Leadership, Education, Personality: An Interdisciplinary Journal ISSN: 2524-6178
www.sibe.rpress.co.in, Vol. 18, Issue 06 June 2023, UGC CARE 1

Predicting Students' Academic Performance in School Education Using Classification Techniques

Rajinder Singh
Research Scholar, Department of Computer Science and Engineering
OM Sterling Global University Hisar
Email – rajindercse191@osgu.ac.in

Dr. Rajinder Singh Sodhi
Associate Professor & HOD, School of Engineering & Technology
OM Sterling Global University Hisar

Abstract

In today's scenario, data mining finds application across various domains, serving to derive insightful analyses. One the main domain in data mining is education to analyze the students' performance and their results. Through data mining, educational institutions can assess student achievements and make policies to enhance them further. Often referred to as Educational Data Mining (EDM), this practice focuses on predicting student performance and devising measures to boost academic success. The main aim of this study was the prediction of students' learning habits and steps that could be taken to improve students' performance. In this study, Four (4) classification algorithms were used, namely J48, Random Forest, Naïve Bayes, and REPTree.

Keywords - Data Mining, Educational Data mining, Accuracy, Naïve Bayes, J48

Introduction

In the present scenario, data mining is a very important field of research in educational institutes. Educational Data Mining (EDM) is a field that applies data mining techniques and methodologies to educational data in order to discover patterns, insights, and trends that can inform and improve educational practices. It involves the collection, analysis, and interpretation of data from various educational sources such as learning management systems, student assessments, and educational software. The main aim of data mining is to predict performance of poor students' that might lead to dropout [1]. There are many factors that influence the performance of students in an institution such as socio-economic factors, non-academic factors and academic factors. Many classification and prediction algorithms of data mining can be used to analyse students' data [2]. Educational data mining affects various areas of the education industry and is undoubtedly very helpful in analyzing data, predicting student performance, grouping and classifying students, and planning and student scheduling

[3]. Several base classifiers can be used in data mining process including j48, KNN, random tree, and naïve bayes. We can also use popular ensemble method viz. Boosting on academic dataset to predict the performance of students [4]. Students' data can be collected from a higher education institution according to various attributes. In their study, various classification, clustering, and different algorithms such as J48 Tree, AD Tree, and Random Tree are applied on data after preprocessing the data. WEKA tool is used to analyze the grouped data with related similarities [5]. More than 1000 records were used in study and this dataset was created by collected from a preparatory male school in Gaza strip. Results obtained from Decision Tree Algorithm are the best. The academic features such as previous year results have more impacts on the students' performance as compare to other social attributes [6]. In some research work, the author examined students' data set with different attributes values using two algorithms which were classification and clustering algorithms. The C4.5 algorithm with the average highest accuracy of 62.7% as compared to other classification algorithms [7]. Some research provides a specialization selection recommendation for students of engineering course through the application of various data mining algorithms. Various important attributes in making predictions were determined by using a selection of characteristics based on the correlation [8]. In some study, the author examines various EDM tools and techniques involved in mining educational data. The author suggests the best tools and techniques from these tools with detail for real-world usage [9].

Related Work

Mubarak Albarka Umar (2019) author examine the problem of Katsina State Institute of Technology and Management (KSITM) like students' dropout and delay in graduation and the various reasons behind these issues. First year students' performance is one of the major reasons behind these problems. The main aim of this study is to predict performance of poor students' that might lead to dropout and thus allow the institution to develop strategic policies that will help those poor students to improve their performance and enable them to graduate in time. A neural network model if proposed by study which is capable of predicting student's performance using attributes like students' personal and academic information, and place of residence etc. Around 61 students' record was used to train and test the model in WEKA software tool and the accuracy of that model was measured using best evaluation criteria. 73.68% of students' performance are correctly predicts by their model.

Pavithra & S. Dhanaraj (2018) Educational data mining (EDM) is an emerging research field in the development of educational psychology and learning sciences. Records of

academic performance play an important role in the development of the institution. Although there are many factors that influence the performance of students in an institution such as socio-economic factors, non-academic factors and academic factors. There are many classification and prediction algorithms in data mining. This article focuses on student performance. The dataset used for this study is primary data from a private institution called "Sree Saraswathy Thyagaraja College" in Pollachi, which is located in a rural area. This research work was conducted using five different algorithms including Naive Bayes, Decision Tree, Representation Tree, J48 and Multi-Layer Perception (MLP). In this study, the authors also highlighted various factors that influence student performance.

Vandna Dahiya (2018) conducted a survey on the various components of educational data mining and their objectives. Educational data mining affects many areas of the education industry and is undoubtedly very helpful in visualizing facts, predicting student performance, grouping and classifying students, predicting profiles, and planning and student scheduling. First and foremost, the authors say, education data from disparate sources is inherently evolving. Secondly, due to the large amount of data, data storage and maintenance also become difficult. Another issue is the order and arrangement of this dynamic educational data and how to understand it. The author also defines some data mining tools such as WEKA, KEEL, R (Revolution), KNIME, and ORANGE.

Mudasir Ashraf et al. (2018) says that the ensemble method approach has produce more accurate results in classifying the correct instances when compared with individual other learning algorithms. To predict the performance of students, author employed used popular ensemble method viz. Boosting on academic dataset. Several other base classifiers are used in this process including j48, KNN, random tree, and naïve bayes. Among this classification method, j48 demonstrated excellent performance of 95.32%. After using filtering procedures of SMOTE and spread sub sampling, j48 show accuracy of 96.44% and naïve bayes exhibited 95.85% accuracy. Also author conclude that oversampling approach produce better results in predicting the outcome of students when compared with other techniques employed.

