An Analysis on Academic Performance of Students using a Hybrid Model for Higher Education

Jai Ruby [#1], K. David [*2]
[#] Research Scholar, Research and Development Center,
Bharathiyar University, Tamilnadu, India
[1] ajairuby@gmail.com
[*] Assistant Professor, H.H. Rajahs College, Pudukottai, Tamilnadu, India
[2] jdbdavid@gmail.com

***Abstract*— Education has been viewed as a key aspect in contributing to the welfare of the country. In modern era, Educational Institutions strive hard to improve the quality of education they render to the student community. Though many factors determine a good institution, the academic performance of the students pay a vital role in it. Data mining is a technique used to bring out such hidden knowledge which exists in the form of raw data in the repository. Many socio economic, Non-academic and academic factors influence the performance of the students. There are many well - known data mining classification algorithms such as ID3, SimpleCART, J48, NB Tree, MLP, Bayesnet etc which are used to predict the student performance. The model proposed here is mainly focused on finding the prediction accuracy of academic performance of students using a hybrid model which is a combination of two classification algorithms ID3 and MLP. The experimental model also prove that the accuracy can be improved by intelligently generating the training dataset for the hybrid model.**

Keyword- Educational Data Mining, Academic Performance, Prediction, Classification, Hybrid, ID3, MLP

I. Introduction

In this modern era, technological revolution and population explosion have changed the scenario of the Higher Education System. Educational institutions are becoming more competitive because of the number of institutions growing rapidly. To stay afloat, these institutions are focusing more on improving various aspects and one important factor among them is quality learning. For providing quality education, the institutions need to know about their strengths which are explicitly seen and which are hidden. To be competitive, the institutions should identify their own strengths that are substantially hidden and implement a technique to bring them out. In recent years, Educational Data Mining has put on a massive recognition within the research realm for extracting knowledge from a huge dataset. Data mining is widely used on educational dataset and it is termed as Educational Data mining (EDM). EDM has become a very useful research area [1]. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting the pattern from large repositories of data generated by or related to people's learning activities in educational settings. Key uses of EDM include learning and predicting student performance in order to recommend improvements to current educational practice. EDM can be considered as one of the learning sciences, as well as an area of data mining [2]. The technique behind the extraction of the hidden knowledge is a Knowledge Discovery process that extracts the knowledge from available dataset and creates a knowledge base for the benefit of the institution.

Many socio economic, non-academic and academic factors influence the performance of the students. Students' data was collected, pre-processed and data mining techniques are applied to discover association, classification, clustering and outlier detection rules and in all tasks knowledge was extracted that describes students' behaviour [3]. The study model [4] is mainly focused on exploring various indicators that have an effect on the academic performance of the students. The extracted information that describes student performance can be stored as intelligent knowledge and can be used by the institutions principally for predicting the students' academic performance in advance. To improve the prediction accuracy is one of the key issues in the Educational Data mining Field. In this paper, the researcher derived a hybrid model that uses two classification data mining algorithms MLP and ID3 on educational domain dataset for predicting students' academic performance. The researcher also made an effort to compare the prediction accuracy of various approaches that was carried out in early research. The experiment strengthens the fact that various factors identified in the study model [4] are really influencing in predicting the results.

This paper makes a new attempt to look into the higher educational domain of data mining to improve the prediction accuracy by using a hybrid model. Section 2 gives the methodology and techniques. Section 3. provides the general account of the model and the dataset under study. Section 4 predicts the academic performance of students using the hybrid model over the datasets and the comparative analysis. Conclusion and a discussion on future work are in the final section.

A. Related Work

Han and Kamber [5] describe data mining as a tool that help the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. J. Shana and T. Venkatachalam [6], used various feature selection methods and have found out the influence of features affecting the student performance. The authors have used a selected number of attributes and have not taken attributes like attendance, theory, laboratory etc. Brijesh Kumar Baradwaj, Saurabh Pal, in [7] conducted a study on a data set of size 50 Post Graduate students for mining educational data to analyze students' performance. Decision tree method was used for classification and to predict the performance of the students. Different measures that are not taken into consideration were economic background, technology exposure etc. Jai Ruby and K. David [4], presented a study on the educational data and identified that 7 factors out of 16 initial factors are high influencing factors for predicting the students' academic performance by using various feature selection techniques like Chi square, Information Gain, Correlation, Linear Regression and Gain Ratio. Ramaswami M., and Bhaskaran in [8] have constructed a predictive model using 772 students' records with 7-class response variables by using highly influencing predictive variables obtained through feature selection. Data mining approach applied here is to discover students' performance patterns in Maths, English and programming courses.

