International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

CRUNCHING THE NUMBERS: COMPARATIVE ANALYSIS OF CLASSIFICATION MODELS FOR ACCURATE GRADUATION PREDICTIONS

Dr. Abdullah A. Al-Majardh
King Khalid University, Kingdom of Saudi Arabia, King Khalid University, Faculty of Science and Arts, Majardh, Computer Science Department

Abstract
Data mining, also known as knowledge discovery in databases (KDD), involves extracting valuable, previously unknown information from large data volumes. This field is gaining significant importance in the educational sector, particularly within universities. This paper aims to predict students' final year grades using classification-based data mining techniques, assessing the performance of three algorithms – Naïve Bayes, J48, and SVM – to improve educational quality. By comparing these classification algorithms, we can evaluate their current efficiency and effectiveness. Various performance measures are utilized to compare the results from these classifiers. Our findings indicate that the J48 classifier achieves the highest accuracy among the tested classifiers, making it a valuable tool in predicting student graduation outcomes.

Keywords: Classification, Naïve Bayes, Educational Data Mining, Prediction, Classification Algorithms.

Introduction

The higher education landscape is becoming increasingly competitive, with institutions vying to attract high-quality students, retain them throughout their academic journey, and ultimately graduate them into the workforce (Tinto, 2017). It has therefore become essential for educational institutions to accurately assess and predict student graduation outcomes to inform their decision-making, resource allocation, and the development of targeted intervention strategies (DiCerbo, 2014). Machine learning techniques, specifically classification models, have emerged as a promising approach to predict student graduation outcomes based on available data (Marquez-Vera et al., 2013). This study aims to assess the predictive performance of various classification models in predicting graduation outcomes, thereby contributing to the development of more effective graduation prediction systems in higher education.

Several factors can influence student graduation outcomes, including academic performance, socio-economic background, and institutional support (Tinto, 1993). Consequently, the development of accurate prediction models requires the consideration of a wide range of variables, as well as the application of appropriate machine learning techniques (Aulck et al., 2016). Classification models, a type of supervised learning technique, have been widely applied in the educational domain to predict student outcomes, such as dropout risk, academic performance, and graduation likelihood (Kotsiantis et al., 2004). These models use a set of input features to classify instances into predefined

categories, such as graduating or not graduating (Witten et al., 2011). Commonly used classification models in predicting student outcomes include logistic regression, decision trees, support vector machines, and neural networks (Huebner, 2018).Previous studies comparing the predictive performance of classification models in the educational domain have reported mixed results. For example, Marquez-Vera et al. (2013) found that decision trees outperformed other models in predicting student dropout risk, while Huebner (2018) reported that logistic regression was the most accurate model for predicting graduation outcomes. The inconsistent findings may be attributed to differences in the input features used, the data preprocessing techniques applied, and the performance metrics employed for model evaluation (DiCerbo, 2014). Therefore, a comprehensive comparative analysis of classification models is necessary to determine their relative strengths and weaknesses in predicting graduation outcomes and to guide the selection of appropriate models for specific contexts. In this study, we conduct a comprehensive comparative analysis of multiple classification models, including logistic regression, decision trees, support vector machines, and neural networks, using a large dataset of student records from a higher education institution. We employ rigorous data preprocessing techniques, such as feature selection and normalization, to ensure the quality and comparability of the input features used in the models (DiCerbo, 2014). We also use a range of performance metrics, including accuracy, precision, recall, and F1 score, to evaluate the predictive performance of the models (Witten et al., 2011).

By comparing the performance of different classification models in predicting graduation outcomes, this study aims to provide insights into the relative strengths and weaknesses of these approaches and to inform the development of more effective graduation prediction systems in higher education. Furthermore, our findings will contribute to the growing body of literature on the application of machine learning techniques in the educational domain and provide guidance for future research in this area.

DATA MINING CLASSIFICATION

Classification is a data mining task that divides data sample into target classes. These techniques based on supervised learning approach which having known class categories and it is used two methods, binary and multilevel. Dataset are partitioned as training and testing dataset and the classifier is trained by using training dataset. The correctness of classifiers could be tested using test dataset. Classification is a data mining task that divides data sample into target classes. These techniques based on supervised learning approach which having known class categories and it is used two methods, binary and multilevel. Dataset are partitioned as training and testing dataset and the classifier is trained by using training dataset. The correctness of classifiers could be tested using test dataset.