K. Kiruthika & S. Sivakumar (2017) used educational data mining methods to analyze the students' learning behavior. They collected students' data from a higher education institution according to various attributes. In their study, they applied various classification, clustering, and different algorithms such as J48 Tree, AD Tree, and Random Tree after preprocessing the data. WEKA tool is used to analyze the grouped data with related similarities. They extracted

knowledge that describes in detail students' learning behavior with two learning methods (1) traditional learning (2) virtual learning.

Hafez Mousa & Ashraf Yunis Maghari (2017) this paper proposes a students' performance prediction model which is based on various Data Mining classification algorithms like Naïve Bayes, Decision Tree and K-NN. Around 1100 records were used in this study and this dataset was created by collected from a preparatory male school in Gaza strip. Results obtained from Decision Tree Algorithm are the best. The academic features such as previous year results have more impacts on the students' performance as compare to other attributes like social case which has little impact or no impact on the students' performance. These results are very helpful in improving students' performance and to minimize students' failure. The data set which is used in this paper has been prepared by collected students' records from preparatory male school for 2014/2015 session year. Author use three classifiers Naïve Bayes, Decision Tree and K-NN) in his study and their results shows that DT classifier gives the best output. Results obtained in this study may help the educators to minimize the failure ratio in students, by determining students that may fail.

Karthikeyan Govindasamy & Velmurugan Thambusamy (2017) In this research work, the author examined two algorithms which were classification and clustering algorithms using students' data set with different attributes values. The C4.5 algorithm with the average highest accuracy of 62.7% outperforms all other classification algorithms. The performance of clustering algorithms is better in prediction of student performance than the classification algorithms.

Rosemarie M. Bautista et al. (2016) says that Educational Data Mining (EDM) can be used to extract patterns which are very useful in student academic performance. The main objective of this research is to provide a specialization selection recommendation to students of engineering course through the application of various data mining algorithms. Different attributes those are important in making predictions were determined by using a selection of characteristics based on the correlation. The comparative analysis between the known algorithms is analyzed and the highest precision has been taken into account. In this study the author found that decision tree classification model using WEKA and J48 produced a precision value of 80.06. The study found that various factors like gender, algebra, calculus, and physics courses have a significant effect on the prediction of engineering specialization.

A. S. Arunachalam & T. Velmurugan (2016) says that educational data mining (EDM) creates a high impact in the field of an academic domain. EDM collect data, explores, analyzes, and gives ideas in understanding behavioral patterns of students and to choose the

right path for their carrier. In this study, the author focuses on various techniques involved in mining educational data. Also, they discuss various EDM tools and techniques in this article. Among these different tools and techniques, the author suggests the best tools and techniques with detail for real-world usage. In conclusion, the author says that most of the classification algorithms perform in a better way of analyzing and describe the current trends of EDM.

Proposed Work

The dataset used in this study was combined of around 300 records collected from various educational institutions. The dataset includes 14 variables for analysis: student's gender, age, father's qualification, father's occupation, mother's qualification, mother's occupation, family income, percentage obtained in 10th grade, percentage obtained in 12th grade, 10th_Medium, 10th_school_type, 12th_Medium, 12th_school_type, 12th_school. To make the variables suitable for analysis, we converted them into categorical attributes by discretizing the numerical attributes. For instance, let's take variable X (X = x0, x1, x2, ...), which represents the passing percentages of students in the 10th grade, 12th grade, and other related factors. We categorized the final grades into three groups: Low-Risk, Medium-Risk, and High-Risk. The table below shows this categorization:

Final Result	Final Grade
X ≥ 80%	Low-Risk
X ≥ 60% and X<80	Medium-Risk
X< 60%	High-Risk

Table-1: Values of final grade

Other different attributes are also descretized, such as the student's present course or stream, the passing percentages for the 10th, and 12th. Finally, the following table lists the most specific attributes:

Attribute	Description	Possible Values
Student ID	Student's Unique Identification	{alphabets Characters}
Gender	Gender of Student	{Male, Female, Other}
Age	Students Age	{Below 16, 16 to 18, Above 18}
Father Qualification	Qualification of father	{10th, 12th, graduation, post graduation, not educated}

Father Occupation	Occupation of Father	{Agriculture, Business, Govt. Service, Labour, Private Service}
Mother Qualification	Qualification of Mother	{10th, 12th, graduation, post graduation, not educated}
Mother Occupation	Occupation of Mother	{Govt. Service, House Wife, Private Service}
Family Income	Income of family	{Under 2 lac, 2 to 4 lac, more than 4 lac}
High School Percentage (10th %)	Percentage of marks obtained in 10th class exam.	{ Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%}
Intermediate Percentage (12th %)	Percentage of marks obtained in 12th class exam.	{ Below 60%, 60% to 70%, 70% to 80%, 80% to 90%, Above 90%}
10th_Medium	Medium of 10^{th} class school	English Hindi
10th_school_type	10^{th} class School type	Private Govt.
12th_Medium	Medium of school education in 12^{th} class	English Hindi
12th_school_type	12^{th} class School type	Private Govt.
12th_school	12^{th} class School type	Co-Education Only for Girls
Result	Final Result obtained after analysis the passing percentage of 10th ,12th and other factors	{High-Risk, Medium-Risk, Low-Risk}

Table - 2: Attributes and their possible values

Weka Tool

A variety of visualization tools, algorithms, and graphical user interfaces for quick access to these capabilities are included in the open source software Weka. It is fully implemented in JAVA language that makes it portable and platform neutral. It can run on almost all computing platform. Major activities of data mining like data preprocessing, classification, clustering, association, visualization, and feature selection are can easily done in WEKA. The WEKA graphical environment has six buttons: Simple CLI, Explorer, Experimenter, Knowledge Flow, ARFF-Viewer, & Log.

