K Nearest Neighbour and Improved Artificial Neural Network Techiniques for Student Academic Performance Prediction

N. Mohamed Farook Ali[1], Dr.M.Elamparithi[2], Dr.V.Anuratha[3]

[1] Research Scholar, Department of Computer Science, Kamalam College of Arts And Science, Anthiyur, Bharathiar University, Coimbatore, Tamilnadu, India.

[2] Associate Professor, Department of Computer Science, Kamalam College of Arts And Science, Anthiyur, Bharathiar University, Coimbatore, Tamilnadu, India.

[3] Associate Professor, Department of Computer Science, Kamalam College of Arts And Science, Anthiyur, Bharathiar University, Coimbatore, Tamilnadu, India.

Abstract: Educational Data Mining has been an emerging topic nowadays due to the growth of educational data. This field makes it possible to develop methods in order to find out hidden patterns from educational data. The methods extracted from Educational Data Mining discipline are then used to understand students including their learning behavior as well as to predict their academic performance. In recent work used artificial neural network to predict academic performance. However in recent work does not used any proper method for missing value imputation and this will affect prediction performance. To mitigate the above mentioned issue in this work introduced an improved framework for student academic performance prediction. In which pre-processing is done by using data formatting, K Nearest Neighbour (KNN) based missing data imputation, min max normalization based data normalization and data filtering methods. Once the data is enhanced using pre-processing, it will be send to the classifier for Student academic performance prediction which is done by using Improved Artificial Neural Network (IANN) method. Experimental results demonstrate the effectiveness of the proposed model interms of precision, recall and f-measure.

Keywords: Educational Data Mining, Student academic performance, Prediction and artificial neural network.

1. INTRODUCTION

Data mining is a concept for analyzing important information of certain data. It helps to extract hidden pattern and to discover relationships between parameters in a huge amount of data. Nowadays, many researchers adopt Data Mining to solve real-world problems in various area such as marketing, telecommunication, health care, medical, industrial and customer relationship. Data mining and machine learning approach can also be used and has been widely used in bioinformatics field[1].

Recently Data Mining has been widely used in educational field. Student's academic performance has become an important part in higher learning institutions. This is because one of the key factor of a high quality learning institutions is based on the record of the students' performance[2]. Students' academic performance prediction is an important concern in the educational field especially education managements. The prediction result could provide an early warning to the students who are at risk by predicting their academic performance. Moreover the prediction result can also be useful in investigating instructor's performance.

The Educational Data Mining can be used to develop a prediction model by exploring educational data and extract hidden pattern for predicting students' academic performance using machine learning techniques [3,4]. Various previous studies regarding prediction of students' academic performance as well as students' learning behaviour are conducted using association rules and clustering methods [5]. However they are producing insufficient results. To overcome those issues in recent work used artificial neural network to predict academic performance. However

in recent work does not used any proper method for missing value imputation and this will affect prediction performance.

To mitigate the above mentioned issues in this work introduced an improved framework for student academic performance prediction. In which pre-processing is done by using data formatting ,K Nearest Neighbour based missing data imputation, min max normalization based data normalization and data filtering methods. Student academic performance prediction is done by using Improved artificial neural network method.

2. Literature review

Burman and Som[2019][6]used multi classifier Support Vector Machine (SVM) to classify the learners in the category of high, average and low according to their academic scores. It is carried out using linear kernel and radial basis kernel. It is noted that RBF produces better results as compared to the linear kernel. Predicting the performance of students in advance can advantage both the institution and learner to take measurable steps in order to enhance the learning process.

Guarín,et al [2015][7]Presented the results of applying an educational data mining approach to model academic attrition (loss of academic status) at the Universidad Nacional de Colombia. Two data mining models were defined to analyze the academic and non-academic data; the models use two classification techniques, naïve Bayes and a decision tree classifier, in order to acquire a better understanding of the attrition during the first enrollments and to assess the quality of the data for the classification task, which can be understood as the prediction of the loss of academic status due to low academic performance. The models aim to predict the attrition in the student's first four enrollments. First, considering any of these periods, and then, at a specific enrollment. Historical academic records and data from the admission process were used to train the models, which were evaluated using cross-validation and previously unseen records from a full academic period. Experimental results show that the prediction of the loss of academic status is improved when the academic data are added.