El-Halees. A [3] did a work on describing student behaviour with a dataset of size 151 that includes only personal and academic details of students. Classification based on Decision tree was done followed by clustering and outlier analysis. Z. J. Kovacic, in [9] presented a study on educational data mining to identify up to what extent the enrolment data can be used to predict student's success by using CHAID and CART data mining algorithms. Mohammed M. Abu Tair and Alaa M. El-Halees [10] applied the data mining for discovering knowledge from data that come from educational environment. Students' data had been collected from the college of Science and Technology for a period of 15 years [1993-2007]. The collected data was pre-processed and data mining techniques are applied for analysing graduate students' performance. MuslihahW.et.al.[11] have compared Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting and classifying student's academic performance. Students' data were collected from the data of the National Defence University of Malaysia (NDUM). H. W. Ian and F. Eibe gave a case study that used educational data mining to identify the behaviour of failing students who are at risk before the final exam [12].

Nguyen N et.al [13], compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of students of Under Graduate and Post Graduate students. The decision tree classifier provided better accuracy than the Bayesian network classifier. Different measures that were not taken into consideration were economic background, technology exposure etc. Bengio Y. et.al [14], discussed that neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification and regression. Neural networks have an advantage over other types of machine learning algorithms. Vasile P. B [15] presented that the use of classification models such as decision tree and Naïve Bayes in the education field to predict students' behaviour. M. Wook, et.al [16] compared two data mining techniques which are: Artificial Neural Network and the combination of clustering and decision tree classification techniques for predicting students' academic performance. Jai Ruby & David [17] compared various data mining algorithms using the student dataset considering only the influencing factors and proved that MLP, a neural network based classification shows a best prediction which is followed by ID3.

Romero and Ventura [18], had a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it had a specific requirement not presented in other domains. Kuyoro' et.al [19] worked on identifying the optimal algorithm suitable for predicting first-year tertiary students' academic performance based on their family background factors and previous academic achievement. Five decision tree algorithms, five rule induction algorithms and an artificial neural network function were taken for the study. It was discovered that random tree performance was better than that of other algorithms used in this study. Jai Ruby & David [20], presented a comparative study model to predict the accuracy of the academic performance of the students using Multi-Layer Perceptron algorithm. The experiment justified that the attributes identified in the study [4] are practically high influencing factors in predicting student performance.

II. DATA MINING AND TECHNIQUES

Data mining also termed as Knowledge Discovery in Databases (KDD) refers to extracting or “mining” knowledge in the form of rules, patterns or models from large amount of data. Fig.1 shows the process of extraction of knowledge as a result of mining the data.

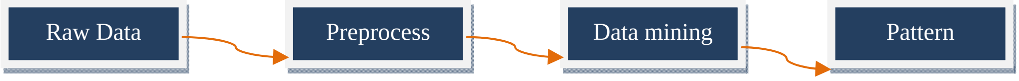


Fig. 1 – Conversion of data into knowledge

Knowledge Discovery process involves various steps like Data pre-processing, data mining, pattern evaluation in extracting knowledge from data. Knowledge Discovery is involved in a multitude of tasks such as association, clustering, classification, prediction, etc. Classification and prediction are functions which are used to create models that are constructed by analysing the data and then the model is used for assessing other data. Two steps are involved in classification. In the first step, a model that describes a predetermined set of classes or concepts is made by examining a set of training dataset. The learning is known as supervised learning as the class labels of all the records of the dataset are known. The models are usually in the form of classification rules or decision tree. In the second step, the model is put to test using a different data set that is used to estimate the predictive accuracy of the model. Various methods like holdout, random sub sampling, k-fold cross validation, stratified cross validation, bootstrapping are used to estimate the accuracy of the model. If the accuracy of the model is considered acceptable, the model can be used to classify the dataset for which the class label is not known in advance [5]. Basic techniques for classification are decision tree induction, Bayesian classification and neural networks. A number of well - known data mining classification algorithms such as ID3, REPTree, Simplecart, J48, NB Tree, BFTree, Decision Table, MLP, Bayesnet, etc., exist.

A. MLP

A neural network is a biologically simulated computational model. Fausett, L [22] had revealed that neural networks have been used in a large number of applications which include pattern recognition, classification and prediction.A neural network is a network that has the ability to learn from its background and improve its performance through learning. Multilayer Perceptron algorithm is one of the most widely used and common neural network. Multilayer Perceptron is a feed forward artificial neural network model trained with the standard back propagation algorithm that maps sets of input data onto a collection of acceptable output. An MLP consists of multiple layers of nodes in a directed graph, with every layer totally connected to the consequent one. These are supervised networks so they require a desired response to be trained. They learn how to transform input data into a desired response, so they are widely used for pattern classification and prediction.