1. J48

J48 is an open source java implementation of the c4.5 algorithm in the WEKA data mining tool.c4.5 is an improvement used to generate a decision tree developed by Ross Quinlan C4.5 is a software extension and thus improvement of the basic ID3algorithm designed by Quinlan. the decision trees generated by C4.5 can be used for classification and for this reason, C4.5 is often referred to as a statical classifier for inducing classification rules in the form of decision trees from a set of given examples C4.5 algorithm was introduced by Quinlan C4.5 is an evolution and refinement of ID3

that accounts for unavailable values continuous attribute value ranges pruning of decision trees, rule derivation and so on ,asset of records are given.

2. Naïve Bayesian algorithm

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

3. SVM learning algorithm

Support vector machine is used for classification which is also a supervised learning method. there are three research papers that have used support vector machine algorithms as their technique to analyzing student's performance.

CLASSIFICATION ACCURACY

Accuracy is defined as the proportion of correct classification from overall number of cases and it depends on confusion matrix. Table 2 shows the confusion matrix that illustrates the number of correct and incorrect predictions made by the classification model compared to the actual value.

1. Correctly classified instance:

The correctly classified instance show the percentage of test instance that were correctly and in correctly classified the percentage of correctly classified instances is often called accuracy or sample accuracy

2. Kappa statistics:

Kappa is chance –corrected measure of agreement between the classification and true classes.

3. Confusion matrix:

A confusion matrix, some times called classification matrix is used to assess the prediction accuracy of model. it measure whether amodel is confused or not, that is whether the model is making mistakes in it predictions. The confusion matrix can be obtained from asset of different scales to compare classifications , including accuracy, which is widely used

The classifiers are evaluated by a confusion matrix which is a combination of four outcomes. In binary classification, the output is either positive or negative. The four different classifications are:

True positives (TP)-accurate positive prediction

False positives (FP)-wrong positives prediction

True negatives (TN)-accurate negative prediction

False negatives (FN)-wrong negative prediction

The effectiveness metrics for classifier used in the research are:-

Precision (p):-

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

Number of true positives classifications divided by the sum of true positives and false positive classifications.

- Recall(R):

TP

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

i.e number of true positives classifications divided by the sum of true positive and false negative classifications.

F1-SCORE-

F1-score is the harmonic mean of precision and recall

$$F1 - \text{score} = \frac{2 * P * R}{(P + R)}$$

Accuracy -

Accuracy is measured by dividing the number of correctly classified instances by the total number of instances.

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Error rate} = \frac{FP + FN}{TP + FP + TN + FN}$$

Mean Absolute Error (MAE):-

MAE measures the average magnitude of errors in asset of prediction it is the summation of the differences between predicted and observation divided by the total number of test samples.

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Root Mean Square Error (RMSE):-

It is the square root of the summation of the squared differences between predicted and actual observations, divided by the number of total test samples.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

ROC curve:

It is another way to evaluate the performance of the classification [12] where FP values are represented on the y axis and TP values on the y axis

$$T_{PR} = \frac{T_P}{TP + FN}$$

$$F_{PR} = \frac{F_P}{TN + FP}$$

Area under curve (AUC):

International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

Another utility called (area under the curve) helps analyze the overall performance of the classification and the ideal classification has AUC.

The ROC is a good visualization tool to identifying the performance of classifier, we times need a numerical value for comparison purpose.

WEKA TOOLS

WEKA is graphical user interface (GUI), it's an open source software developed at Waikato University in New Zealand. It contains four applications; explorer, experimental, knowledge flow and the command line interface (CLI) and also contains tools for data pre-processing, classification, clustering, regression and visualization.

The pre-processing is an important step that is used to extract and improve the quality of data. WEKA tool import dataset from a proper file like attribute relation file format which is the preferable one. Figure 2 and Figure 3 show the output of data pre-processor and model visualization in WEKA, respectively.

3. RELATED WORK

Many research studies have been done in educational data mining to predict the students' performance

In [7], the final CGPA of students was predicted using multiple linear regression and correlation to analyse the yearly GPA, and various inferential statistics were developed. The study determined the correlation between the first-year result and the final-year result of the student. With the aid of a regression plot, the students' GPA for the five years of study was fitted using multiple linear regressions in order to explain how the GPA for each year contributed to the variations in the final CGPA of the students at graduation.