Mishra,et al [2014][8]used different classification techniques to build performance prediction model based on students' social integration, academic integration, and various emotional skills which have not been considered so far. Two algorithms J48 (Implementation of C4.5) and Random Tree have been applied to the records of MCA students of colleges affiliated to Guru Gobind Singh Indraprastha University to predict third semester performance. Random Tree is found to be more accurate in predicting performance than J48 algorithm.

Miguéis et al [2018][9]proposed a two-stage model, supported by data mining techniques, that uses the information available at the end of the first year of students' academic career (path) to predict their overall academic performance. Unlike most literature on educational data mining, academic success is inferred from both the average grade achieved and the time taken to conclude the degree. Furthermore, this study proposes to segment students based on the dichotomy between the evidence of failure or high performance at the beginning of the degree program, and the students' performance levels predicted by the model. A data set of 2459 students, spanning the years from 2003 to 2015, from a European Engineering School of a public research University, is used to validate the proposed methodology. The empirical results demonstrate the ability of the proposed model to predict the students' performance level with an accuracy above 95%, in an early stage of the students' academic path. It is found that random forests are superior to the other classification techniques that were considered (decision trees, support vector machines, naive Bayes, bagged trees and boosted trees). Together with the prediction model, the suggested segmentation framework represents a useful tool to delineate the optimum strategies to apply, in order to promote higher performance levels and mitigate academic failure, overall increasing the quality of the academic experience provided by a higher education institution.

Hamsa et al [2016][10]developed student's academic performance prediction model, for the Bachelor and Master degree students in Computer Science and Electronics and Communication streams using two selected classification methods; Decision Tree and Fuzzy Genetic Algorithm. Parameters like internal marks, sessional marks and admission score were selected to conduct this work. Internal marks are the combination of attendance marks, average marks obtained from two sessional exams and assignment marks. Admission score for degree students is the weighted score obtained from 10^{th} and 12^{th} examination marks and entrance marks. In the case of master's degree students, it includes degree examination marks and entrance marks. Resultant prediction model can be used to identify student's performance for each subject. Thereby, the lecturers can classify students and

take an early action to improve their performance. Systematic approaches can be taken to improve the performance with time. Due to early prediction and solutions are done, better results can be expected in final exams. Students can view their academic information and updates. Reputed companies having a tie-up with the institution can search students according to their requirements.

3. Proposed methodology

This section discusses the proposed Student academic performance prediction model. Which consist of two phases first one is pre-processing is using data formatting, Nearest Neighbour based Missing Data imputation, min max normalization based data normalization and data filtering methods and second one is Student academic performance prediction by using Improved artificial neural network method. Overall architecture of the proposed model is shown in figure 1.

