ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

**Indian Journal of Science and Technology,** *Vol 12(14), DOI: 10.17485/ijst/2019/v12i14/142792, April 2019*

# Classification and Prediction of Student Academic Performance in King Khalid University-A Machine Learning Approach

**B. Prasanalakshmi[1]* and A. Farouk[2]**

[1]Department of Computer Science, King Khalid University, Guraiger, Abha 62529, Saudi Arabia; drsanaksa@gmail.com,
[2]Department of Chemistry, King Khalid University, Guraiger, Abha 62529, Saudi Arabia; afahmad@kku.edu.sa

## Abstract

**Objectives**: Universities accumulate huge amount of student's data in electronic form. Based on the information stored in the database filtering a data on certain criteria becomes difficult, when executed manually. Hence implementing tools that analyses the data in statistical, descriptive or computational ways are quite important to be considered. **Methods/Statistical Analysis**: This study presents an analysis on top ten machine learning algorithms used in classification and prediction. WEKA tool is used to conduct the experiment to know the accuracy and other result parameters on evaluating the categorical prediction of student performance. Also an analysis has been done to estimate the parameters based on the number of samples. **Findings:** The comparative analysis on the classification accuracy of around 12 classifiers of WEKA involving Rep Tree, Naive Bayes, J48, Bagging, lBK, Multilayer Perceptron, Random Forest, Random Tree, Stacking, AdaBoost, Logistic and SMO were analysed on datasets in varying number of instances. Based on the results obtained best 5 methods are chosen and compared on the entire dataset for prediction results. Ten machine learning algorithms were considered wherein the results such as accuracy in classification, Kappa statistic, and Mean absolute error are considered and compared. Bagging, Random Forest, lBK, Random Tree was filtered at the first level based on kappa statistic. In the second level filter based on accuracy lBK, Random Tree was considered as the final suitable models for the provided dataset. **Application/Improvements:** Developing a questionnaire among students and teachers is to be done to evaluate and predict the results in various angles based on various parameters. The positive factors and the negative factor contribution for the result of the institution are to be analysed.

**Keywords:** Educational Data Mining, Analysis, Prediction, Machine Learning, Student Performance, WEKA

## 1. Introduction

Machine learning uses the educational data mining techniques to predict the exact results on the student performance thereby creates an initiative for the educational institutions to rise up the results of their institution by looking over the parameters that affects their academic position in global market. This area on educational data mining improves the pedagogical strategy. Students' academic performance is a crucial deciding factor in building their future[1,2]. Machine learning includes developing a new model for the proposed work. Even though, many machine learning algorithms exists, some algorithms are of concern in all fields of research. For categorical analysis and prediction 43 algorithms are available for classifying data , but ease of consolidation only 10 algorithm of peak performance on considerable parameters are analyzed in this work. Many tools exist to test on the data for machine learning algorithms, but WEKA seems to be user interactive and easy to be used even for nonprogrammers, hence WEKA is chosen as a tool to identify the algorithm which can be used as a base for development of new model in predicting student performance. As in Figure 1 the entire process of machine learning depicts in to following steps in major.

**Author for correspondence*

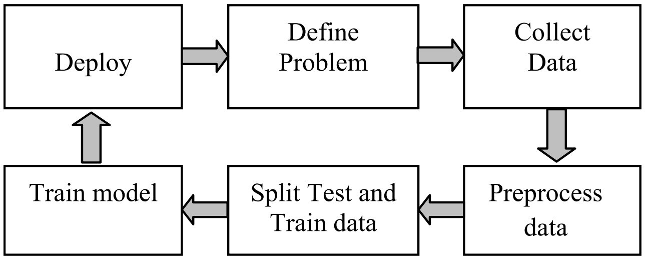

**Figure 1:** Machine learning process.

Steps used for predicting a data in machine learning involves:

- Data Gathering involves collecting data from real-time environment and segregating the data according to the requirement of the prediction result.
- Pre-processing data
- Classify using model
- Save model that train data
- Apply saved model for test data
- Predict result and estimate accuracy parameters

## 2. Related Works

Praneet et al.[4] in his work projects out the importance of predicting the results of students in the field of education. The real-time dataset of student academic records is tested and applied on various classification algorithms such as multilayer Perception, Naïve Bayes, SMO, J48 and REP Tree using WEKA an Open source tool. As a result, statistics are generated based on all classification algorithms and comparison of all five classifiers is also done in order to predict the accuracy and to find the best performing classification algorithm among all.

Ameerah et al.[5] provides the overview of data mining techniques that have been used to predict students' performance. The prediction algorithm used to identify the important attributes in a student's data is identified. Factors like Internal assessments, psychometric factors, CGPA, Social network interaction, Student demographic were considered.