The Explorer interface features a number of panels that provide access to the Weka software essential elements.

- The Preprocess panel works to import data from databases, ARFF, CSV files etc. and all the preprocesses functions are done on this panel like using filtering method that can change the data's format, turning numerical attributes into discrete ones and other task related to data. Also on this preprocess screen, it is also possible to update or delete records and attributes in accordance with particular criteria. We can also able to view graph for a specific attribute.
- We can use Most of classification and regression techniques (such as the NaiveBays algorithm, ID3 Tree, ADTree, J48 Tree, and ZeroR rules, among others) to the dataset. Model accuracy can be estimated using the Classify panel. Additionally, we can also view incorrect predictions, ROC curves, etc. Classification results can be seen in the classifier output area.
- Most of the clustering method such as the simple k-means algorithm, DBScan, EM, and XMeans algorithm can be use from the Cluster panel. Omit attribute button in clustering process can be used to omit some specific attributes according to our requirement.
- The associate panel is provided in Weka software to access the different algorithm, such as the Apriori and Predictive Apriori algorithms. When the proper parameter for the association rule has been selected, we can view the result and also we can save the result set.
- We can use the Select Attributes panel to explore all possible combinations of attributes within the dataset and to find the optimal attribute subset for accurate predictions.
- 2D plots of the present relation can be used to visualized using the visualize panel.

Result and Discussion

The data set used in this study contains around 300 students records obtained from the various schools whether they are in rural or urban area and are of Govt. and private school type. All the schools are either in Co-Educational or in girls only category. To study this data set we are using various classification methods in Weka tool.

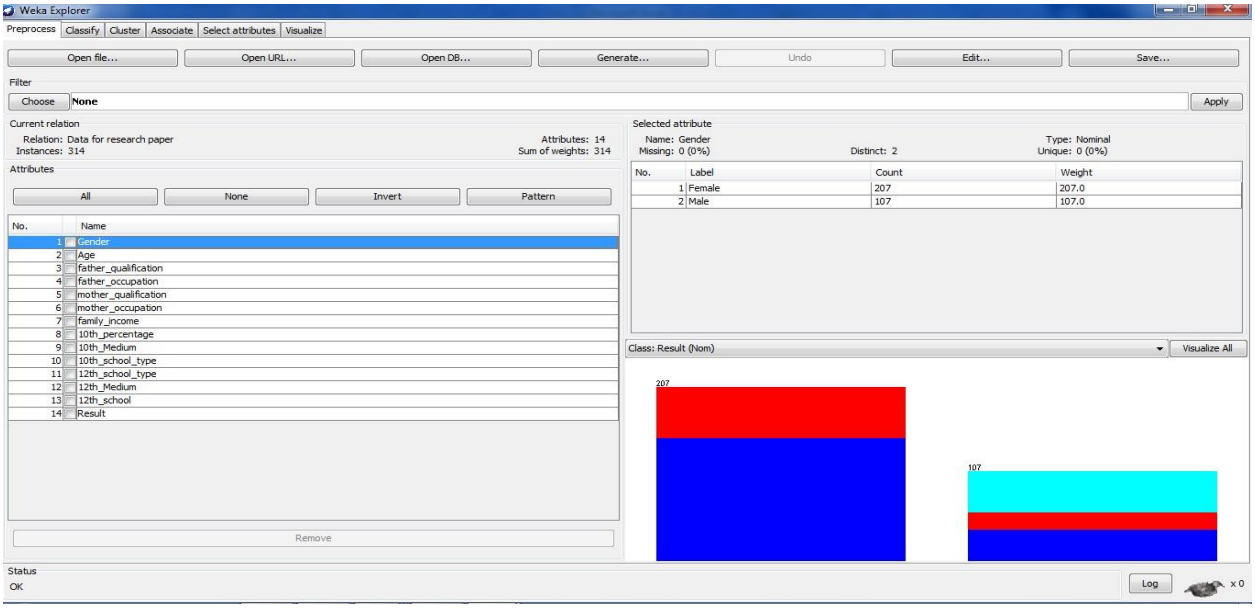

Figure 1: Weka 3.9.6 with explorer window opens with student's database

This main window is used to show all the variables and their related information. In the left panel this window shows the variable name where as right panel shows count and weight of selected variable. We can remove unwanted variable by unselecting that variable from the left panel if that variable is of no use in our study.