In [8] features such as student attendance, average scores, relevant course data, the level of student participation in class etc. were deployed in a data mining model for predicting the performance of 908 students.

In [9] a decision tree model was applied to predict the probability of failure of 1,547 students such that relevant knowledge can be acquired that will enable the management team to be able to deploy adequate and early intervention. In the study, the student grades were classified into five categories, and these are: excellent, very good, good, acceptable and fail. Ten input features that include the student's department, high school grades, level of participation in class, attendance, midterm scores, lab reports, homework grades, seminar score, completion of assignments and the overall grades were applied in the decision tree model developed In [10] by using decision tree classifiers, the likelihood of a student to drop out of an institution was predicted through educational data mining.

In [11], association, classification, clustering and outlier detection data mining techniques were applied to analyse 3,314 graduate student performance records over a fifteen-year period. The dataset was analysed using Rule Induction, Naïve Bayesian classifier, K-Means clustering algorithm followed by density-based and distance-based outlier detection methods. 18 attributes of the student dataset were considered, and only 6 attributes: matriculation GPA, gender, specialty of the students, the city of the student, the grade and the type of secondary school attended were selected for the data mining analysis. The remaining 12 attributes were dropped due to their large variances and because some of the attributes are personal information that did not provide useful knowledge.

In [12] The unsupervised clustering analysis performed, identified four unique clusters in the dataset using k-means algorithm. Data mining method was applied by to evaluate student data towards identifying the key attributes that influence the academic performance of students.

This provides an opportunity for improving the quality of higher education.

In [13], data mining technique was Applied to analyse student data at a Bulgarian university. The student dataset that was analysed, contained the personal and pre-admission attributes of each student. The Decision Tree Classifiers (J48), k-Nearest Neighbour, Bayesian, Naïve Bayes classifiers, the OneR, and the JRip Rule learners were applied to extract knowledge from the student dataset, and accuracy of 52e67% was achieved. The result showed that the number of courses failed in the first academic year and the admission score of the student are two major features among the very influential features in the classification analysis.

In [14] the authors used WEKA data mining software for the prediction of final student mark based on parameters in two different datasets. Each dataset contains information about different students from one college course in the past fourth semesters. The IBK shows the best accuracy among other classifiers

In [15] the authors represents a study that will be helped to the students and the teachers to improve the result of the students who are at the risk of failure. Information's like Attendance,

Seminar and assignment marks were collected from the student's previous database, to predict the performance at the end of the semester. The authors used Naïve Bayes classification algorithm that shows a highest accuracy compared to other classification algorithms.

The researchers in [16] conducted a comparative research to test multiple decision tree algorithms on an educational dataset to classify the educational performance of students. The study mainly focuses on selecting the best decision tree algorithm from among mostly used decision tree algorithms, and provides a benchmark to each one of them and found out that the Classification and Regression Tree method worked better on the tested dataset, which was selected based on the produced accuracy and precision using 10-fold cross validations

Researchers in [17] provided an overview on the data mining techniques that have been used to predict students' performance and also it focused on how the prediction algorithm can be used to identify the most important attributes in a student's data. Under the classification techniques, Neural Network and Decision Tree are the two methods highly used by the researchers for predicting students' performance.

In [18], predictive analysis was carried out to determine the extent to which the fifth year and final Cumulative Grade Point Average (CGPA) of engineering students in a Nigerian University can be determined using the program of study, the year of entry and the Grade Point Average (GPA) for the

International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

first three years of study as inputs into a Konstanz Information Miner (KNIME) based data mining model. Six data mining algorithms were considered, and a maximum accuracy of 89.15% was achieved. The result was verified using both linear and pure quadratic regression models, and R2 values of 0.955 and 0.957 were recorded for both cases. This creates an opportunity for identifying students that may graduate with poor results or may not graduate at all, so that early intervention may be deployed.

In [19] analyze and evaluate the university students' performance by applying different data mining classification techniques by using WEKA tool. The highest accuracy of classifier algorithms depends on the size and nature of the data. Five classifiers are used NaiveBayes, Bayesian Network, ID3, J48 and Neural Network Different performance measures are used to compare the results between these classifiers. The results show that Bayesian Network classifier has the highest accuracy among the other classifiers.

