saurabh[3]

[1]*Assistant Professor, Information Science and Engineering, NMAMIT, Nitte, Karnataka, India*
[2]*Assistant Professor, Information Science and Engineering, NMAMIT, Nitte, Karnataka, India*
[3]*Assistant Professor, Information Science and Engineering, NMAMIT, Nitte, Karnataka, India*

Abstract - *There is a great concern for the analysis of student performance in educational sectors. Student data can be analyzed to predict the results, placements and the week points of the student can be addressed. So that student can focus on those areas to get better results or placement. For predicting the student result different machine learning algorithms are used on the student data set. Several classification algorithms like Naive Bayes, Support Vector machine (SVM), J48 , Random Forest are used to test the student data.*

***Key Words*: Support Vector machine (SVM), Random Forest, J48, Naive Bayes**

1. INTRODUCTION

It is a tedious task for humans to summarize the huge amount of data in order to get useful information. For this purpose, data mining can be used where the large data is analyzed to form meaningful patterns. Data mining can be used for different activities like Association rules, Classification, Estimation, Description and visualization. Classification, Estimation and prediction are all the examples of supervised learning. Academic success is essential as it determines the positive outcome of a student. Educational data mining is focused on using methods for analyzing student data in order to analyze the student performance [1].

It is also important to know that, there are advantages of using data mining with regard to statistical modeling [12]. There are several stages in data mining including statistics. Data collected is preprocessed, evaluated and finally results are interpreted in Knowledge Discovery process. In this paper the student data set is tested by using weka tool. Weka provides tools for preprocessing also various machine learning algorithms are implemented [13]. It is required to have several considerations on the data in order to train and analyze them. Firstly data should be clean and should not contain any null values. Once the data is preprocessed, you can develop machine learning model by selecting one of the option from classify, cluster and associate. Also there are options to reduce the features and apply the machine learning models on them. There are different test methods supported such as percentile split, training set, testing set and cross validation.

Predicting student performance can be benefited by analyzing about slow and fast learners. Student data can be tested and the results can be categorized such as fast and slow learners, placement prediction, identifying students who are likely to drop out, weak, needs improvement, good in academia but lately deteriorated and so on. There are different types of machine learning techniques such as supervised learning based on the given labelled data for input and output for predicting the future output and in unsupervised learning, the model is trained on the unlabelled data for only input. Data classification can be categorized under supervised learning where data is classified into different classes. In this paper classification algorithms like Naive Bayes, SVM, J48, and Random Tree are utilized to test the student data which is downloaded from Kaggle.

2. RELATED WORKS

Based on the current and previous performance of students fourth year results are predicted[2]. Classification algorithms like C4.5, ID3 and improved ID3 algorithm are compared. 74% accuracy was obtained using improved ID3 Algoritm. In secondary school, the prediction was performed in order to predict mathematical performance using Naive Bayes, Multilayered Perceptron, J48, Decision Tree and Random forest [3]. Data set consists of background of student, coursework result and social activities. Here 395 real data set of students are used by considering 33 attributes. Also two and five level classification were done. Vrushali Mhetre and Prof. Mayura Nagar focused on identifying slow, average and fast learners using Naive Bayes, ZeroR, J48 and Random Tree [4]. Data set of MCA students are used in WEKA tool. Random Tree technique achieves 95% accuracy.

Senthil Kumar Thangavel, Divya Bharathi P and Abijith Sankar developed generalized framework to predict student placement [5]. It predicts the students to have one of the dream company, core company, mass recruiters. Data Meter and WEKA tool comparison is done over this. Accuracy of 71% is obtained for 289 instances. In order to predict right class for students from science, social and literature class knowledge discovery data mining is used [6]. Different classification algorithms like J48, SimpleCart, Kstar, SMO, NaiveBayes and OneR are used in testing the student data set. Out of 6 algorithms, J48 provides good results of 79.61% accuracy. Sagardeep Roy and Anchal Garg research aims at identifying students who are likely to drop out, weak, neads

improvement, good in academics but lately deterioated [7]. Classification algorithms like Naive Bayes, MLP and J48 are used. J48 algorithm obtaines a better result of 73% accuracy which is higher compared with other algorithms.