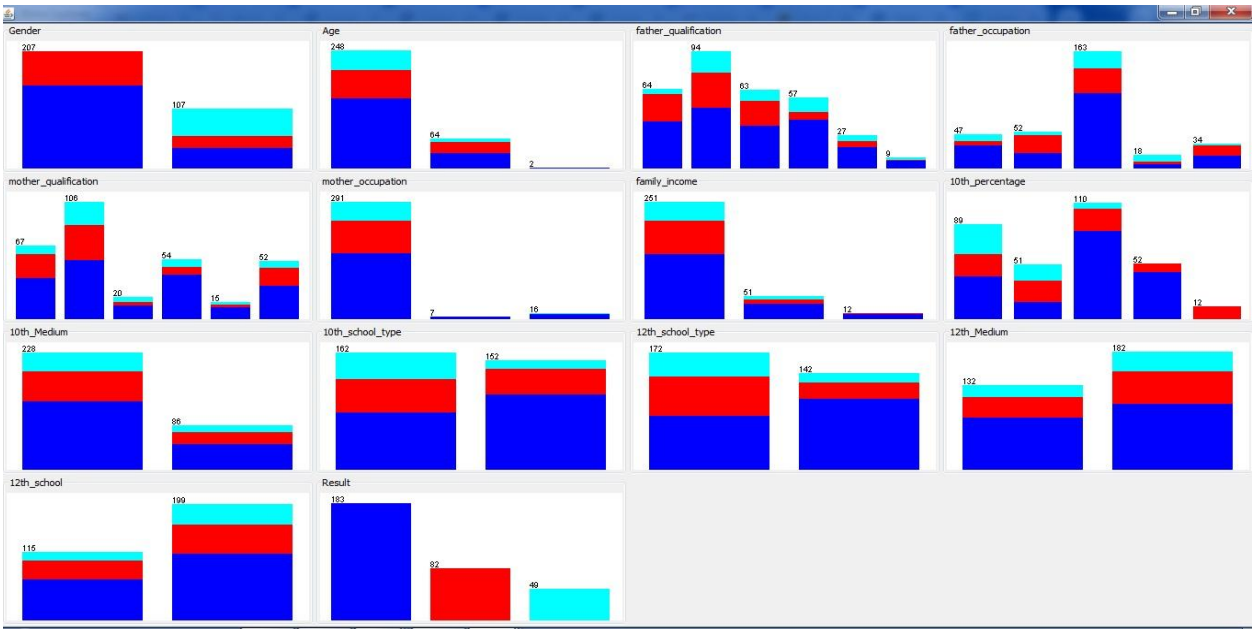

Figure 2: Visualized Result of Student's Dataset from Weka

This window shows the visualization of all variables with their respective count. We can also check variable count in separate window for each variable.

In Weka, the visualization window is an important tool for exploring data as well as evaluating the results of machine learning models. After loading dataset in the Explorer interface, when we navigate to the "Visualize" tab to see a scatter plot matrix that showcases the relationships between different pairs of attributes. We can also interactively select attributes for the X and Y axes to focus on specific relationships, and if dataset includes a class attribute, different colors can represent different classes. Also, Weka offers features like histograms for individual attributes, advanced 2D and 3D plotting, and interactive tools for more detailed analysis. For model results, Weka gives features to visualize classifier errors and cluster distributions. These visualizations as output can be saved as image files for reporting and further analysis.

Using J48

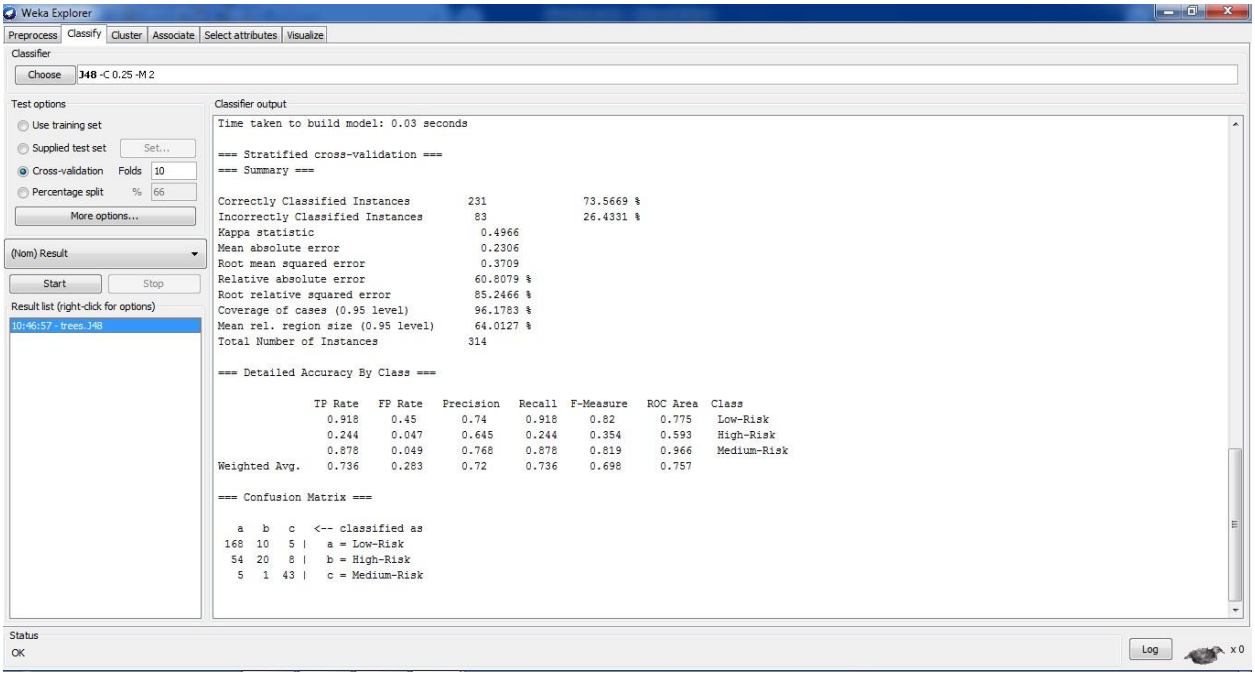

Figure 3: J48 Classifier output window

This algorithm is an extended version of ID3. Accounting of missing values, decision trees pruning and other various features are added in J48. J48 builds a decision tree by recursively splitting the dataset into subsets based on the attribute that provides the highest information