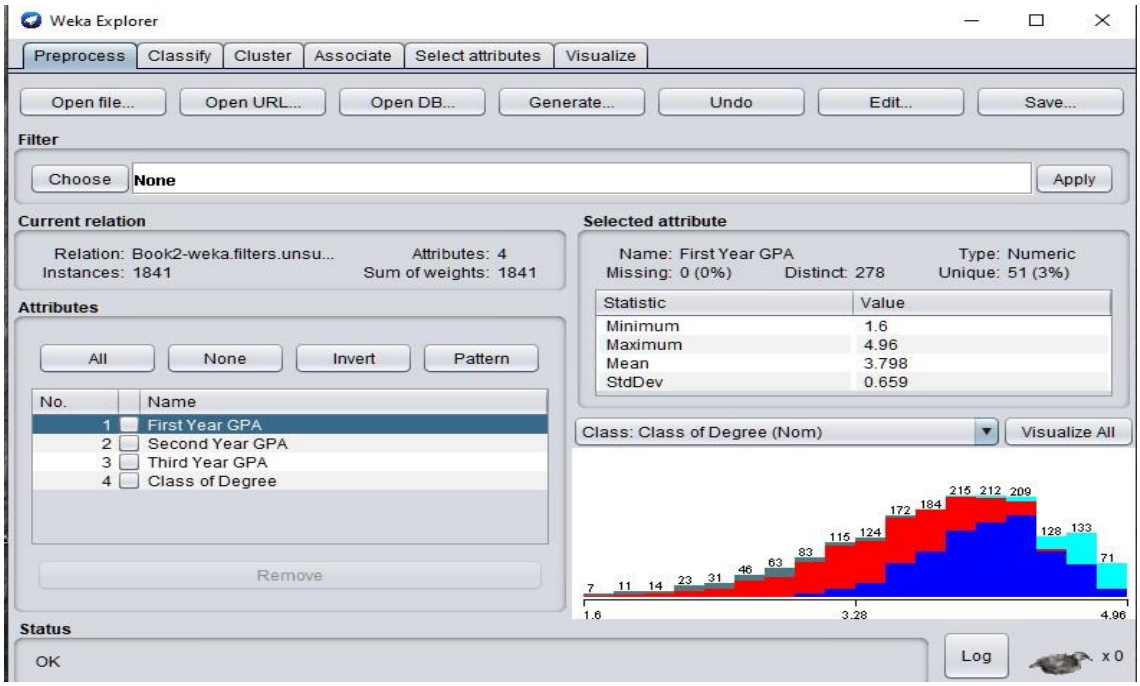
4. Experiment & results

Data set:

The data set name is full data .CSV this consist of the following 10 feature. [sample, college , year of entry , first year GPA, second year GPA ,third year GPA , fourth year GPA ,fifth year
GPA, final CGPA ,class of degree]with 1281 instance see figure 4.1

4. EXPERIMENTAL RESULTS AND DISCUSSION

The experimental results and discussion have done on selecting 1841 instance three selected classification algorithms were used, Naive Bayes, J48 and SVM and each one has its own characteristics to classify the data set. The data set consist of the following 4 feature. [first year GPA, second year GPA, third year, class of degree] with 1841 instance see figure 4.1



Data set figure (4.1)

4.1 Experiment with data set:

International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

The following part describes the measurement of the performance using (Naïve Bayes, J48.SVM) algorithm evaluating their results using training and testing technique Implementation algorithms of classification by using WEKA tools.

4.1.1 Result of decision tree (J48)

The experiment was conducted using decision tree c4.5 algorithm in WEKA to classify data.

