**PREDICTING STUDENTS PERFORMANCE IN EXAMS**

**By Nedhunuri Rohith Reddy**

**INTRODUCTION**

The ultimate goal of any educational institution is offering the best educational experience and knowledge to the students. Identifying the students who need extra support and taking the appropriate actions to enhance their performance plays an important role in achieving that goal. In this research, a machine learning technique has been used to build a model that can predict the performance of the students in his exams by producing the output as grade. The machine learning technique include Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression. This research pays extra attention to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students’ performance. These effects introduced by using features that measure whether the student uses the internet for learning and the time spent on the social networks by the students. The models have been compared using the ROC index performance measure and the classification accuracy. In addition, different measures have been computed such as the classification error, precision, recall, and the F measure. The dataset used to build the models is collected based on a survey given to the students and the students grade book.

The economic success of any country highly depends on making higher education more affordable and that considers one of the main concerns for any government. One of the factors that contributes to the educational expenses is the studying time spent by students in order to graduate. For example, the loan debt of the American students has been increased due to the failure of many students in getting graduated on time. Higher education is provided for free to the students in Iraq by the government. Yet, failing of graduating on time costs the government extra expenses. To avoid these expenses, the government has to ensure that the student graduate on time. Machine learning techniques can be used to forecast the performance of the students and identifying the at risk students as early as possible so appropriate actions can be taken to enhance their performance. One of the most important steps when using these techniques is choosing the attributes or the descriptive features which used as input to the machine learning algorithm. The attributes which I selected for this model is gender, parental level of education, lunch, ethnicity/race, math score, math score, reading score, writing score, percentage. The activities of this research includes the students dataset, data collecting, data preprocessing, creating and evaluating machine learning model using auto AI, and finding the best metric and analyzing the results.

**LITERATURE REVIEW**

**Existing Problem**

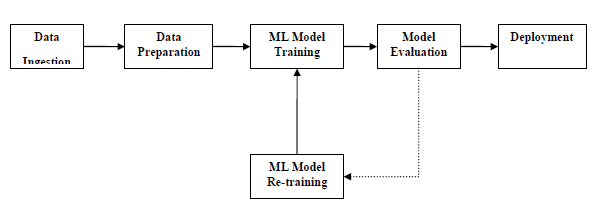
Much research has been done in the area of educational data mining where a predictive model is built to forecast the performance of students to identify the at risk students. This problem can be considered a hard problem because the performance depends on many characteristics related to the students. These characteristics can be categorized into student's GPA and grades, demographics, psychological profile, culture, academic progress, and educational background. The student's GPA is the most important attribute used to predict the performance. The GPA can represent the real value for the future educational and career possibilities and progression. In addition, the academic potentials can be evaluated by the student GPA. The demographics information that consists of the family background, the gender, disability, and age is also considered an important attribute. Till now, many machine learning and data mining techniques have been used to predict the students' performance such as: Artificial Neural Network (ANN); K-Nearest Neighbor (KNN); Support Vector Machine (SVM); Linear Regression; Logistic Regression; Decision Tree (DT); Random Forest (RF); Principal Component Analysis (PCA); Naïve Bayes (NB); Neuro-Fuzzy classification (NF); Decision List (DL); Bayesian Network (BN); and Discriminant Analysis (DA). Table 1 shows a summary of the research papers that relate to this study.