gain, a measure of how well an attribute separates the data according to their target classifications. At each node, the algorithm selects the attribute that best reduces the entropy or disorder of the dataset, and this process continues until a stopping condition is met, such as when all instances in a subset belong to the same class or when no more attributes are left to split on.

This algorithm is used in Weka to analyze data of around 300 students. The values within the confusion matrix denote the number of instances that belong to a particular actual class and were classified into a specific predicted class. From Confusion matrix of J48, instances in Low-Risk category were classified as Low-Risk correctly 168 times (true positives), but there were 10 instances misclassified as High-Risk and 5 instances misclassified as Medium-Risk. Similarly, High-Risk instances were correctly classified 20 times, but 54 were mistakenly classified as Low-Risk and 8 as Medium-Risk and Medium-Risk instances were correctly classified 43 times, but there were 5 misclassifications as Low-Risk and 1 as High-Risk.

Using Random Forest

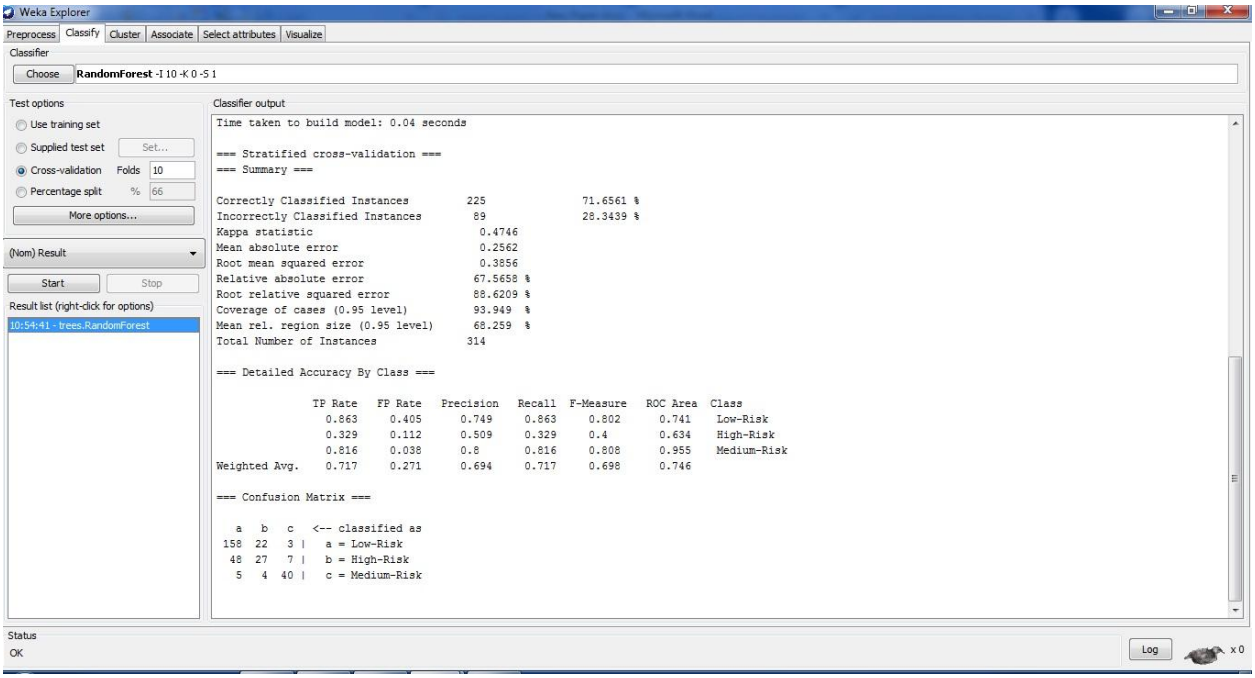

Figure 4: Random Forest Classifier output window

This supervised classification algorithm is both flexible and user-friendly. As its name shows forest-like structure composed of numerous trees. It delivers robust results with increased accuracy as more trees are added to the forest. Essentially, this algorithm constructs multiple decision trees that can be amalgamated to enhance stability and precision in predictions. Each

tree in a random forest is built from a different bootstrap sample of the training data, and at each split in the tree, a random subset of features is considered, which helps to ensure diversity among the trees. This randomness helps to reduce the variance of the model, making Random Forests more robust and less likely to overfit compared to a single decision tree.

From above matrix, instances labeled as Low-Risk were correctly classified as such 158 times (true positives), but there were 22 instances misclassified as High-Risk and 3 instances misclassified as Medium-Risk, Similarly, instances labeled as High-Risk were correctly classified 27 times, but 48 instances were mistakenly classified as Low-Risk and 7 as Medium-Risk and instances labeled as Medium-Risk were correctly classified 40 times, with 5 misclassifications as Low-Risk and 4 as High-Risk.