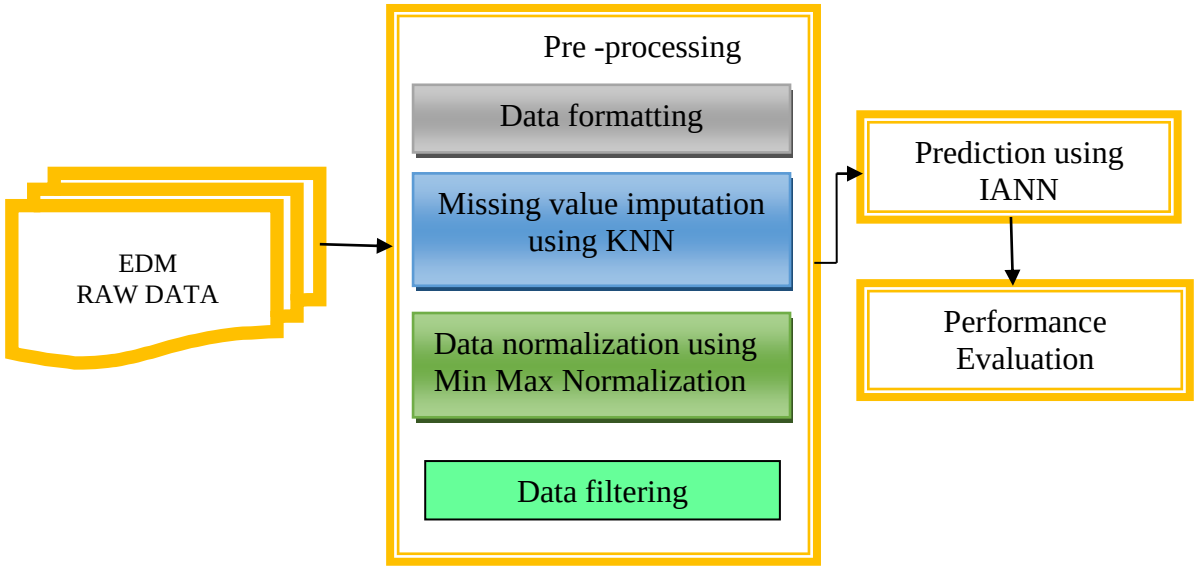

Figure 1.Overall architecture of the proposed model

3.1. Input Dataset

This work uses educational dataset from UCI repository –modified. This dataset contain 230318 Number of Instances, Number of Attributes 13.

3.2. Data formatting

In this step covert raw Educational data into machine understandable data by applying some preprocess steps. This pre-processing tasks is to integrate the data obtained from different sources into one single dataset. Then convert the format of the source data file into the format of destination data file. Have converted our data file into .ARFF format of Google Colab.

3.3. Missing value imputation using KNN

After converting into the machine readable form should impute the missing values before prediction. In general, the KNN imputation is an appropriate choice when have no prior knowledge about the distribution of the data. Given an incomplete instance, this method selects its k closest neighbours according to a distance metric, and estimates missing data with the corresponding mean or mode. The mean rule is used to predict missing numerical features and the mode rule is used to predict missing categorical features [11].

K-Nearest Neighbor Classifier KNN is simple clustering algorithm which impute data based on their similarity with neighbors. K stands for number of data that are considered for the imputation. When a dataset is given, it chooses the k nearest samples from the training data and determines the instance considering the most representative samples for imputation. In this system, Euclidean Distances (ED) are used to find the distances between test sample and database samples. The Euclidean distance between X = (x1, x2, x3, · · · xn) and Y = (y1, y2, y3, · · · yn) is defined as

$$ED(x,y)=\sqrt{\sum_{j=1}^{k}(X_i - Y_i)^2} \quad (1)$$

To effectively apply the KNN imputation approach, a challenging issue is the optimal value of k, and the other is selecting neighbours. The optimal k-value can be selected using only non-missing parts[12]. This k-value estimating procedure considers some elements of the non-missing parts as artificial missing values, and finds an expected k-value that produces the best estimation ability for the artificial missing values. In the proposed approach we determine this parameter optimally using cross validation.

3.4. Min max normalization

After Missing value imputation it needs to normalize their scale .Because input educational data might have scale variations which lead to provide inaccurate results to avoid this issues it is required to normalize the data. This work uses Min-max Normalization model and the process of normalization entails converting numerical values into a new range using a mathematical function. Min-max normalization is one of the most common ways to normalize data. The values in the dataset are normalized within the given range minimum and maximum value from dataset and each value are replaced according to the following formula[13].

$$v' = \frac{v-\min_A}{\max_A-\min_A}\left(\text{new\_max}_A - \text{new\_min}_A\right) + \text{new\_min}_A \quad (2)$$

Where,

A - Attribute data,

Min (A), Max (A) - minimum and maximum absolute value of A respectively

v' - New value of each entry in data

v - Old value of each entry in data

New_ max (A), new_ min (A) - max and min value of the range (i.e boundary value of range required) respectively.

3.4. Data filtering

Have applied the filtering mechanism to extract the data of our interest by using the built in operators in filters. If a column reveals low information according to label attributes then can exclude it from data.

3.5. Student academic performance prediction using Improved weighted Artificial Neural Network

After pre-processing, input data will send to the ANN for prediction. Artificial neural networks (ANN), which is a technique that mimics the properties of biological nervous system and the functions of adaptive biological learning, is used in this work to detect and classify Student academic performance. A neural network is one which consists of an interconnected group of simulated neurons to process and compute information. ANN consists of an input layer, an output layer and one hidden layer [14]. Each layer is formed of nonlinear processing units, termed neurons, and the connections between neurons in successive layers are weighted.