B. ID3

ID3 - Iterative Dichotomiser 3 is a decision tree Induction algorithm. The ID3 algorithm was originally developed by J. Ross Quinlan at the University of Sydney. The ID3 algorithm induces decision trees from data. It is a supervised learning algorithm that is trained by examples for different classes. After being trained, the algorithm should be able to predict the class of a new item. To determine which attributes are the most important, ID3 uses the statistical property of entropy. Entropy measures the amount of information in an attribute. This is how the decision tree, which will be used in testing future cases, is built.

C. Feature Selection Techniques

An educational dataset may contain a number of student oriented attributes. Various attribute selection methods do exist to identify the attributes that make great impact. Some of the notable methods are chi-square, information gain, correlation, gain ratio, and regression. Chi-square test is a statistical method used to identify degree of association between variables [21].Information Gain and Gain Ratio are used to determine the best attribute. Linear Regression involves finding the best line to fit two variables so that one variable can be used to predict another and to find a mathematical relationship between them. Correlation is used to assess the degree of dependency between any two attributes.

III. DATASET, TOOL AND METHODOLOGY

The dataset used for this study was taken from PG Computer Application course offered by an Arts and Science College between 2007 and 2012. The data of 165 students were collected. Student personal and academic details along with their attendance were collected from the student information system. The collected information was integrated into a distinct table. Among the different attributes initially present using feature selection techniques like chi square, info gain, gain ratio, correlation and regression it was found that the high impact attributes that contribute for the performance of the students are Theory, Medium of Study, Previous Course studied, UG Percentage, Stay, Extra Curricular Activities and Family Income. The influencing attributes are selected and are used to classify and predict the student performance using weka data mining tool.