Raheela et al.[6] made a case study on the student academic performance prediction using the cohort performance system considering only pre-university marks and marks of 1st and 2nd year courses, no socio-economic or demographic features, to predict the graduation performance in 4th year at university

## 3. Student Performance Model

In order to choose a tool and a best algorithm to serve as a base in developing a new model for the academic performance prediction WEKA is chosen. The academic results of previous semester based on 5 attributes like Student id, name, Mid-semester 1 and Mid-semester 2 contributing a major part in semester_internal marks are used to predict the final exam results. Even though the results of this analysis will turn up to be more positive when other factors contributing to the results like assignment, quiz are included. As per the curriculum of King Khalid university the entire marks of the course is split into two major equal halves semester_internal and semester_final marks each sharing the 100 marks of total equally. The semester_internal marks includes not only the Mid_semester 1 and Mid_semester 2 marks but also includes lab exams (if any), assignment, quiz, activities based on the course specification allocated for each course. As an initiative part this research work starts with prediction of the results of the exam that has been completed previous semester considering only the Mid_semester 1 and Mid_semester 2 marks. In the later case the results of this semester are to be predicted as a proposed future direction of this research considering various other factors.

The major classifiers designed in WEKA for machine learning purpose includes[3]:

- weka.classifiers.IBk: k-nearest neighbour learner
- weka.classifiers.j48.J48: C4.5 decision trees
- weka.classifiers.j48.PART: rule learner
- weka.classifiers.NaiveBayes: naive Bayes with/without kernels
- weka.classifiers.OneR: Holte'sOneR
- weka.classifiers.KernelDensity: kernel density classifier
- weka.classifiers.SMO: support vector machines
- weka.classifiers.Logistic: logistic regression
- weka.classifiers.AdaBoostM1: AdaBoost
- weka.classifiers.LogitBoost: logit boost
- weka.classifiers.DecisionStump: decision stumps (for boosting)

In the perspective of machine learning application, there are ten major algorithms that suits to the classification process of any research problem. Three major categories of Machine learning algorithms exist as Linear, Non-linear and Ensemble as shown in Table 1. Linear algorithms assume that the predicted attribute is a linear combination of the input attributes. The relationship between the input attributes and the output attribute being predicted are not considered into assumptions in Non-linear algorithms, whereas, Ensemble methods combines the predictions from multiple models in order to make more robust predictions.

### 3.1 Data Collection and Preparation Phase

For the analysis and prediction of the academic results only three main attributes Mid_semester 1, Mid_semester 2 marks and Semester_internal marks were taken as dependent attributes for classifying and predicting the results of final exams. The odd semester marks of 2018 in the College of Arts and Science, AhdRufidah a female branch of King Khalid University is incorporated for analysis. A total of marks of 2350 students were analyzed out of which around1880 data were considered as training data contributing to 80% of population data and the remaining 20% were used as test data for prediction. The prediction results are provided in the forthcoming section, based on which the appropriate method is to be chosen for future implementation.

### 3.2 Data Analysis Phase

The entire research process on predicting the student academic performance involves two main steps. The first step is to find a suitable machine learning algorithm that supports our requirement in predicting the academic performance based on the available dataset. The comparative analysis on the classification accuracy of around 12 classifiers of WEKA involving RepTree, NaiveBayes, J48, Bagging, lBK, MultilayerPerceptron, RandomForest, RandomTree, Stacking, AdaBoost, Logistic and SMO were analysed on datasets in varying number of instances. Based on the results obtained best 5 methods are chosen and compared on the entire dataset for prediction results. The training set is used to create the model for prediction and the testing set is used to check the model accuracy.

## 4. Results and Discussion

Ten machine learning algorithms were considered wherein the results such as accuracy in classification, Kappa statistic, and Mean absolute error are considered and compared. Initially the twelve classifiers are considered on evaluating the prediction results of 2350 instances, in which the results are observed as shown in Table 2. Based on the results of Table 2 an inference on rejecting the classifiers is arrived. Basically Kappa statistic is a measure to show the agreement of prediction with the true results. Numerical implications of Kappa statistic is to be very high which shows the coincidence or absorption of attribute values in predicting the results.

The Kappa statistic varies from 0 to 1, where,

- 0 = agreement equivalent to chance.
- 0– 0.20 = slight agreement.
- 0.21 – 0.40 = fair agreement.
- 0.41 – 0.60 = moderate agreement.
- 0.61 – 0.80 = substantial agreement.
- 0.81 – 0.99 = near perfect agreement.
- = perfect agreement.

$$k = \frac{Po - Pe}{1 - Pe} = 1 = \frac{1 - Po}{1 - Pe},$$

**Table 1:** Categories of ML algorithms for prediction and their respective methods in WEKA

| Machine learning algorithms category | Algorithm in ML | Function in WEKA |
|---|---|---|
| Linear | Linear Regression | function.LinearRegression |
| | Logistic Regression | function.Logistic |
| Nonlinear | Naive Bayes: | bayes.NaiveBayes |
| | Decision Tree | trees.J48 |
| | k-Nearest Neighbors | lazy.IBk |
| | Support Vector Machines | functions.SMO |
| | Neural Network | functions.MultilayerPerceptron |
| Ensemble | Random Forest: | trees.RandomForest |
| | Bootstrap Aggregation | meta.Bagging |
| | Stacked Aggregation | meta.Stacking |

Where:
$P_o$ = the relative observed agreement among raters.
$P_e$ = the hypothetical probability of chance agreement.