Using Naïve Bayes

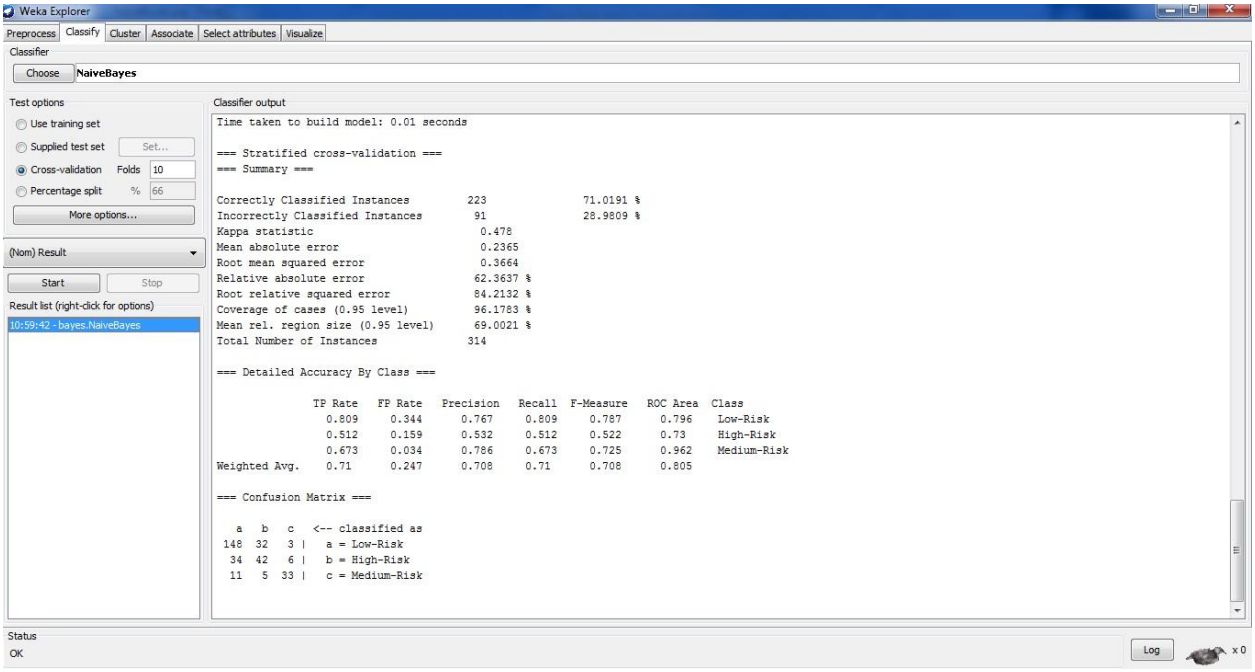

Figure 5: Naïve Bayes Classifier output window

The algorithm calculates the probability of a given instance belonging to each class based on the feature values observed in the training data. It does this by first estimating the probability distributions of each feature within each class. Then, when presented with a new instance to classify, Naive Bayes calculates the likelihood of that instance belonging to each class based on the observed feature values and combines this with the prior probability of each class to

compute the posterior probability using Bayes' theorem. The class with the highest posterior probability is then assigned as the predicted class for the instance.

Simplicity and efficiency are the key advantages of Naive Bayes, making it particularly suitable for large datasets. Additionally, it can handle missing values and categorical features effectively. However, its assumption of feature independence may not hold true in all datasets, which can lead to suboptimal performance in some cases. Nonetheless, Naive Bayes remains a popular and effective choice for classification tasks, especially in scenarios where computational resources are limited or where real-time predictions are required.

For example, instances labeled as Low-Risk were correctly classified as such 148 times (true positives), while 32 instances were misclassified as High-Risk and 3 instances as Medium-Risk. Similarly, instances labeled as High-Risk were correctly classified 42 times, but 34 instances were mistakenly classified as Low-Risk and 6 as Medium-Risk. Likewise, instances labeled as Medium-Risk were correctly classified 33 times, with 11 misclassifications as Low-Risk and 5 as High-Risk.