V. Shanmugarajeshwari and R. Lawrance used C5.0 algorithm is used to predict the student's performance [8]. It acts like a warning system to improve the students performance. The dataset includes marks in higher secondary, previous semester and performance in last semester that is Pass or Reappear. The C5.0 algorithm have accuracy of 100% which is then comapred with Naive and Decision Tree algorithms.

Ahmad Afif Supianto et.al tested student data set against Random Tree, REP Tree, and C4.5 decision tree algorithms [9]. C4.5 obtains accuracy of 77% where as Random Tree has 74% and REP Tree has 76%. Entry method and Gender could be removed as it shows no negative influence on the accuracy of the method. Student performance is predicted using hybrid classification technique which give accuracy of 62% [10]. ID3 and J48 are used on the dataset voting technique is applied. Out of 500 data set, 300 dataset is used for predicting the student's performance.

3. METHODOLOGY

3.1 System Design

Fig -1: System Design

Data is collected in the first phase as depicted in Fig. 1. From the raw data required attributes are selected and classification algorithm is used to construct the prediction model. Out of several classifiers, proposed work make use of SVM, J48, Random Forest, Naive Bayes algorithms. The results of the different classification algorithm are compared.

3.1.1 Data Collection

The data set used for prediction is obtained from Kaggle. WEKA tool is used for testing the data using different classification algorithms. The attribute description is shown in Table 1 where target attribute is used as a class predicting Pass or Fail values from the data set.

Table -1: Data set attributes Description

Sl. No.	Attributes	Description
1	Gender	Nominal Value {male,female}
2	General_Science	Score in General_Science
3	Maths	Score in Maths
4	Computer	Score in Computer
5	Result	Target attribute {Pass, Fail}

3.1.2 Classification Algorithm

Classification can be performed on types of data such as structured and unstructured. Categorizing the data into the given classes is called Classification. The proposed model classifies the data into two categories Pass and Fail. For this study, out of several classifiers proposed work make use of SVM, J48, Random Forest and Naive Bayes which is best suited. SVM can be used to find a hyperplane which clearly classifies data into given classes. There may be multiple possible hyperplanes which can be chosen to separate the two classes of data points. J48 classifier is a simple classifier which creates a binary tree.

Bayes Theorem is based on Naive Bayes algorithm. The presence of a feature is not related to the presence of another feature. This model can be easily built and it is useful for larger data set. Random forest algorithm comprises of a large number of individual decision trees. It is much better than single decision tree. The algorithm first selects the random samples and constructs decision trees for every sample. Prediction result from every sample is obtained. Finally it will select the best most voted prediction. Using ID3 algorithm J48 classification algorithm is developed. J48 algorithm was developed in order to deal with missing data, continuous data, pruning, splitting and generating rules. For splitting purpose the method uses Gain Ratio instead of Information Gain.

3.1.3 Performance Metrics

Accuracy

In order to measure the algorithm correctness of different classifiers, accuracy can be used. It is ratio of sum of exact prediction to the total number of instances that has to be predicted.

$$\text{Accuracy} = \frac{(\text{Tpositives} + \text{Tnegatives})}{(\text{Tpositives} + \text{Tnegatives} + \text{Fpositives} + \text{Fnegatives})} \quad (1)$$

Kappa Statistics

In comparison with any random model, Kappa statistics value would determine the accuracy of the model.

$$\text{Kappa} = \frac{K(OA) - K(EA)}{(1 - K(EA))} \quad (2)$$

Where K (OA) is Observed accuracy, K (EA) is Expected accuracy

MAE (Mean Absolute Error)

It is a comparison between predicted and observed values which can be calculated by the formula.

$$MAE = 1/n \sum_{i=1}^{n} [|fi - yi|] \quad (3)$$

where f_i = prediction, y_i = true value

Root Mean Squared Error (RMSE)

It is the average amount of error which occurs when the dataset is tested. It is calculated by the formula

$$MAE = \sqrt{(1/n) \sum_{i=1}^{n} ((P(i,j) - Tj)/Tj)^2} \quad (4)$$

where P_i = predicted output, i= fit instance= target value of fit instance

Receiver Operating Characteristic(ROC) Area

ROC is used in order to represent the performance in a graphical manner. It can be used to plot the graph by showing true positive rate of different classifiers against the false positive rate.