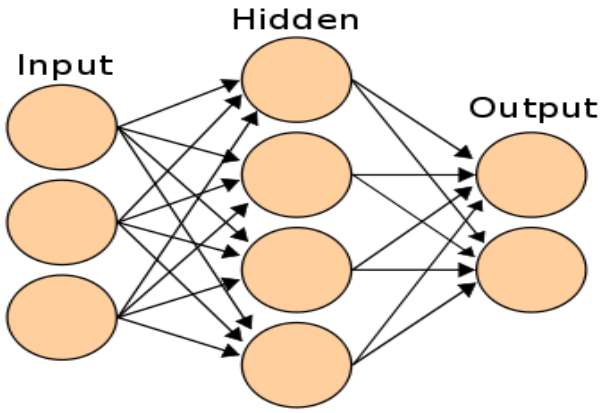

Figure 2.ANN

Non-linear processing is achieved by applying an activation function to the summed inputs to each unit. And the output layer produces the class label as high, low and medium for every input subject. Figure 2.shows the ANN structure. Before training the ANN model the whole dataset is divided into two parts; training & testing. So, in this model, a ratio of 60:40 is used to achieve better results[15]. The hidden layer of a Neural network typically consists of sigmoid function. However this will not perform well with multi label classification. To mitigate the above mentioned issues in this work used tanh activation function. The tanh will squash the range of each neuron between -1 and 1. Such nonlinear elements provide a network with the ability to make soft decisions. The tanh activation function is given by

$$\tanh(i) = \frac{(e^{i})-(e^{-i})}{(e^{i})+(e^{-i})} \quad (3)$$

where i is the sum of the input patterns.

4. 4. RESULTS AND DISCUSSION

This section discusses the experimental results of the proposed model which is implemented using python. Proposed Multiple Metrics – Improved Artificial Neural Networks (MM-IANN) is compared with ANN and Random Tree (RM) interms of precision, recall, accuracy and f-measure for the educational dataset from UCI repository.

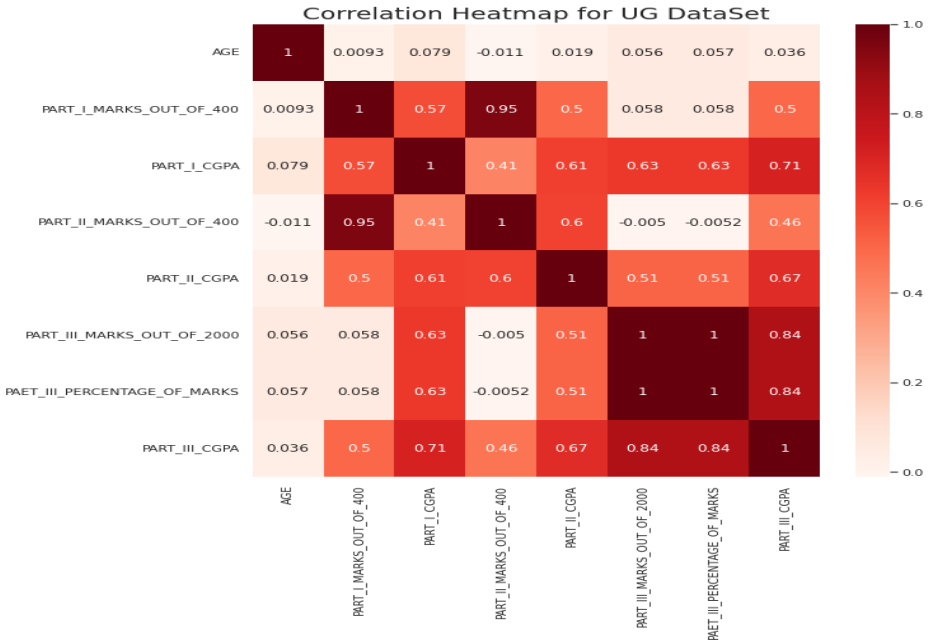

Figure 3.Correleation heat map for UG dataset

Above figure 3.shows the correlation heat map of UG dataset for students age and Marks in all parts. It represents these coefficients to visualize the strength of correlation among variables. It helps find features that are best for Neural network model building. The heat map transforms the correlation matrix into color coding.