**Proposed Solution**

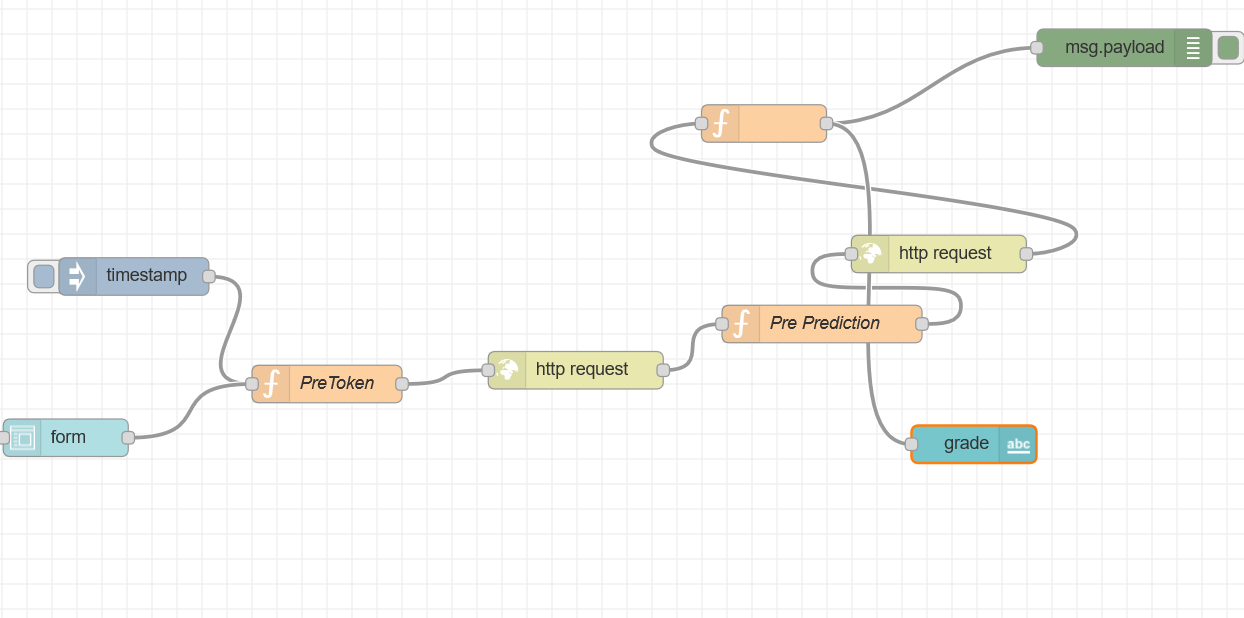
Many models have been build to predict the performance, but no one had used parental level of education as a attribute which I think as a crucial one. The accuracy rate produced by my model is more than 90 percentage. I have used the auto AI to built up my model and the corresponding metric.

**THEORITICAL ANALYSIS**

**Block Diagram**

[](https://www.google.com/url?sa=i&url=https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_pipelines_automatic_workflows.htm&psig=AOvVaw01e81WmUxMAnfo6B9k2ahq&ust=1593072150180000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCOD9kdb-meoCFQAAAAAdAAAAABAD)

**Hardware / Software Designing**

****

**EXPERIMENTAL INVESTIGATIONS**

The dataset used in this research is collected from the Kaggle Website.There are 16 attributes including the dependent variable and 1000 records that is 1000 students information. The dependent variable is grade which needs to be predicted based on the given 15 independent attributes. 70 percentage of the students are good and rest of them are weak. The grades produced is in the alphabetical format that is “A”,”B” ,”C”, “D”,”E”.

In this research, three folds cross validation method has been used. In this method, the dataset is divided into three equal size sets. The learning and testing are executed three times. At each fold or execution, the machine learning algorithm selects one set to be the test set and the remaining two sets as the training sets. The accuracy and the performance measures is aggregated over all the folds in order to calculate the final performance and the final accuracy of the model. The ROC index, the area under the curve, performance measure has been used to evaluate the performance of the classification models. This measure is a well-known measure that is relying on the ROC curve and it is calculated by using the prediction scores. Equation 3 is used to calculate the ROC index. In addition to the ROC index, many important measures have been used such as the accuracy, the classification error, and the F Measure. Equation 4 is used to calculate the F Measure. The F Measure is a useful alternative to the misclassification rate measure.

𝑅𝑂𝐶 𝑖𝑛𝑑𝑒𝑥 = ∑ ( 𝐹𝑃𝑅(𝑇[𝑖])−𝐹𝑃𝑅(𝑇[𝑖 −1]))×(𝑇𝑃𝑅(𝑇[𝑖])+𝑇𝑃𝑅(𝑇[𝑖 −1])) 2 ⁄| 𝑇| 𝑖=2

Where |𝑇| represents the number of thresholds that are used, 𝐹(𝑇[𝑖]) represents the false positive rate at the threshold i, and 𝑇𝑃𝑅(𝑇[𝑖]) represents the true positive rate at the threshold i. A larger ROC index indicates a better classification model. A model with ROC index above 0.7 considered a strong model while a model with ROC index below 0.6 considered a weak model.