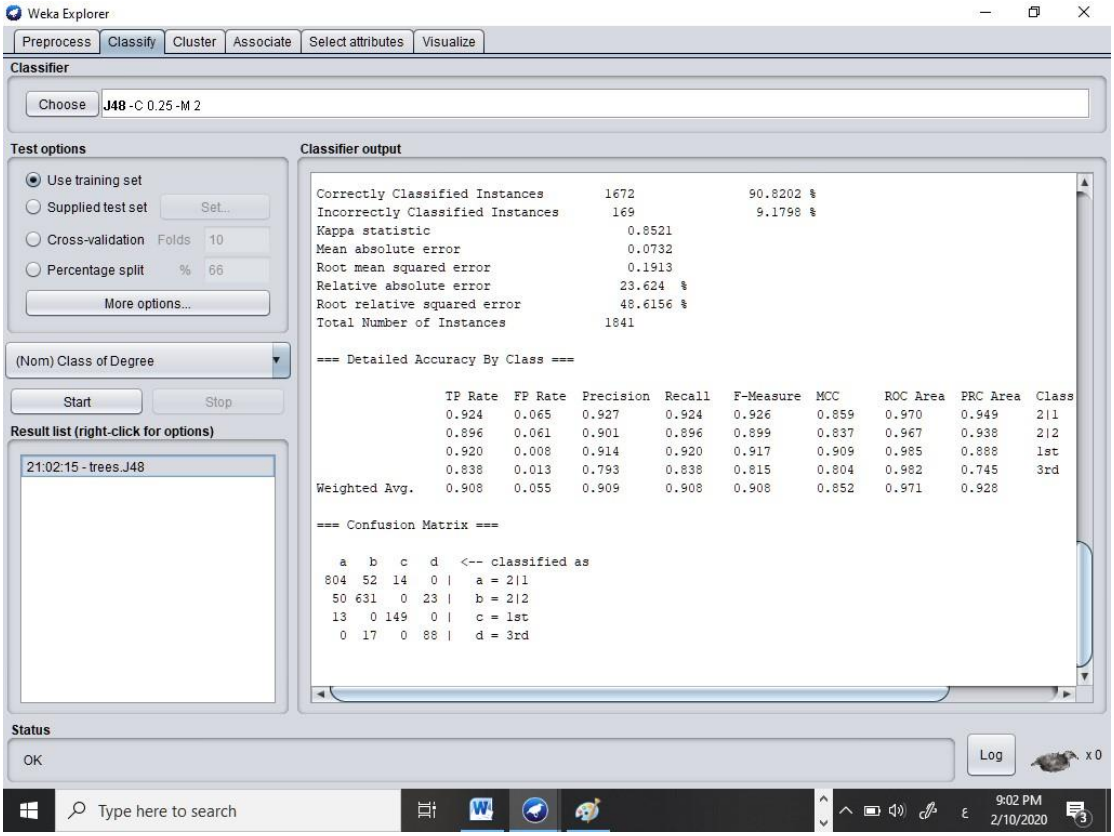


Figure (4.2) show the result of classification model using c4.5 algorithm

4. 1.2 Result of Naïve Bayes:

The experiment was conducted using Naïve Bayes algorithm in WEKA to classify data figure 4.3 show the result of classification model using Naïve Bayes algorithm.