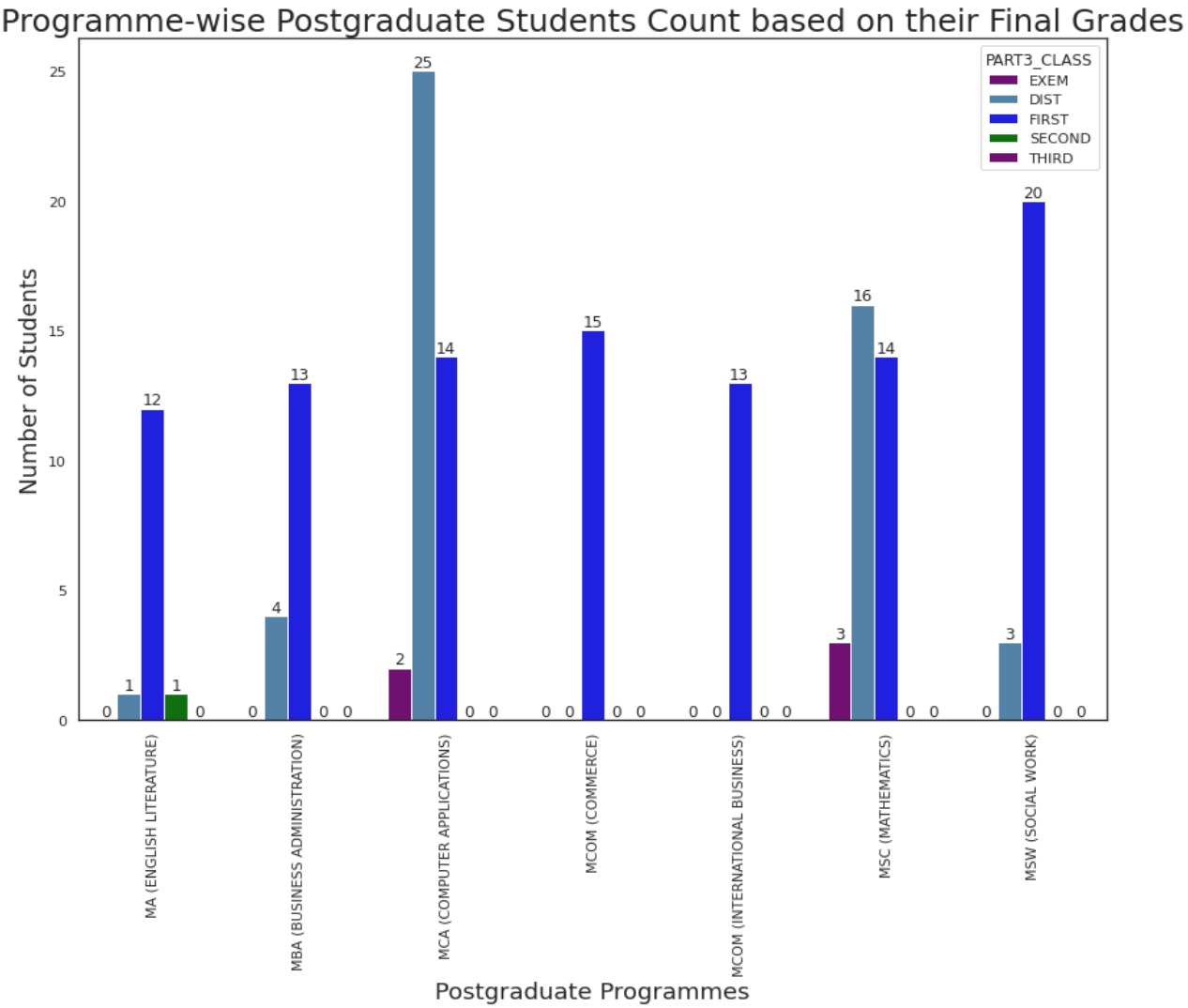

Figure 4.program-wise postgraduate students count based on their final grades

Above figure shows the program wise students count based on their final grades for MA, MBA, MCA, COMMERCE, MCOM, MSC and MSW Programs. From the above figure it is observed numbers of students are high in MCA.