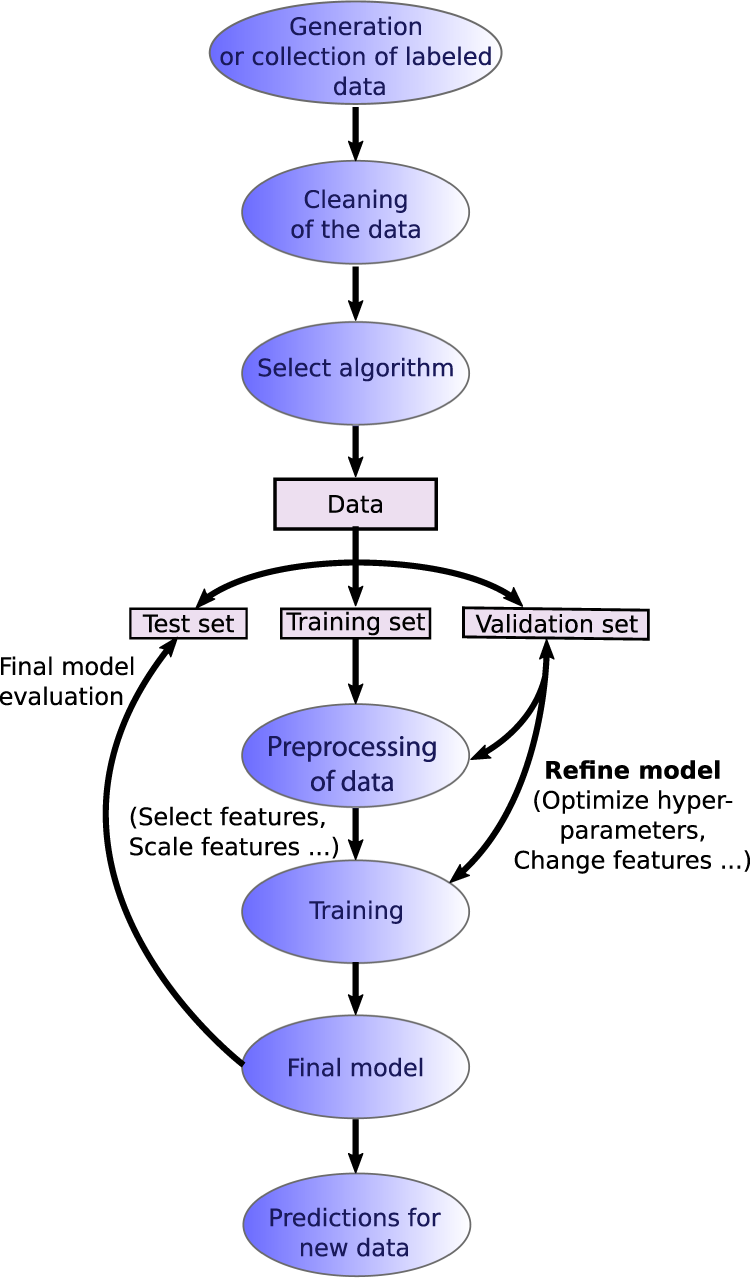
𝐹 𝑀𝑒𝑎𝑠𝑢𝑟𝑒 = 2∗( 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛∗𝑅𝑒𝑐𝑎𝑙𝑙 )/( 𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛+𝑅𝑒𝑐𝑎𝑙𝑙 )

𝑃𝑟𝑒𝑐𝑖𝑠𝑖𝑜𝑛 = 𝑇𝑃 /(𝑇𝑃 +𝐹𝑃)

𝑅𝑒𝑐𝑎𝑙𝑙 = 𝑇𝑃 /(𝑇𝑃 +𝐹𝑁)