As far as the error measures are considered, they are expected to be at the least value which reveals the genuinity of prediction. The next parameter considered is accuracy which is depicted by (TP+TN)/(TP+TN+FP+FN). With a conclusion of these three factors the classifiers RandomTree, lBK, RandomForest, Bagging and J48 are taken into consideration for predicting results on test data in the ranking order suitable for the provided train data. The graphical result on the three factors considered for Analyzing and deciding the model is shown in Figure 2.

The chosen 5 classifiers are tested over the test data of 500 instances which shows perfection in deciding the prediction model. Table 3 shows the elimination of J48 classifier on predicting results since the remaining 4 classifiers as Bagging, Random Forest, lBK, RandomTree has an accuracy of approximately 60% and the kappa statistic to be approximately 0.37. On further filtering it is clear that lBK and RandomTree shows less mean absolute error when compared to the other classifiers as shown in Table 4. Hence lBK a WEKA implementation of K-nearest neighbor algorithm and RandomTree are chosen for final evaluation.

Since the confusion matrix decides on the accuracy of parameter statistic achieved, the comparative measure of the two classifiers are shown in Figure 2. The parameters arrived from the confusion matrix includes:

- Accuracy= (TP+TN)/(TP+TN+FP+FN)
- precision=TP / (TP + FP)
- sensitivity = TP / (TP + FN)
- specificity = TN / (FP + TN)
- F-score = 2*TP /(2*TP + FP + FN)

The resultant confusion matrix of the experiment is shown. To get a better understandability on the confusion matrix since all the predictions are on the diagonal of the matrix and the misclassifications outside the diagonal. In order to improve the accuracy of prediction all data were remodeled again and again in different classifier methods to create a balanced classification.

Best rules found on prediction using Apriori algorithm:

1. mid1_grade=A Mid2_grade=A Final_grade=A 643 ==>Sem_grade=A 638 <conf:(0.99)> lift:(1.71) lev:(0.11) [265] conv:(45.13)
2. mid1_grade=A Mid2_grade=A 820 ==>Sem_grade=A 809 <conf:(0.99)> lift:(1.7) lev:(0.14) [334] conv:(28.78)
3. mid1_grade=A Final_grade=A 787 ==>Sem_grade=A 755 <conf:(0.96)> lift:(1.66) lev:(0.13) [299] conv:(10.04)
4. Mid2_grade=A Final_grade=A 775 ==>Sem_grade=A 740 <conf:(0.95)> lift:(1.65) lev:(0.12) [291] conv:(9.07)
5. mid1_grade=A 1119 ==>Sem_grade=A 1015 <conf:(0.91)> lift:(1.57) lev:(0.16) [367] conv:(4.49)

**Table 2:** Comparative statistics for classifiers

| Classifiers | Correctly Classified Instances (%) | Kappa statistic | Mean absolute error | Time taken to build (sec) |
|---|---|---|---|---|
| Stacking | 47.63% | 0 | 0.2779 | 0 |
| AdaBoost | 51.55% | 0.2625 | 0.3182 | 0 |
| NaiveBayes | 51.55% | 0.2668 | 0.2182 | 0 |
| SMO | 53.21% | 0.2794 | 0.2768 | 0.03 |
| Logistic | 53.51% | 0.2706 | 0.2362 | 0.03 |
| MultilayerPerceptron | 54.36% | 0.3174 | 0.2302 | 0.01 |
| REPTree | 54.62% | 0.3046 | 0.2353 | 0 |
| J48 | 55.04% | 0.31 | 0.2342 | 0 |
| Bagging | 55.93% | 0.3267 | 0.2324 | 0 |
| RandomForest | 56.19% | 0.3307 | 0.224 | 0.09 |
| lBK | 56.19% | 0.3317 | 0.2232 | 0.39 |
| RandomTree | 56.19% | 0.3317 | 0.2232 | 0 |

## 5. Future Works and Conclusion

The deciding factor of the academic result is not just a single factor. It depends on several other factors. Based on the analysis made by this research a decision has been arrived on use a specific algorithm in predicting the results based on several factors. For this a future scenario on developing a questionnaire among students and teachers is to be done to evaluate and predict the results in various angles based on various parameters.