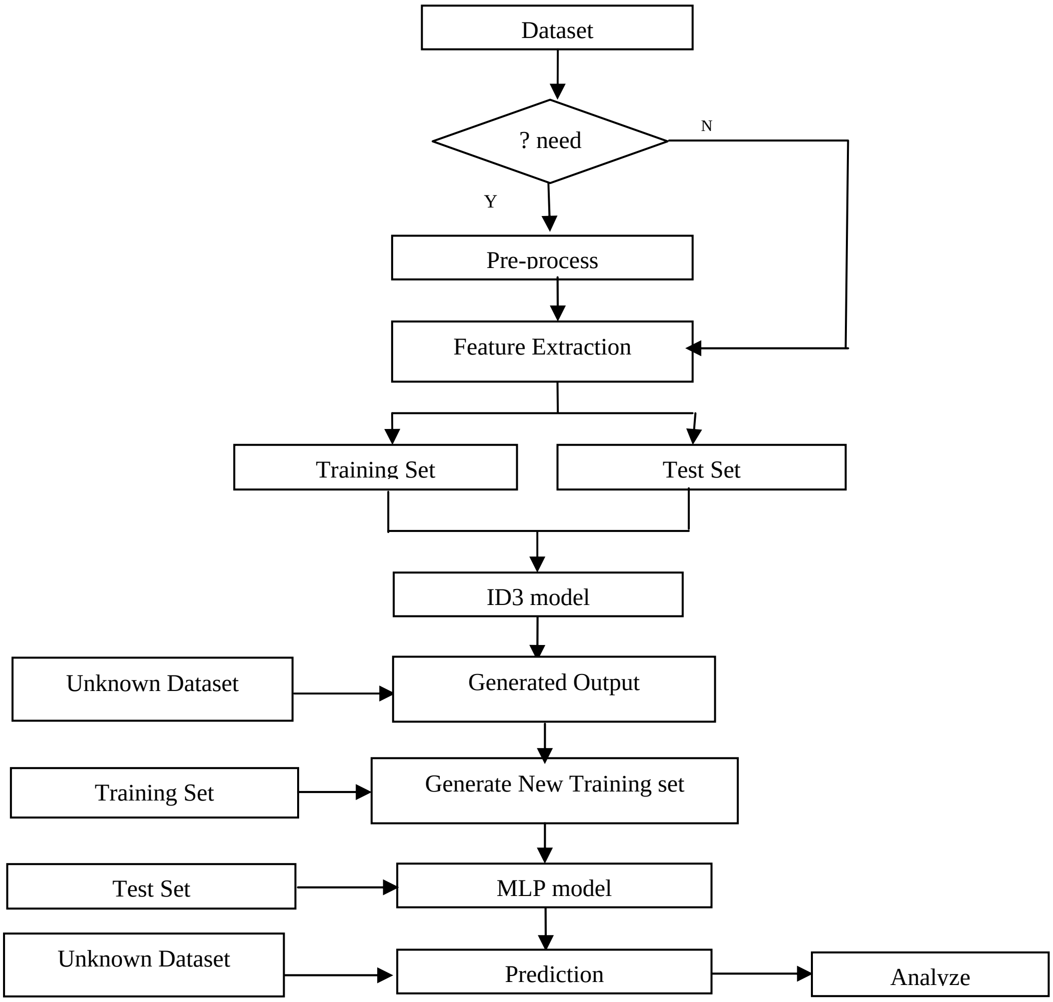


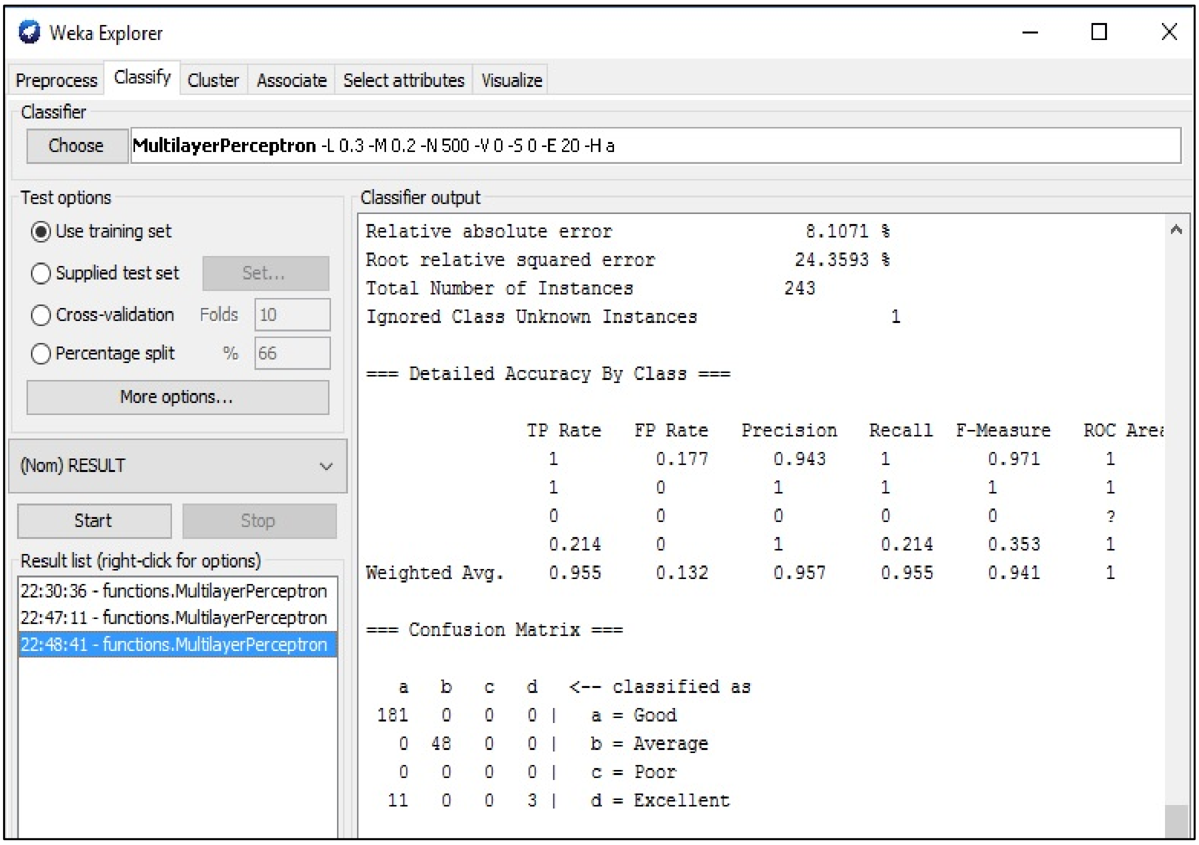
Fig. 2 Working of proposed hybrid model

Weka (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java under GNU General Public License, developed at the University of Waikato, New Zealand. The Weka workbench contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality. Weka tool contains many packages which include Filters, Classifiers, Clusters, Associations, and Attribute Selection. The Visualization tool in weka allows datasets and the predictions of Classifiers in a pictorial form. In Weka, datasets should be formatted to the ARFF format. The initial dataset of 165 records was split up into two sets. Two thirds of the data are allocated to the training set and the remaining one third is allocated to the test set. The training set helps in building the model and it is used for classification. Training and testing is performed k-times. The classify panel in the weka tool facilitates applying classification algorithms and to estimate the accuracy of the predictive model. Two different classifiers ID3 and MLP (Multi Layer Perceptron) were used in the model. This hybrid approach helps to analyse the enhancement in the prediction accuracy on student dataset than by using just an existing data mining classification algorithm.