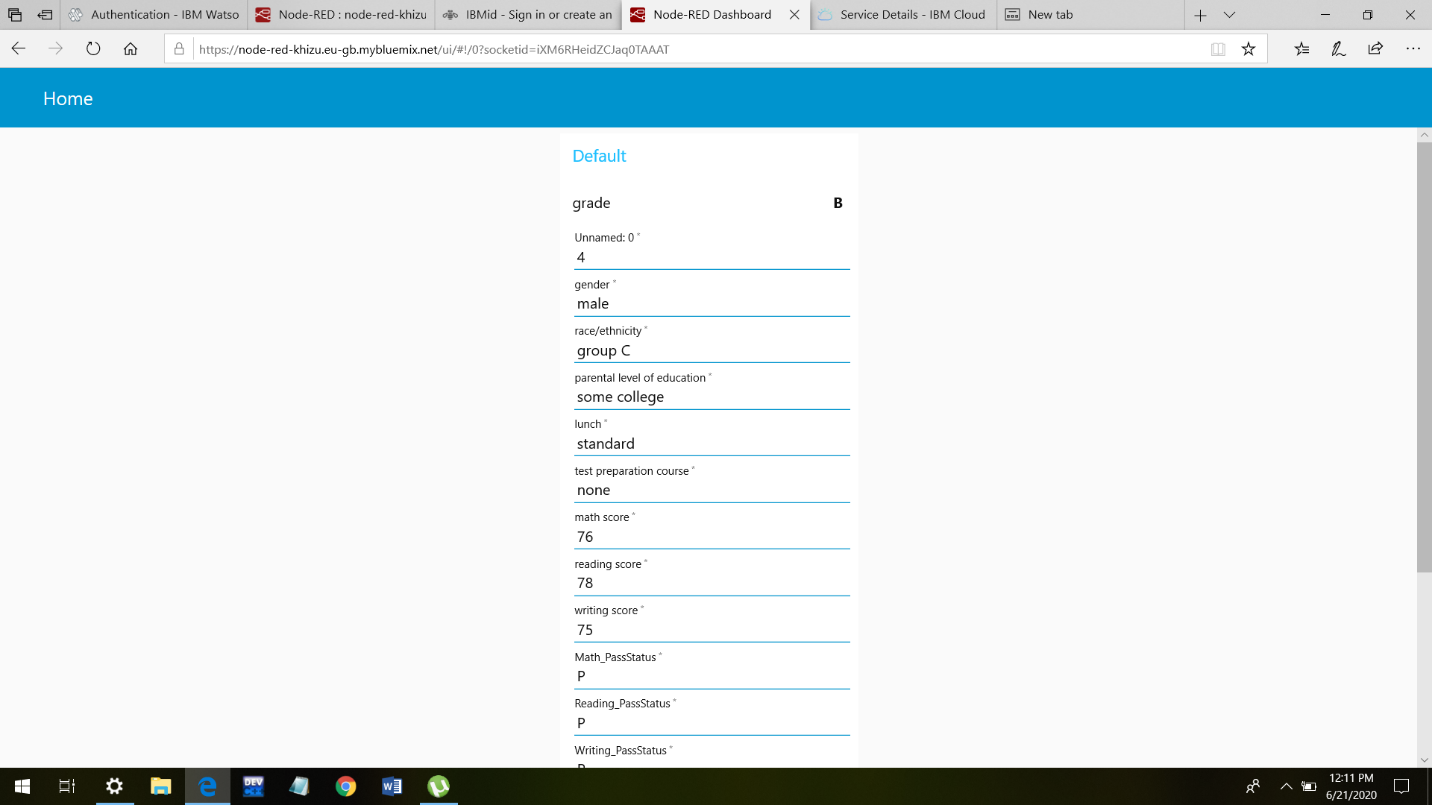
TP, True Positives, is the number of data rows in the test set which had a positive target and that were predicted to have a positive target. TN, True Negatives, is the number of data rows in the test set that had a negative target and that were predicted to have a negative target. FP, False Positives, is the number of data rows in the test set which had a negative target but that were predicted to have a positive target. FN, False Negative, is the number of data rows in the test set that had a positive target but that were predicted to have a negative target.

**FLOW CHART**

[](https://www.google.com/url?sa=i&url=https://www.nature.com/articles/s41524-019-0221-0&psig=AOvVaw0qdfT7mP33gfXaAg3rFuOg&ust=1593071103692000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNi1n-H6meoCFQAAAAAdAAAAABAD)

**RESULT**

The model has been created and tested using the machine learning technique and the output is predicted by deploying the model in a web application built by using Node Red App Service present in IBM Cloud.



**ADVANTAGES & DISADVANTAGES**

**ADVANTAGES**

* The ultimate goal of any educational institution is offering the best educational experience and knowledge to the students which can be achieved by Identifying the students who need extra support and taking the appropriate actions to enhance their performance plays an important role in achieving that goal
* It can easily predict the performance of each student with given inputs.

**DISADVANTAGES**

* Choosing the inputs will be difficult because we can’t understand which inputs plays an major rule.
* The model can’t be prepared with 100 percent accuracy

**APPLICATIONS**

* This model can be deployed into an web application easily. Node Red Service can be used to built the web application’s easily.
* All Education Institutions can use this model to predict their student performances and can improve their student capabilities.

**CONCLUSION**

To solve the problem of identifying the students who have a poor academic performance in the academics, an ML model has been built to predict the performance of the students. An ML model purely built by auto AI model has been developed. The models have been compared to one another using the ROC index performance measure and the classification accuracy. ANN model has the highest ROC index that equals to 0.807 and accuracy of 77.04. In addition, the decision tree model showed that not all the attributes involve in the classification process. Computer Grades-Course1, Accommodation, Interest in studying computer, Educational Environment Satisfaction, and the Residency are the attribute used by the decision tree model.