2019 4th International Conference on Information Systems and Computer Networks (ISCON)
GLA University, Mathura, UP, India. Nov 21-22, 2019

Student Performance Assessment and Prediction System using Machine Learning

Mehil B Shah
Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
shahmehil6@gmail.com

Maheeka Kaistha
Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
kaisthamaheeka@gmail.com

Yogesh Gupta
Computer Science and Engineering
Manipal University Jaipur
Jaipur, India
yogesh.gupta@jaipur.manipal.edu

***Abstract*—Student Performance Analysis System is an emerging field and is very crucial to schools and universities in helping their students and professors. Most of the pre-existing methods are based only on past academic performance of students. This paper aims to develop models which can predict the student's performance and grades while keeping in mind other equally essential personality factors like interests, attributes and opinions (IAO variables) which affect their lifestyle. It uses various machine learning and deep learning techniques to predict the performance of students, and basic exploratory data analysis to derive various correlations of student's performance with psychographic attributes.**

Keywords—Exploratory Data Analysis, Classification Techniques, Multi Class Classification, Boosting Techniques, Neural Networks.

I. INTRODUCTION

Machine learning is a specialisation under the vast artificial intelligence. Machine learning works towards comprehending the complexity of various types of collected data and identifying the correct model for the data by trying several models. This is done for easier interpretation and use by people.

Machine learning lies in the computer science field but is different from basic computing algorithms which are used for problem solving. In the process of machine learning, the algorithms are designed in a way which allows the system or computer to evaluate the data inputs, create training sets and produce the required range specified output using statistical estimation [1].

1. One of the fundamental parts of a student's personal and professional development is performance evaluation.
2. Performance evaluations emphasize students' strong suits and their forte. This acts as an essential tool in augmenting their strengths and distinguishing areas that need improvement.
3. Performance evaluation is an equally essential tool for teachers as well, for helping students in attaining their goals. By being able to analyse the performance of their students, teachers can divert their attention to the necessary areas, advice and guide the students along the correct path and acknowledge and reward their achievements.

II. LITERATURE REVIEW

Literature reviews act as identifiers and evaluators of the documentation of the studies done within a specific field of research. It helps in enumerating the evolution and progress in the given field and summarizes all the previous work done and the latest current knowledge available about the topic.

Pauziah Mohd Arsad, et al. [2] used the method of Artificial Neural Network for predicting student academic performance. Their work used cumulative grade points as the standard of measurement. The results of students from the first semester were taken as the initial input variable making it the independent variable and marks of the eighth semester grade points as the output the dependent variable. The correlation coefficient R and the mean square error were used to measure the performances of the models.

Midhun Mohan M G, et al. [3] uses Learning and Predictive Analytics on huge, expansive data for predicting the performance of students. They collected the data from CBSE schools, using MySQL server and then pre-processed, cleaned it and performed required transformations using the Apache Hive framework.

Madhav S. Vyas, et al. [4] made use of a decision tree model for academic performance prediction. The continuous values were converted to discrete values and the null values eliminated in the collection and pre-processing phase. Further, the decision tree prediction model was used to distinguish the students with poor performance from the students with good performance using the CART algorithm.

Huda Al-Shehri, et al. [5] used a dataset from the University of Minho, Portugal, consisting of 395 data samples for performance in maths. Support Vector Machine and K Nearest Neighbour algorithm were applied on the dataset to predict the student's grade and get a better result compared to the previous work done on that dataset. The final conclusion obtained from the empirical studies was that Support Vector Machine gave a better result with correlation coefficient of 0.96, in comparison to K-Nearest Neighbour which gave a correlation coefficient of 0.95.

978-1-7281-3651-6/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Dalhousie University. Downloaded on July 12,2024 at 03:43:10 UTC from IEEE Xplore. Restrictions apply.

III. DATASET DESCRIPTION

A. Data Collection

We obtained our dataset from UCI Machine Learning Repository [6], of two different schools in Portugal comprising of secondary education. The dataset which was collected using the school reported data and surveys in the form of questionnaires is vast with numerous attributes including grades, social and co-curricular activities, living demographics as well as home and parental information. The datasets are of two subjects: Mathematics and Portuguese language which were designed according to binary three-level classification and regression.