The working of the proposed hybrid model shown in Fig. 2 consists of two phases. In the first phase, the given dataset is scrutinized for the need of pre-processing. If needed, it is pre-processed and it is subjected to feature extraction. This minimises the number of attributes that are not so relevant in prediction of academic results in a student dataset. The refined dataset is now split into two specific sets. One is termed as training set and other as test set as this forms the base of classification in data mining. Then a model is built applying the ID3 classification algorithm. The model is then used for predicting the result of the unknown dataset. In the second phase, using the existing training set and the result obtained from the previous phase, a new training set is generated. This is again used to develop a model using MLP classification algorithm. This model is finally used for predicting the result of unknown dataset. The results obtained are then used for analysis of the prediction accuracy of the hybrid model.

IV. EXPERIMENTAL RESULT ANALYSIS

ID3 and MLP were the two classification algorithms used in this hybrid model. The student data set of 165 records were pre-processed and converted from excel to Attribute Relation File Format. The high influencing attributes were identified using feature selection technique and a final dataset was prepared. The dataset was then split into two sets consisting of two-thirds as training set and one-third as a testing set. The training set is used to build a model based on ID3 and the test set is used to estimate the accuracy of the classifier. The Fig. 3 shows the classification of the training data set using MLP algorithm and Fig. 4 shows the classification of test data using MLP algorithm via weka tool. If the accuracy of the model is acceptable then it is used for the prediction of data for which the class label is unknown.



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

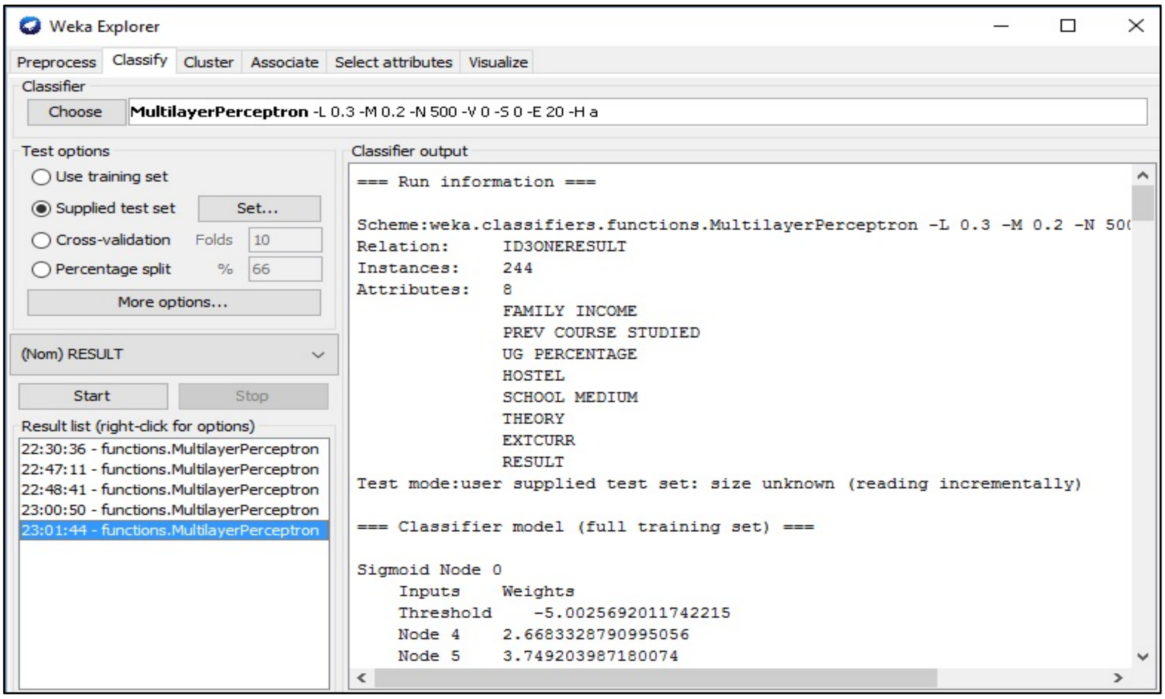
● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) RESULT