Using REPTree

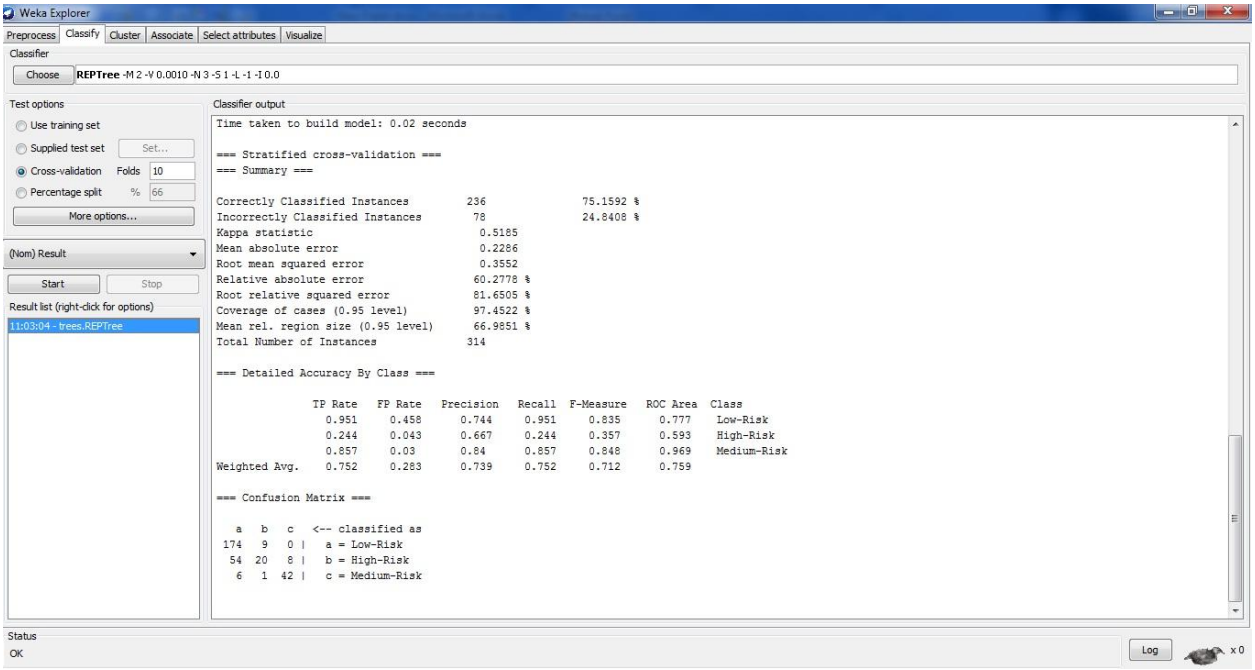

Figure 6: REPTree Classifier output window

The REPTree (Repeated Incremental Pruning to Produce Error Reduction) algorithm is a decision tree-based classification algorithm. It constructs a decision tree recursively by splitting the dataset based on the attribute that maximizes the reduction in error. Unlike

traditional decision tree algorithms, REPTree employs a pruning strategy that aims to reduce overfitting by iteratively removing nodes from the tree if doing so improves its predictive accuracy on a separate validation set.

From the above data we see that out of the instances known to be Low-Risk, 174 were correctly classified as Low-Risk (true positives), while 9 were incorrectly classified as High-Risk (false negatives), and none were misclassified as Medium-Risk. Similarly, out of the instances known to be High-Risk, 20 were correctly classified as High-Risk (true positives), while 54 were misclassified as Low-Risk (false positives), and 8 were misclassified as Medium-Risk. Additionally, out of the instances known to be Medium-Risk, 42 were correctly classified as Medium-Risk (true positives), while only 6 were misclassified as Low-Risk and 1 as High-Risk. This data shows that REPTree algorithm works better for medium-risk category as compared to high-risk and low-risk.

Results and Discussion

	Classifier			
Criteria	J48	Random Forest	Naïve Bayes	REPTree
Correctly Classified Instances	231	225	223	236
Incorrectly Classified Instances	83	89	91	78
Precision	0.72	0.694	0.708	0.739
Recall	0.736	0.717	0.71	0.752
F-Measure	0.698	0.698	0.708	0.712

Table – 3: Comparison of results of all classifier

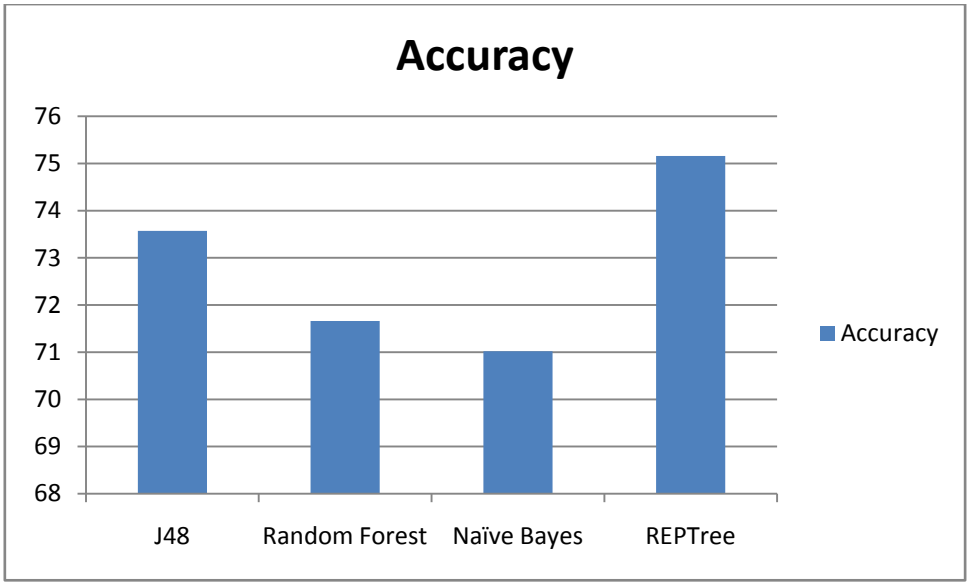

Figure 7: Accuracy Graph for all Classifier

In the evaluation of classification models in WEKA, the detailed accuracy by class provides insights into the performance of the classifier across individual classes. The "Correctly Classified Instances" metric denotes the number of instances accurately classified by the model, while "Incorrectly Classified Instances" indicates the number of which are not correctly classified by the model. Precision, Recall, and F-Measure are performance metrics that further quantify the classifier's effectiveness. Precision represents the proportion of correctly predicted instances for a particular class among all instances classified as that class. In this context, for instance, J48 achieved a precision of 0.72, indicating that 72% of instances classified as belonging to a certain class were correctly classified. Recall, also known as sensitivity, measures the proportion of actual instances of a class that were correctly predicted by the classifier. For example, in the case of Random Forest, a recall value of 0.717 signifies that 71.7% of instances belonging to a particular class were accurately identified by the model.