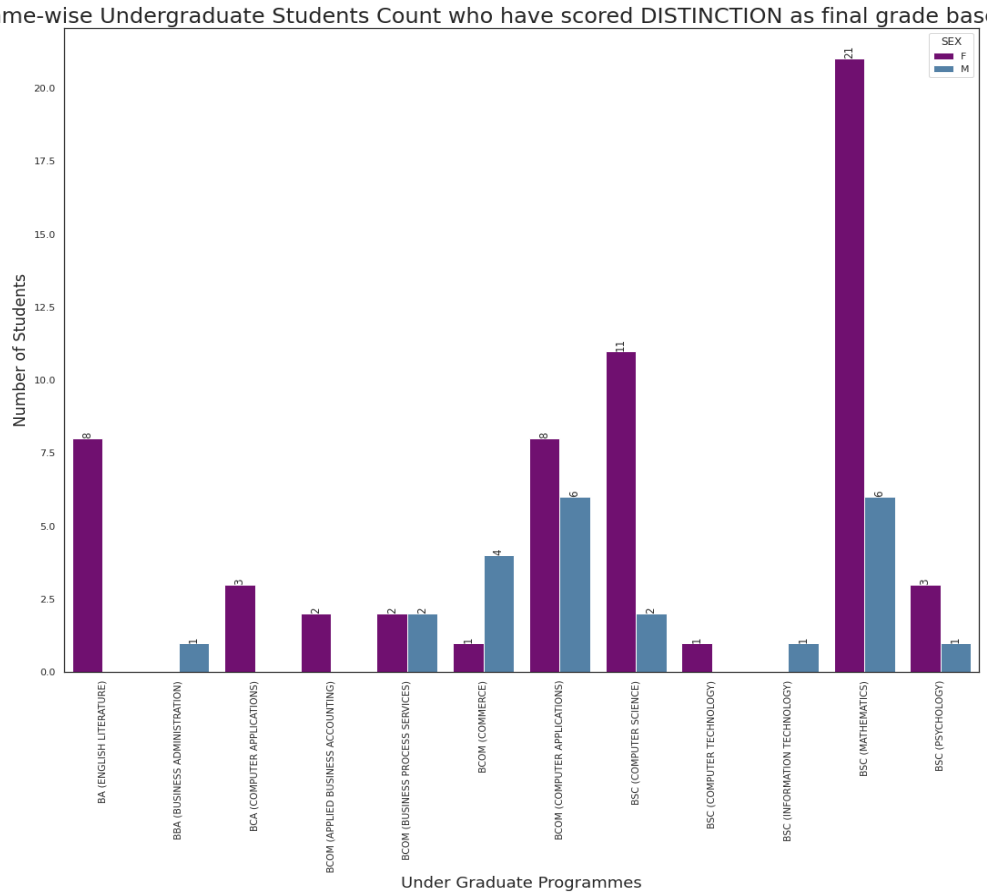

Figure 5.Program-wise undergraduate students count who have scored DISTINCTION as final grade based on their Gender

Program-wise undergraduate students count who have scored DISTINCTION as final grade based on their Gender is shown in the above figure 5. From the above figure it is observed that BSC (Mathematics) has higher number of students who have scored DISTINCTION as final grade than other programs.