B. Data Description

We used 33 different attributes in this dataset. These attributes are: the school, student's gender, age, living address, size of the family, parental status, the level of education of mother and father, the job status of mother and father, reason to choose school, student's local guardian, daily travel time, weekly time devoted to studies, past class failure records, extra external tutoring, educational support provided by family members, extra paid classes of the given course, co-curricular activities, whether the student attended pre-primary school, interest in higher education, connectivity of the internet at home, romantic status, family relationships, after school self-time availability, frequency of going out, weekday and weekend alcohol consumption, health status, frequency of taking absences from school, first test grade, second test grade and the final grade.

C. Data Preprocessing

Since the data was clean, with no missing rows or attributes, we skipped the data cleaning phase, and started preprocessing the data. In the Data Preprocessing phase, we added the **Grade** column and assigned values to it depending on the **G3** scores, and we converted it into numerical values using Label Encoder. We finalized on **Grade** as our dependent variable and all other variables as the predictor variables.

IV. METHODOLOGY

A. Exploratory Data Analysis

We started our analysis by plotting various graphs on our dataset. We used the Kernel Density Plot which uses a non- parametric method called kernel density estimation, in which the probability density function is calculated of any continuous random variable [7], to detect the correlations between education of the student and parents' education levels.

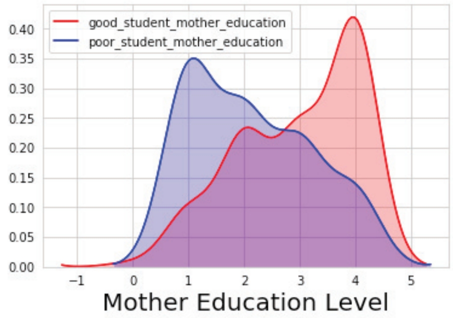

Fig. 1. Kernel Density Plot of students' education level based on mother's education level

From the given kernel density estimation graph, it can be inferred that the mother's level of education plays a significant role in the student's academic performance.

A Box and Whisker plot depicts grouped numerical data based on their quartiles and indicates the degree of dispersion, skewness in the distribution and helps to eliminate the outliers of the data [8]. From Fig. 2, it can be inferred that the students whose frequency of going out is 2 have the maximum median final score. Students going out 3-4 times also perform well but the students who go out only once or never have the least final grade. This led to the inference that that weekly outings should be an optimal amount based on the student's learning and focusing abilities

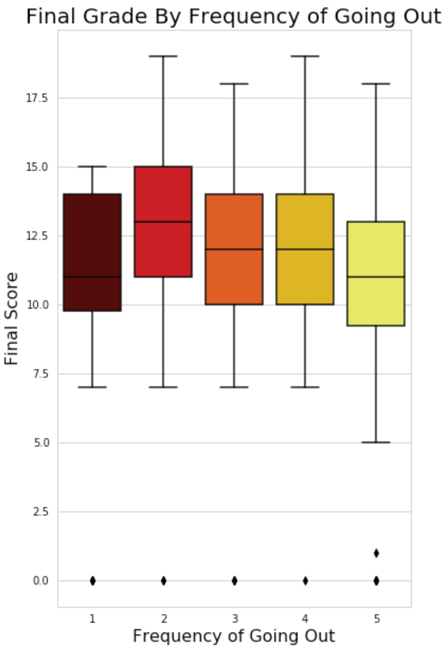

Fig. 2. Box Plot depicting students' education level with the frequency of going out

A Violin plot is a plot showing numeric data along with the probability density distribution at different values. The graph below indicates that the students in age range of 18-19 are more inclined towards higher education and their study time is more compared to others. After the age of 19, it is observed that the median of study time reduces as less number of students are interested in higher education.

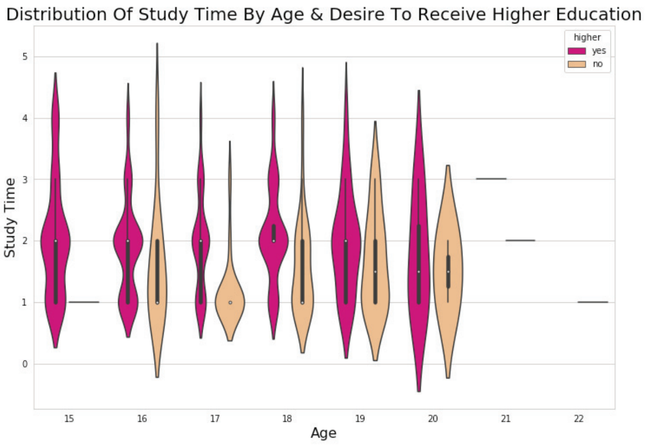

Fig. 3. Violin Plot showing the correlation with study time and age

Authorized licensed use limited to: Dalhousie University. Downloaded on July 12,2024 at 03:43:10 UTC from IEEE Xplore. Restrictions apply.