22:30:36 - functions.MultilayerPerceptron
22:47:11 - functions.MultilayerPerceptron
22:48:41 - functions.MultilayerPerceptron

```
Relative absolute error                  8.1071 %
Root relative squared error             24.3593 %
Total Number of Instances              243
Ignored Class Unknown Instances                  1

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall   F-Measure   ROC Area
                 1         0.177     0.943      1         0.971      1
                 1         0         1          1         1          1
                 0         0         0          0         0          ?
                 0.214     0         1          0.214     0.353      1
Weighted Avg.    0.955     0.132     0.957      0.955     0.941      1

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 181   0   0   0 |   a = Good
   0  48   0   0 |   b = Average
   0   0   0   0 |   c = Poor
  11   0   0   3 |   d = Excellent
```

Fig. 3 Classification of Training Dataset using MLP



MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

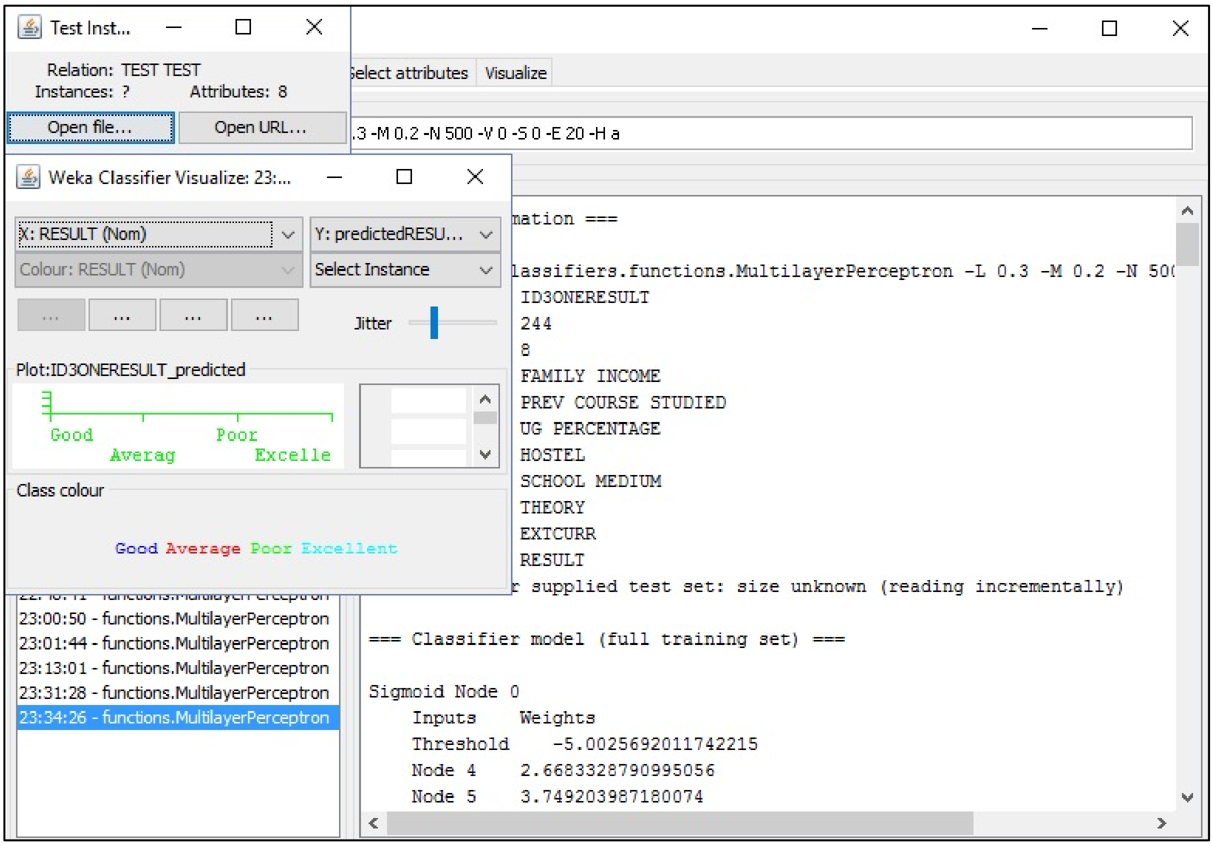
(Nom) RESULT

Result list (right-click for options)
22:30:36 - functions.MultilayerPerceptron
22:47:11 - functions.MultilayerPerceptron
22:48:41 - functions.MultilayerPerceptron
23:00:50 - functions.MultilayerPerceptron
23:01:44 - functions.MultilayerPerceptron

```
=== Run information ===

Scheme:weka.classifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 50(
Relation:     ID3ONERESULT
Instances:    244
Attributes:   8
              FAMILY INCOME
              PREV COURSE STUDIED
              UG PERCENTAGE
              HOSTEL
              SCHOOL MEDIUM
              THEORY
              EXTCURR
              RESULT
Test mode:user supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    -5.0025692011742215
    Node 4    2.6683328790995056
    Node 5    3.749203987180074
```

Fig. 4 Classification of Test Dataset using MLP

The sample set was divided into 5 sets of distinct two-thirds records and 5 sets of distinct one-third records. They are the data set used in Run1 through Run5 respectively. In each run, three new sets of data whose class labels were unknown was given for prediction. Since we used three different sets of new data for prediction, the average of the three results was considered for each run. Fig. 5 shows the prediction of new test data whose label is unknown using MLP algorithm via weka. The same experimental setup is repeated using ID3 classification algorithm and prediction accuracy obtained during different runs are recorded.