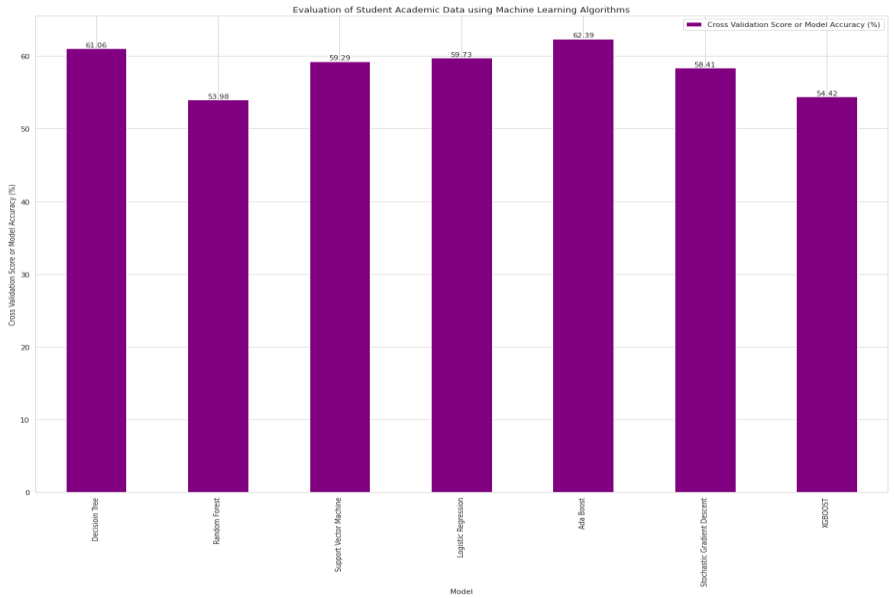

Figure 6.Accuracy results comparison of existing methods

Above figure shows the performance comparison results for the existing Decision tree, Random forest, Support vector machine, Logistic regression, Ada boost, stochastic gradient decent methods interms of accuracy. From the figure it is observed that the Ada boost produces the higher accuracy results than other models. From the figure it is concluded that the Ada boost model produces 62.39% accuracy while the other models such as Decision tree, Random forest, Support vector machine, Logistic regression, Stochastic gradient decent methods produces 61.06%,53.98%,59.29%,59.73%,58.41%,54.42% accordingly.

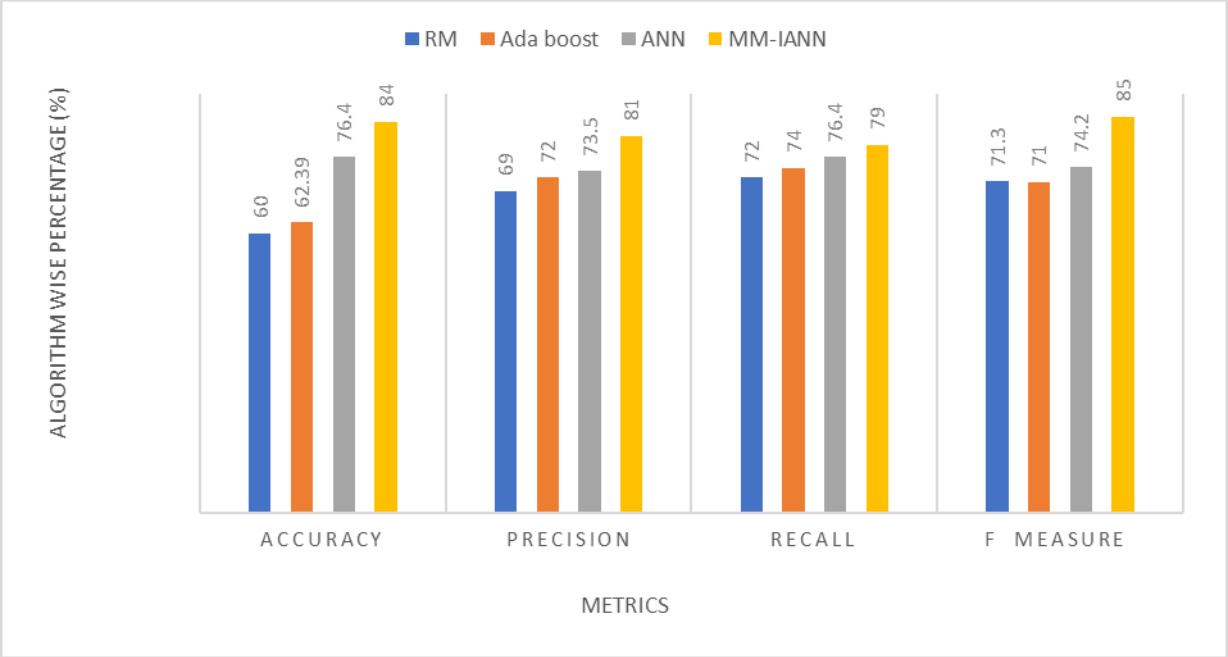

Figure 7.Accuracy, Precision, Recall and F- Measure results comparison

Above figure shows the performance comparison results of the proposed MM-IANN and the existing RM , Ada boost and ANN in terms of Accuracy, Precision, Recall and F- Measure. Proposed system uses min max normalization as pre-processing and its increases the results of proposed MM-IANN model. From the figure it is concluded that the proposed MM-IANN produces higher results interms of Accuracy, Precision, Recall and F-Measure than existing RM, Ada boost and ANN. For example proposed MM-IANN achieves 84% accuracy while the other existing RM, Ada boost and ANN methods achieves 60%, 62.39% and 76.4%.