International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

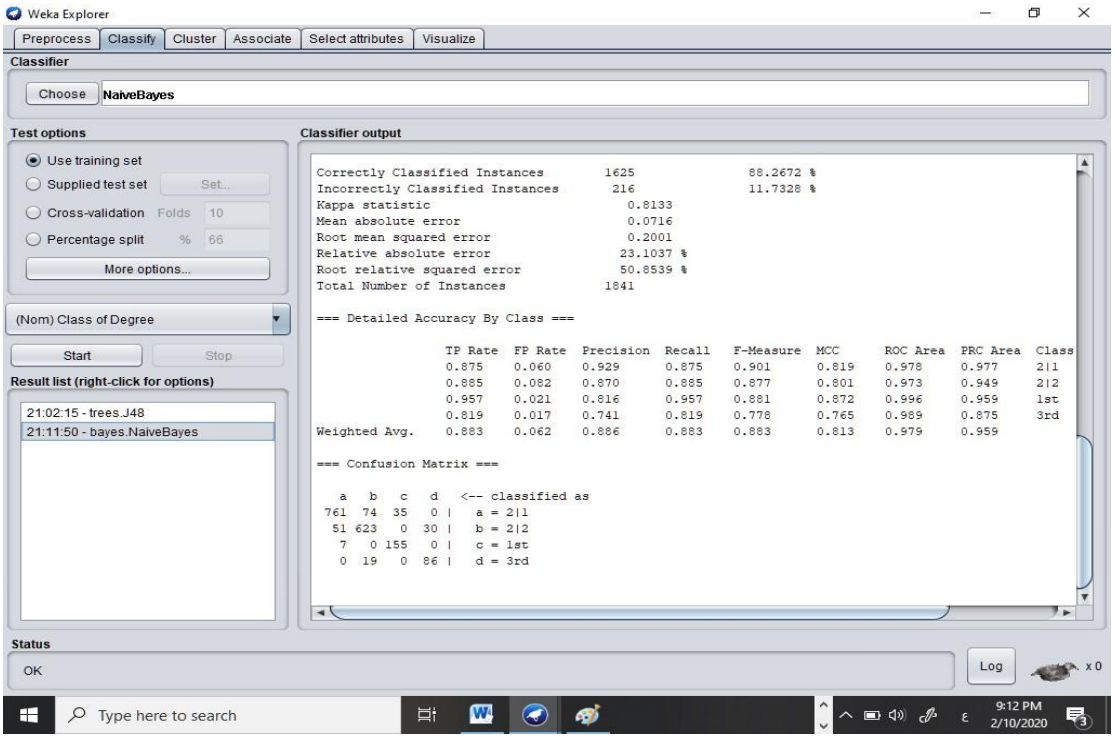


Figure (4.3) Result of classification model using Naïve Bayes algorithm

4. 1.3 Result of SVM:

The experiment was conducted using SVM algorithm in WEKA to classify data figure 4.4 show the result of classification model using SVM algorithm.

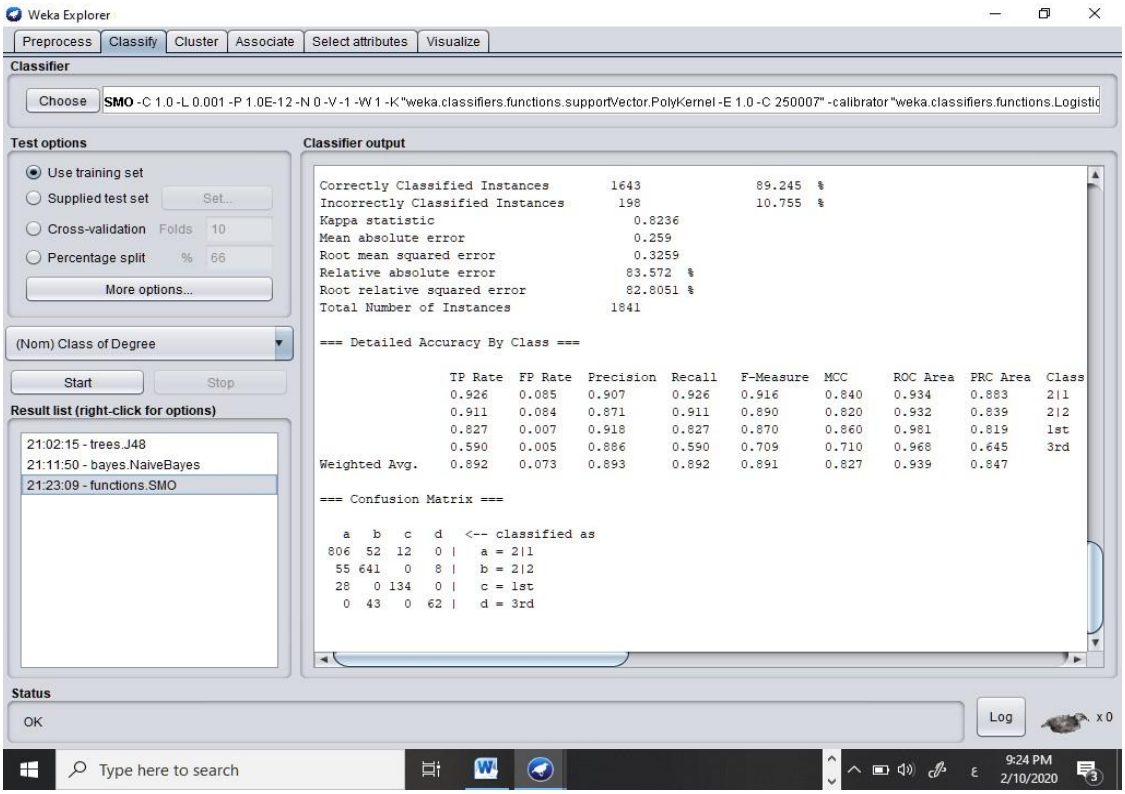


Figure 4.4 classification model using SVM algorithm.