2. EXPERIMENTAL RESULT ANALYSIS

The experiment was conducted by using 1000 instances. To predict the student result, the classification algorithms like Naive Bayes, SVM, Random Tree and J48 are used. It is not so easy to keep different set of data for training and testing. If the training data is reduced then there may increase in risk of losing important patterns. This may introduce error. As the k-fold validation is divided into subsets, these subsets are used for training as well as testing which gives better result than other testing options. By considering result as the target attribute 10 folds cross validation test was done on data set.

Table -2: Classification results obtained from SVM, Naive Bayes, J48, Random Tree classifiers

Algorithms	Accuracy	ROC	Kappa statistic	RMSE	MAE	Time (s)
Naive Bayes	96.5%	0.815	0.7637	0.151	0.0376	0
SVM	96.8%	0.669	0.6591	0.1766	0.0312	0.01
J48	99.6%	0.822	0.9735	0.0555	0.0032	0
Random Tree	99.6%	0.821	0.9735	0.0558	0.0031	0

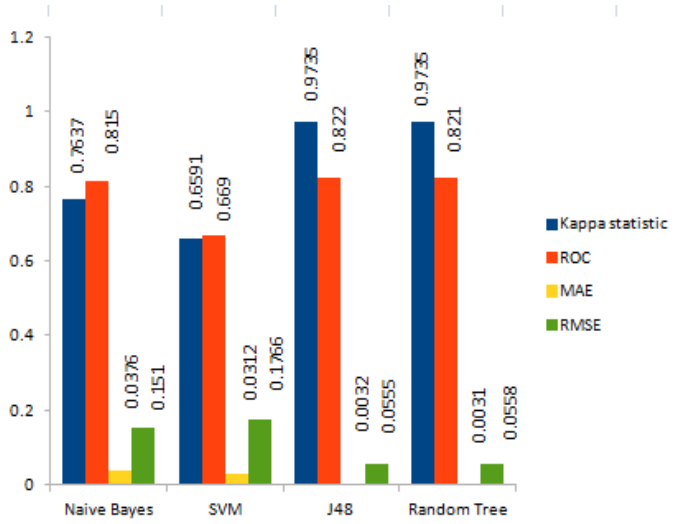

Chart -1: Comparison of Performance Metrics

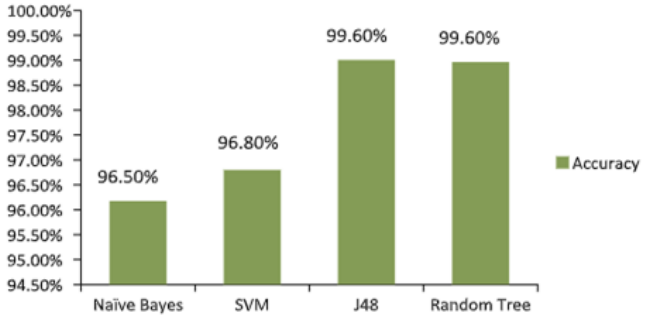

Chart -2: Comparison of Accuracy

SVM comparatively has good accuracy but time consumed is more compared to other classification algorithms. Kappa statistics result is comparatively high in J48 and Random tree algorithms. It is noticed that ROC values for J48, Naive Bayes and Random tree have corresponding values. Also Mean absolute error is very lowest in Random Tree. Root mean square error must be low for a good classifier. Here we observe that J48 and Random tree has low values for RMSE. From Chart 2 and 3, it is evident that Random Tree and J48 have the higher accuracy.