The questionnaire will include the contributing factor of result such as difficulties faced by students in the period of course, difficulties faced by teachers which contributes in pull down of results whereas on the other side the positive factors will also be collected to estimate the favouring factors of the academic performance.

## 6. Acknowledgement

Credit to college of arts and science, AhdRufaidah, Abha, Kingdom of Saudi Arabia.

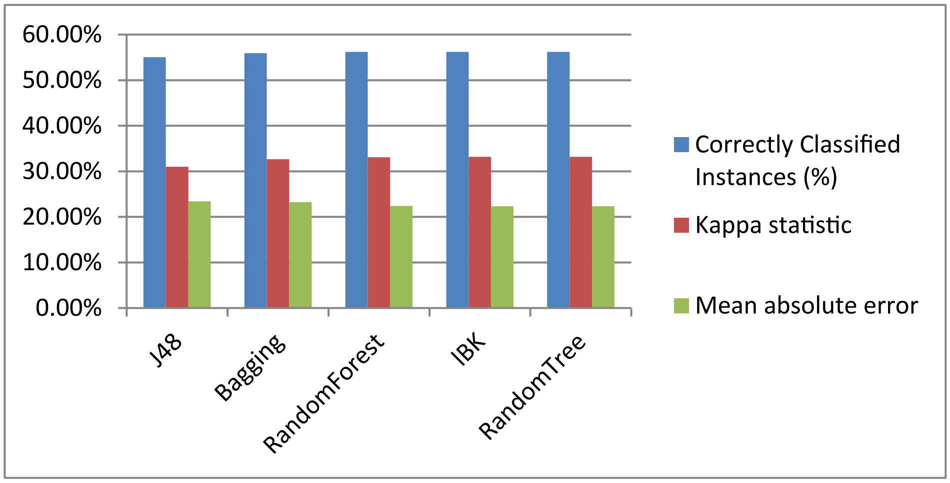

**Figure 2:** Model decision based on three parameters.

**Table 3:** Filtered result based on perfection

| Classifiers | Train (2350 instances) | | | Test (500 instances) | | |
|---|---|---|---|---|---|---|
| | Correctly Classified Instances (%) | Kappa statistic | Mean absolute error | Correctly Classified Instances (%) | Kappa statistic | Mean absolute error |
| J48 | 55.04% | 0.31 | 0.2342 | 59.12% | 0.3501 | 0.2245 |
| Bagging | 55.93% | 0.3267 | 0.2324 | 59.92% | 0.3657 | 0.222 |
| RandomForest | 56.19% | 0.3307 | 0.224 | 59.92% | 0.3666 | 0.2156 |
| lBK | 56.19% | 0.3317 | 0.2232 | 60.32% | 0.3737 | 0.215 |
| RandomTree | 56.19% | 0.3317 | 0.2232 | 60.32% | 0.37378 | 0.215 |

**Table 4:** Class precision and class recall of lBK and random forest

| lBK & RandomTree<br>D | | Actual | | | | | Class Precision |
|---|---|---|---|---|---|---|---|
| | | B | C | A | F | | |
| Predicted | D | 9 | 7 | 19 | 5 | 0 | 0.321 |
| | B | 6 | 41 | 24 | 40 | 0 | 0.471 |
| | C | 5 | 14 | 27 | 12 | 1 | 0.27 |
| | A | 8 | 22 | 15 | 221 | 0 | 0.789 |
| | F | 0 | 3 | 15 | 2 | 3 | 0.75 |
| Class Recall | | 0.225 | 0.369 | 0.458 | 0.831 | 0.13 | -- |

## 7. References:

1. Baker RS, Corbett AT, Koedinger KR. Detecting Student Misuse of Intelligent Tutoring Systems. Proceedings of the 7th International Conference on Intelligent Tutoring Systems; 2004. 531–40. https://doi.org/10.1007/978-3-540-30139-4_50.
2. Tang T, McCalla G. Smart recommendation for an evolving e-learning system: Architecture and experiment, International Journal on E-Learning, 2005; 4(1):105–29. https://www.learntechlib.org/primary/p/5822/.
3. https://github.com/Waikato/weka-3.8/blob/master/weka docs/README.
4. ParneetKaur, Manpreet Singh, Gurpreet Singh Josan. Classification and prediction based data mining algorithms to predict slow learners in education sector, Procedia Computer Science. 2015; 57:500–08. https://doi.org/10.1016/j.procs.2015.07.372.
5. Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid. A review on predicting student's performance using data mining techniques, Procedia Computer Science. 2015; 72:414–22. https://doi.org/10.1016/j.procs.2015.12.157.
6. Raheela Asif, Agathe Merceron, Mahmood K. Pathan. Predicting student academic performance at degree level: A case study, I.J. Intelligent Systems and Applications. 2015; 01:49–61.