International Research Journal of Biology and Biotechnology
Volume 10 Issue 3, July-September 2022
ISSN: 2995-4487
Impact Factor: 7.01
http://kloverjournals.org/journals/index.php/bb

In the Table 4.1 shows performance results of all classifiers.
Table 3 shows performance results of all classifiers by using WEKA, and Figure 4.5 shows the accuracy performance of classification techniques.
Table (4.1) performance results of all classifier

Criteria	**classifier**		
	(J48)	**Naïve Bays**	**SVM**
Correctly classified instance	**1672**	**1625**	**1643**
Incorrectly classified instance	**169**	**216**	**198**
Time to build model(sec)	**0.12**	**0.01**	**0.09**
Accuracy (%)	**% 90.8**	**% 88.3**	**% 89.3**

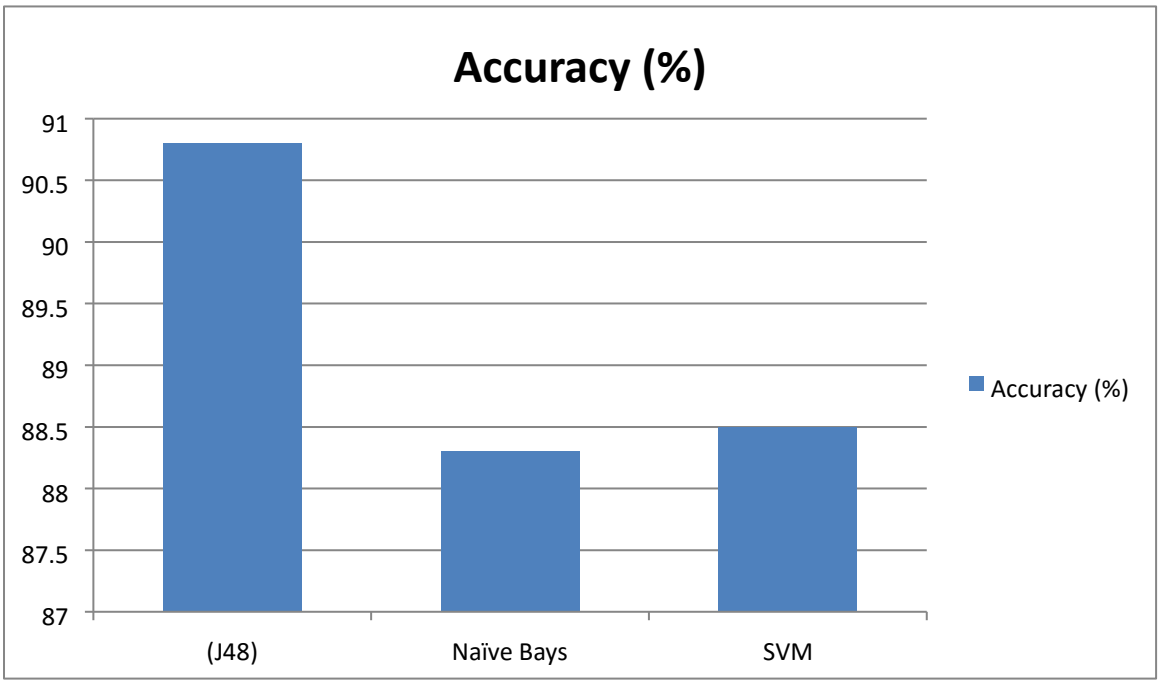


Figure 4.5 accuracy performance of classification techniques
Table 2. Error measures in weak

Criteria	**classifier**		
	(J48)	**Naïve Bays**	**SVM**
Kappa statistics	**0.8521**	**0.8133**	**0.8236**
Mean absolute error	**0.0732**	**0.0716**	**0.259**
Root mean squared error	**0.1913**	**0.2001**	**0.3259**
Relative absolute error	**% 23.624**	**% 23.1037**	**% 83.572**
Root relative squared error	**% 48.6156**	**% 50.8539**	**% 82.8051**

According to result In table 3, the **J48** classifier has more correctly classified instances than other classifiers, which is usually referred to the best accuracy model. The graphical representation in Figure 4 shows that the best classifier of students' performance based on their dataset is the **J48** classifiers. In the result, **J48** has an efficient classification among other classifiers.

CONCLUSION

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in.

This paper predict the grad of final year of students using the data mining technique classification in order to evaluate the current performance of Algorithms and compares between the classification algorithms and take efficient to enhance the quality of education

Data set contains of 1842 instance and for attributes. three classifiers are used and the comparisons are made based on the accuracy among these classifiers and different error measures are used to determine the best classifier. Experiments results show that Bayesian Network has the best performance among other classifiers. In future work, more dataset instance will be collected and will be compared and analyzed with other data mining classification in deferent data set size and tools.