Table -3: Comparison of Correct and Incorrect classification

Algorithms	Correctly Classified	Incorectly classified	TP Rate	FP Rate
Naive Bayes	929	33	0.996	0.018
SVM	932	30	0.969	0.453
J48	959	3	0.997	0.016
Random Tree	959	3	0.997	0.016

From the Table 3, we can observe that J48 and Random Tree have the higher accuracy. Where the correct prediction is 959 out of 1000 instances for J48 and Random Tree. Incorrectly classified instances are three for both J48 and Random forest.

3. CONCLUSIONS

A prediction model to predict the student result is implemented using weka tool. Classification algorithms like SVM, Naive Bayes, Random Tree and J48 were used in the experiment. Using various performance metrics performance of these algorithms were analyzed. After analyzing the results it was found that J48 and Random tree have highest accuracy of 99.6% compared with other algorithms. It was also observed that next best accuracy algorithms are SVM and Naive Bayes but time consumed is more compared to other classification algorithms. In future more student attributes can be tested and analyzed against the feature selection algorithms that can be used to reduce the features.

REFERENCES

[1] Azwa Abdul Aziz, Nur Hafieza Ismail, Fadhilah Ahmad, Mining Students' Academic Performance. Journal of Theoretical and Applied Information Technology ,2013.

[2] Rahul Patil et.al., Prediction system for student performance using Data Mining Classification. IEEE , 2018.

[3] Ching-Chieh Kiu, Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities. IEEE , 2018.

[4] Vrushali Mhetre, Prof. Mayura Nagar, Classification based data mining algorithms to predict slow, average and fast learners in educational system using Weka. Proceedings of the IEEE. International Conference on Computing Methodologies and Communication , 2017.

[5] Senthil Kumar Thangavel et.al., Student Placement Analyzer: A Recommendation System Using Machine Learning. International Conference on Advanced Computing and Communication System Coimbatore 2017.

[6] Oktariani Nurul Pratiwi, Predicting Student Placement Class using Data Mining. IEEE International Conference on Teaching, Assessment and Learning for Engineering, pp 26-29 , 2013.

[7] Sagardeep Roy, Anchal Garg, Predicting Academic Performance of Student Using Classification Techniques. 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics. GLA University, Mathura, 2017.

[8] V. Shanmugarajeshwari, R. Lawrance, Analysis of Students' Performance Evaluation using Classification Techniques. IEEE, 2016.

[9] Ahmad Afif Supianto et.al, Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University. IEEE, 2018.

[10] A.Dinesh Kumar et.al, Prediction of Student Performance using Hybrid Classification. International Journal of Recent Technology and Engineering. ISSN: 2277-3878, Vol 8 Issue 4, 2019.

[11] K.Manikandan , S.Sivakumar, M.Ashokvel, A Classification Model for Predicting Campus Placement performance Class using Data Mining Technique. International Journal of advance research in science and Engineering . Vol 7, Special Issue 06, 2018.

[12] C. M. Vera, C. R. Morales and S. V. Soto, Predicting School Failure and Dropout by Using Data Mining Techniques. IEEE Journal of Latin-American Learning Technologie, Vol. 8, No. 1 , 2013.

[13] Aaditya Desai, Sunil Rai, Analysis of Machine Learning Algorithms using Weka. International Journal of Computer Applications , 2013.

BIOGRAPHIES

Ms. Prathyakshini is an Assistant Professor in Information Science and Engineering department at NMAM Institute of Technology, Nitte, Karkala, Karnataka, India. She has 3 years of Teaching experience. Her research interests include Machine Learning and Image processing.

Ms. Pratheeksha Hegde N is an Assistant Professor in Information Science and Engineering department at NMAM Institute of Technology, Nitte, Karkala, Karnataka, India. She has 4 years of Teaching experience. Her research interests include Machine Learning and Image processing.

Ms. Nikitha Saurabh is currently working as an Assistant Professor in the Dept. of Information Science and Engineering at NMAM Institute of Technology, Nitte, Karkala, Karnataka, India. She has 10 years of teaching and 1 year of industry experience. Her research interests include Machine Learning, Deep Learning and Data Science.