F-Measure is the harmonic mean of precision and recall, providing a balanced measure of a classifier's performance. A higher F-Measure indicates better overall performance. For REPTree, the F-Measure of 0.712 suggests a good balance between precision and recall. These detailed accuracy metrics allow for evaluation of the classifier's performance across different classes, providing valuable insights into its strengths and weaknesses. They help in identifying which classes the classifier excels in classifying accurately and which may require further tuning or improvement.

Conclusion

In education field, data mining plays a crucial role in conducting predictive analyses on student datasets. Through data mining techniques, valuable insights such as predicting student performance and achievements can be derived, guiding subsequent actions for those students which are at risk.

In this study, contains records of around 300 students from various schools situated in Haryana. Around 14 attributed are taken to collect the data about the students. For analysis, four classification algorithms are used: J48, Random Forest, Naïve Bayes and REPTree. Each algorithm possesses distinct features. Consequently, all algorithm shows good accuracy levels (>65%).

For future study, dataset size can be increase and exploring other data mining methodologies are recommended to get some more valuable information about the students' performance.

References:

1. Umar, M. A. (**2019**), "Student Academic Performance Prediction using Artificial Neural Networks: A Case Study**",** *International Journal of Computer Applications*
2. Pavithra, A.; Dhanaraj, S. (**2018**), "Prediction Accuracy on Academic Performance of Students Using Different Data Mining Algorithms with Influencing Factors", *International Journal of Scientific Research in Computer Science Applications and Management Studies*
3. Dahiya, V. (**2018**),"A Survey on Educational Data Mining", *International Journal of Research in Humanities, Arts and Literature (IMPACT: IJRHAL)*, ISSN (P): 2347-4564; ISSN (E): 2321-8878 Vol. 6, Issue 5, p 23-30.
4. Ashraf, M.; Zaman, M.; Ahmed, M. (**2018**), "Using Predictive Modeling System and Ensemble Method to Ameliorate Classification Accuracy in EDM", *Asian Journal of Computer Science and Technology*, ISSN: 2249-0701 Vol.7, No.2, pp.44-47
5. Kiruthika, K.; Sivakumar, S. (**2017**), " Analysis of Students' Behaviour And Learning Using Classification Of Data Mining Methods", *International Journal of Computational and Applied Mathematics*, ISSN 1819-4966, Volume 12, Number 1.
6. Mousa, H.; Maghari, A. (**2017**), "School Students' Performance Predication Using Data Mining Classification", *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 6, Issue 8.
7. Govindasamy, K.; Thambusamy, V. (**2017**), "A Study on Classification and Clustering Data Mining Algorithms based on Students Academic Performance

Prediction", *International Journal of Control Theory and Applications,* Volume 10, Number 23.

8. Bautista, R.M.; Dumlao, M.; Ballera, M.A. (**2016**), "Recommendation System for Engineering Students' Specialization Selection Using Predictive Modeling", Third International Conference on Computer Science, Computer Engineering, and Social Media (CSCESM2016), Thessaloniki, Greece.
9. Arunachalam, A.S.; Velmurugan, T. (**2016**), "A Survey on Educational Data Mining Techniques", *International Journal of Data Mining Techniques and Applications,* Volume: 05, Page No.167-171 ISSN: 2278-2419.
10. Algur, S.P.; Bhat, P.; Ayachit, N. (**2016**), "Educational Data Mining: RT and RF Classification Models for Higher Education Professional Courses", *I.J. Information Engineering and Electronic Business*, 2, 59-65.
11. Sivakumar, S.; Venkataraman, S.; Selvaraj, R. (**2016**), "Predictive modeling of student dropout indicators in educational data mining using improved decision tree", *Indian Journal of Science and Technology*, Vol 9, Issue 4, pp 1–5.
12. Alharbi, Z.; Cornford, J.; Dolder, L.; Iglesia, B.D.L (**2016**) "Using data mining techniques to predict students at risk of poor performance", Proceedings of Science and Information Organization Computing Conference, IEEE, pp. 523-531.
13. Kumar, M.; Shambhu, S.; Aggarwal, P. (**2016**), "Recognition of slow learners using classification data mining techniques", *Imperial Journal of Interdisciplinary Research*, Vol 2, Issue. 12.
14. Al-Barrak, M.A.; Al-Razgan, M. (**2016**), "Predicting Students Final GPA Using Decision Trees: A Case Study", *International Journal of Information and Education Technology*, Vol. 6, No. 7.
15. Devasenapathy, K.; Duraisamy, S. (**2016**), "Performance analysis of teaching assistant using decision tree classification algorithm", *Asian Journal of Information Technology*, Vol 15, No. 19, pp 3820-3825, ISSN: 1682-3915.
16. Parneet, K.; Manpreet, S.; Gurpreet, S.J. (**2015**), "Classification and prediction based data mining algorithms to predict slow learners in education sector", *3rd International Conference on Recent Trends in Computing (ICRTC).*