The grouped histogram below indicates that majority number of students involved in a romantic relationship perform fairly. Whereas the number of students scoring poor and good grades is approximately the same. There is not much difference observed in a particular given grade category of students who are involved in a relationship and those who are not.

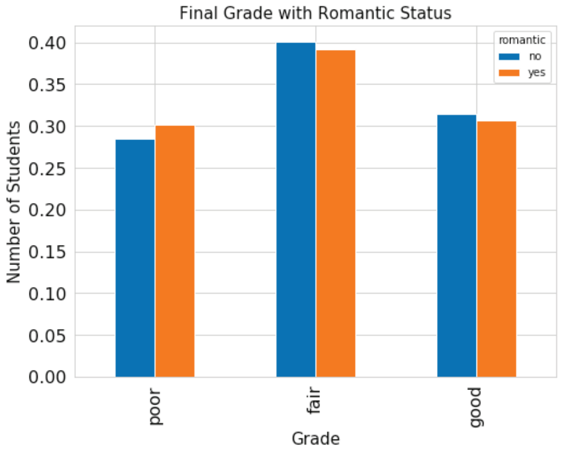

Fig. 4. Correlation with romantic status

From the correlation matrix, we can make certain observations:

1. The grades of future have a very high correlation with the previous grades.
2. Grades have a strong positive correlation with mother's education, study time and desire to receive higher education.
3. Grades have a strong negative correlation with alcohol consumption, past failures and going out.
4. There is a strong positive correlation between mother's education and father's education, daily alcohol consumption, weekend alcohol Consumption and going out.

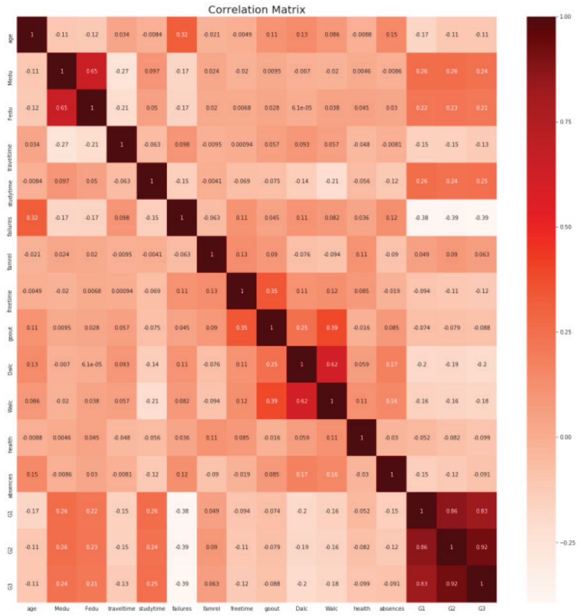

Fig. 5. Correlation Matrix

There were few more correlation which we derived:

1. There is a slight difference in performance of urban-area and rural-area students, with urban-area students outperforming the rural-area students.
2. A grouped histogram of Final Grade based on weekend alcohol consumption indicated that, majority of students drinking more on the weekend score a Fair grade in the final exam.
3. It was observed that majority of students score a Fair grade.

B. *Machine Learning Algorithms*

We implemented four classification Machine Learning Models and further three Machine Learning Boosting Algorithms on our dataset.

Decision Tree Model

Decision Trees use the method of splitting data continuously according to a certain parameter and are of two types of decision trees, "Classification trees and Regression trees". The tree can be explained by decision nodes and leaves. This algorithm helps to identify which attributes are to be considered as the root node at each level, this is called attribute selection. There are different attribute selection measures like Information Gain and Gini Index. Information gain is used for categorical attributes, to estimate the information contained by each attribute.

$$Gain(S,A) = Entropy(S) - \sum_{v\epsilon Values(A)} \frac{|S_v|}{|S|}.Entropy(S_v) \quad (1)$$

$$GiniIndex = 1 - \sum_j p_j^2 \quad (2)$$

Support Vector Model

It is a supervised algorithm which is primarily used to solve classification problems. The objective of this algorithm is to find a hyperplane in an N-dimensional space that empathetically groups the data points. To divide the classes of data points, there are many alternative hyperplanes that can be selected. We need to select the plane that maximizes the margin, the paramount distance between the points. Hyperplanes are decision boundaries that help classify the data points; data points lying at different peripheries of the hyperplane can be assigned to their respective classes. Support vectors are data points that are adjacent to the hyperplane and affect the position and orientation of the hyperplane. Using these, we maximize the margin and removal of these support vectors will have a significant impact on the position of the hyperplane.