Relation: TEST TEST
Instances: ? Attributes: 8

.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

X: RESULT (Nom)
Y: predictedRESU...
Colour: RESULT (Nom)
Select Instance

Plot:ID3ONERESULT_predicted

23:00:50 - functions.MultilayerPerceptron
23:01:44 - functions.MultilayerPerceptron
23:13:01 - functions.MultilayerPerceptron
23:31:28 - functions.MultilayerPerceptron
23:34:26 - functions.MultilayerPerceptron

```
mation ===

lassifiers.functions.MultilayerPerceptron -L 0.3 -M 0.2 -N 50(
 ID3ONERESULT
 244
 8
 FAMILY INCOME
 PREV COURSE STUDIED
 UG PERCENTAGE
 HOSTEL
 SCHOOL MEDIUM
 THEORY
 EXTCURR
 RESULT
r supplied test set: size unknown (reading incrementally)

=== Classifier model (full training set) ===

Sigmoid Node 0
    Inputs    Weights
    Threshold    -5.0025692011742215
    Node 4    2.6683328790995056
    Node 5    3.749203987180074
```

Fig. 5 Prediction of unknown dataset using MLP

In the proposed hybrid model there are two phases. The same experimental setup is used in both the phases. In the first phase, ID3 classification algorithm is used to build the model and then it is used for the prediction of new test data whose label is unknown. In the second phase, from the generated output, the optimized results were clubbed along with the training set and a new training set is generated. Using the generated new training set, a new model is created using MLP algorithm. The new model is then used to predict the academic performance of the unknown student dataset.

TABLE I. Prediction Accuracy of the Various Classification Models in Percentage for 5 Different Dataset

Models	Run - 1	Run – 2	Run - 3	Run - 4	Run – 5
ID3	64	63	77	50	71
MLP	74	60	58	49	83
Hybrid ID3+MLP	89.7	87.5	88.6	46	90

Table I shows prediction accuracy of the Classification Algorithms in percentage for a different combination of training data set, testing data set and new unknown datasets for which prediction of class label was carried out. The average of the 5 runs of the ID3, MLP and the hybrid models were tabulated in Table II.

TABLE II. Average Accuracy of 5 Run of ID3, MLP And Hybrid Classification Algorithms in Percentage