5. Conclusion and Future Work

Accurately predicting students' future performance based on their ongoing academic records is crucial for effectively carrying out necessary pedagogical interventions to ensure students' on-time and satisfactory graduation. This work aimed to provide an automated model for student's academic performance prediction. This work performs some pre-processing such as data formatting, K Nearest Neighbour (KNN) based missing data imputation, data normalization using Min-Max normalization and data filtering to enhance the input data quality. Finally Student's academic performance prediction is done by using Improved Artificial Neural Network (IANN) method. Experimental results demonstrate that the proposed model produces 84% accuracy which is better than other existing methods. However proposed model consumes more time for computation and it can be enhanced using feature selection to reduce the time consumption and this could be focused in the future research plan.

REFERENCES

[1] Liu, Q., Huang, Z., Yin, Y., Chen, E., Xiong, H., Su, Y. and Hu, G., 2019. Ekt: Exercise-aware knowledge tracing for student performance prediction. IEEE Transactions on Knowledge and Data Engineering, 33(1), pp.100-115.

[2] Xu, J., Moon, K.H. and Van Der Schaar, M., 2017. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11(5),

pp.742-753.

[3] Mishra, T., Kumar, D. and Gupta, S., 2014, February. Mining students' data for prediction performance. In 2014 Fourth International Conference on Advanced Computing & Communication Technologies (pp. 255-262). IEEE.

[4] Guarín, C.E.L., Guzmán, E.L. and González, F.A., 2015. A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje, 10(3), pp.119-125.

[5] Shahiri, A.M. and Husain, W., 2015. A review on predicting student's performance using data mining techniques. Procedia Computer Science, 72, pp.414-422.

[6] Burman, I. and Som, S., 2019, February. Predicting students academic performance using support vector machine. In 2019 Amity international conference on artificial intelligence (AICAI) (pp. 756-759). IEEE.

[7] Guarín, C.E.L., Guzmán, E.L. and González, F.A., 2015. A model to predict low academic performance at a specific enrollment using data mining. IEEE Revista Iberoamericana de tecnologias del Aprendizaje, 10(3), pp.119-125.

[8] Mishra, T., Kumar, D. and Gupta, S., 2014, February. Mining students' data for prediction performance. In 2014 Fourth International Conference on Advanced Computing & Communication Technologies (pp. 255-262). IEEE.

[9] Miguéis, V.L., Freitas, A., Garcia, P.J. and Silva, A., 2018. Early segmentation of students according to their academic performance: A predictive modelling approach. Decision Support Systems, 115, pp.36-51.

[10] Hamsa, H., Indiradevi, S. and Kizhakkethottam, J.J., 2016. Student academic performance prediction model using decision tree and fuzzy genetic algorithm. Procedia Technology, 25, pp.326-332.

[11] Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), pp.1774-1785.

[12] Deng, Z., Zhu, X., Cheng, D., Zong, M. and Zhang, S., 2016. Efficient kNN classification algorithm for big data. Neurocomputing, 195, pp.143-148.

[13] Gumaei, A., Hassan, M.M., Hassan, M.R., Alelaiwi, A. and Fortino, G., 2019. A hybrid feature extraction method with regularized extreme learning machine for brain tumor classification. IEEE Access, 7, pp.36266-36273.

[14] Kumari, C.U., Prasad, S.J. and Mounika, G., 2019, March. Leaf disease detection: feature extraction with K-means clustering and classification with ANN. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1095-1098). IEEE.

[15] Chauhan, N., Isshiki, T. and Li, D., 2019, February. Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS) (pp. 130-133). IEEE.

DOI: https://doi.org/10.15379/ijmst.v10i2.3054