References

Agarwal, R., Imielinski, T. and Swami, A. (1993) 'Database Mining: A

Performance Perspective', IEEE: Special issue on Learning and Discovery in Knowledge- Based Databases, pp. 914-925.

Agarawal, R. and R. Srikant, (1994) 'Fast algorithms for mining association rules', Proceedings of the 20th International Conference on Very Large Data Bases, San Francisco, CA., USA., pp: 487-499.

Maindonald, J. H. (1999) 'New approaches to using scientific data statistics, data mining and related technologies in research and research training' Occasional Paper 98/2, The Graduate School, Australian National University.

Quinlan, J. (1986), "Induction of Decision Trees," Machine Learning, vol. 1, pp.81106.

Berson, A., Smith, S. J. and Thearling, K. (1999) Building Data Mining Applications for CRM McGraw-Hill.

Oguntunde, Pelumi, Okagbue, Hilary, Oguntunde, Mooneye Adedoyin, Opanuga, A., 2018. Analysis of the inter-relationship between students' first year results and their final graduating grades. Int. J. Adv. Appl. Sci. 5 (10), 1e6.

Bucos, M., Dr_agulescu, B., 2018. Predicting student success using data generated in traditional educational environments. [Article]. TEM J. 7 (3), 617e625.

Ahmed, A.B.E.D., Elaraby, I.S., 2014. Data mining: a prediction for student's performance using classification method. World J. Comp. Appl. Technol. 2 (2),43e47.

Yadav, S.K., Bharadwaj, B., Pal, S., 2012. Mining Education Data to Predict Student's Retention: A Comparative Study arXiv preprint arXiv: 1203.2987.

Tair, M.M.A., El-Halees, A.M., 2012. Mining educational data to improve students performance: a case study. Int. J. Inf. Commun. Technol. Res. 2 (2), 140e146.

Al-Radaideh, Q., Al-Shawakfa, E., Al-Najjar, M., 2006. Mining student data using decision trees. In: Paper Presented at the the 2006 International Arab Conference on Information Technology. ACIT'2006

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

DiCerbo, K. E. (2014). Game-based assessment of persistence. Educational Technology & Society, 17(1), 17-28

Huebner, R. (2018). Predicting college success: Comparing data mining methods for predicting college success. Journal of Applied Research in Higher Education, 10(3), 285-298.

Kotsiantis, S. B., Pierrakeas, C. J., & Pintelas, P. E. (2004). Predicting students' marks in Hellenic Open University. In International Conference on Web-Based Learning (pp. 786-795). Springer, Berlin, Heidelberg.

Marquez-Vera, C., Romero, C., & Ventura, S. (2013). Predicting school failure and dropout by using data mining techniques. IEEE Transactions on Learning Technologies, 5(4), 266-278.

Tinto, V. (1993). Leaving college: Rethinking the causes and cures of student attrition. University of Chicago Press.

Tinto, V. (2017). Reflections on student persistence. In Higher Education Governance and Policy (pp. 5-23). Routledge.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2011). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Kabakchieva, Dorina, 2013. Predicting student performance by using data mining methods for classification. Cybern. Inf. Technol. 13 (1).

I. Milos, S. Petar, V. Mladen and A. Wejdan, Students' success prediction using Weka tool, INFOTEH-JAHORINA Vol. 15, March 2016. 684.

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

S.K Yadav, B. Bharadwaj, and S. Pal, 2012. Data Mining Applications: A Comparative Study for Predicting Student's Performance. International Journal of Innovative Technology & Creative Engineering (ISSN: 2045-711), Vol. 1, No.12, December.

A.Mohamed Shahiria,, W. Husaina, N. Abdul Rashida, "A Review on Predicting Student's Performance using Data Mining Techniques" Procedia Computer Science 72 (2015) 414 – 422, ELSEVIER.

Aderibigbe Israel Adekitan, Odunayo Salau, The impact of engineering students' performance Hilal Almarabeh, I.J. Modern Education and Computer Science, 2017, 8, 9-15 Published Online August 2017 in MECS (http://www.mecs-press.org/) DOI:10.5815/ijmecs.2017.08.02