Classification Models	Prediction Accuracy
ID3	65.0
MLP	64.8
Hybrid ID3-MLP	80.36

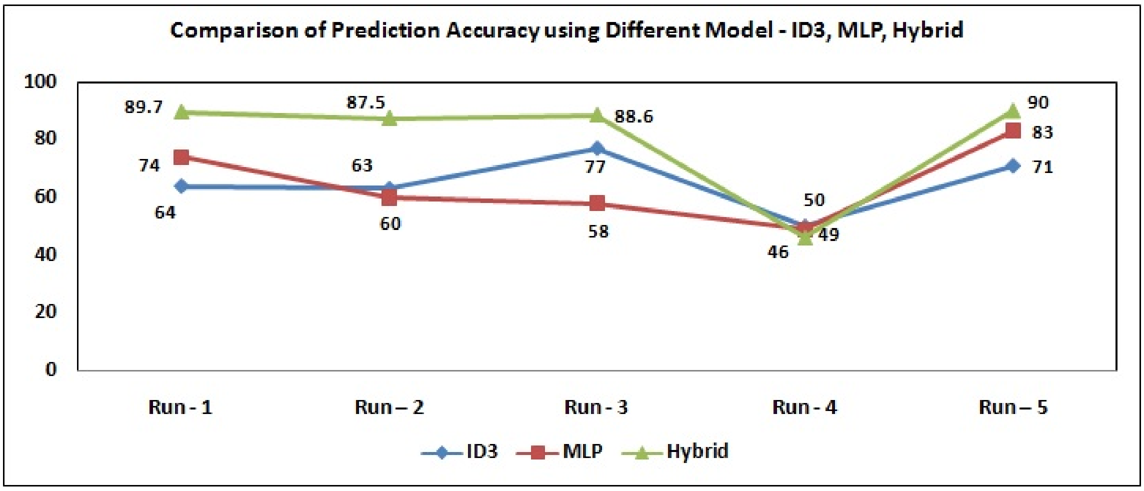


Fig. 6 Comparison of Prediction Accuracy using ID3, MLP and Hybrid model

Fig. 6 shows the comparison of prediction accuracy of different datasets using ID3, MLP and proposed Hybrid classification models. The prediction accuracy percentage of dataset using hybrid model shows a better result than the other models of ID3 and MLP. The average accuracy percentage of 5 different datasets is found to be almost same for both ID3 and MLP. There is a drastic improvement in the prediction percentage if a hybrid model is used. The experimental results show that the prediction accuracy highly depends on the classification algorithm and the generated training dataset.

V. CONCLUSION

This experimental model is mainly focused on analyzing the prediction accuracy of the academic performance of the students using different classification algorithms like ID3, MLP and a hybrid model. The experimental results prove the attributes chosen from the original dataset are really highly influential and it is enough to improve the prediction accuracy of the unknown classes. The proposed hybrid model proved to be a better performing model when compared with the existing models. This analysis helps the institution to know the academic status of the students in advance and can concentrate on weak students to improve their academic results. To develop an algorithm to generate a better training set and to represent the knowledge created in a standard form would be a future work.

REFERENCES

[1] Baker R.S.J.D., & Yacef K, 2009, "The state of educational data mining in 2009: A review and future vision", Journal of Educational Data Mining, I, 3-17.
[2] Monika Goyal & Rajan Vohra, "Applications of Data Mining in Higher Education" IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp.130-120.
[3] El-Hales-A.(2008),"Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia,Dec 15-18.
[4] Jai Ruby & K. David, "A study model on the impact of various indicators in the performance of students in Higher Education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May-2014, pp.750-755.
[5] Han. J & Kamber. M, "Data mining concepts and techniques", San Francisco, USA, Morgan Kaufmann, 2001.
[6] J. Shana, T. Venkatachalam, "Identifying Key Performance Indicators and Predicting the Result from Student Data." International Journal of Computer Applications (0975 – 8887) Vol.25-No.9 July 2011,
[7] Brijesh Kumar Baradwaj, Saurabh Pal," Mining Educational Data to Analyze Students' Performance" IJACSA, Vol.2, No.6,2011
[8] Ramaswami M., & Bhaskaran R., "CHAID Based Performance Prediction Model in Educational Data Mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, 2010.
[9] Kovacic Z. J., "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference 2010.
[10] Mohammed M. Abu Tair & Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", 2012
[11] Muslihah W., Yuhanim Y., Norshahriah W., Mohd Rizal M., Nor Fatimah A., & Hoo Y. S., 'Predicting NDUM Student's Academic Performance Using Data Mining Techniques', In Proceedings of the Second International Conference on Computer and Electrical Engineering, IEEE computer society, 2009.
[12] Ian H. W. & Eibe F., "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," California: Morgan Kaufmann, 2005
[13] Nguyen N., Paul J., & Peter H., "A Comparative Analysis of Techniques for Predicting Academic Performance". In Proceedings of the 37th ASEE/IEEE Frontiers in Education Conference. pp. 7-12, 2007.
[14] Bengio Y., Buhmann J. M., Embrechts M., & Zurada J. M., "Introduction to the special issue on neural networks for data mining and knowledge discovery," IEEE Trans. Neural Networks, vol. 11, pp. 545-549, 2000.
[15] Vasile P. B., "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment". Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, IEEE, (2007).
[16] M. Wook, Y. H. Yahaya, N. Wahab, M. R. M. Isa, N. F. Awang, and H. Y. Seong, "Prediction NDUM student's academic performance using data mining techniques," presented at the International Conference on Computer and Electrical Engineering, 2009.
[17] Jai Ruby & K. David, "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study", IJRASET International Journal for Research in Applied Science & Engineering Technology, Volume 2 Issue XI, November 2014
[18] Romero, C. and Ventura, S. (2007) "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146.
[19] Kuyoro 'shade o, Nnicolae Goga, Oludele Awodele and Samuel Okolie "Optimal algorithm for predicting students' academic performance" International journal of computers & technology volume 4 no. 1, JAN-FEB, 2013 ISSN 2277-3061
[20] Jai Ruby & K. David,"Analysis of Influencing Factors in Predicting Students Performance Using MLP - A Comparative Study ", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 2, February 2015
[21] Anne F. Maben, 2005, "Chi-square test adapted from Statistics for the Social Sciences".
[22] Fausett, L., "Fundamentals of Neural Networks", Prentice-Hall, Englewood Cliffs, NJ, 1994.