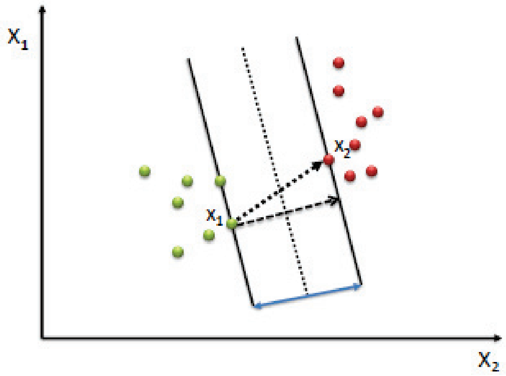

Fig. 6. SVM Formulation

Authorized licensed use limited to: Dalhousie University. Downloaded on July 12,2024 at 03:43:10 UTC from IEEE Xplore. Restrictions apply.

Random Forest Model

Random Forest is simple yet versatile machine learning algorithm that produces exceptional output in majority of the instances and finds a widespread use in various problem statements, owing to its simplicity and its ability to perform both classification and regression. Forest is a collection of Decision Trees, primarily trained by the "bagging" method, in which conjunction of various learning algorithms help in improving accuracy. Instead of probing for the essential features while divaricating the node, it probes for the best feature among a random sub-group of features which produces extensive diversity that leads to a model with great accuracy and minimum error.

$$G(t)=1-\sum_{k=1}^{Q} p^2(k|t) \tag{3}$$

Logistic Regression

Logistic regression is one of the most prominent algorithm and it is majorly used to solve classification problems. It is named after the logistic or the sigmoid function. It is a curve that takes a real-valued number and maps it into a value in the range of 0 and 1, excluding the limits. Logistic regression can be classified as binomial, multinomial or ordinal.

AdaBoost Classifier

Adaptive Boosting or "AdaBoost" used for Binary Classification, incorporates numerous weak learners into solitary well-built learner. When initial "decision stump" is constructed, all observations are considered to be of equal weights. To improve the accuracy and correcting the previous mistakes, the observations that were wrongly classified previously now have more influence than the observations that were correctly classified.

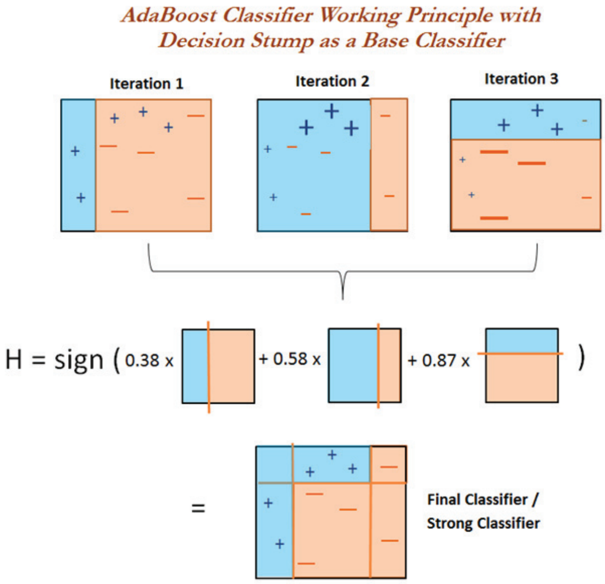
AdaBoost Classifier Working Principle with Decision Stump as a Base Classifier
Iteration 1
Iteration 2
Iteration 3
H = sign (0.38 x
+ 0.58 x
+ 0.87 x
)
=
Final Classifier / Strong Classifier

Fig. 7. Working of Adaboost

XGBoost Classifier

"Extreme Gradient Boosting" or "XGBoost" is the execution of multiple gradient boosted decision trees which is quicker and gives better results. It uses hyper parameters which can be used to fine tune the performance. It cuts off Gradient Boosting when it reaches a certain level, because Gradient Boosting takes a lot of time and XGBoost tends to reduce the computation time by cutting off Gradient Boosting at a certain satisfactory level.

$$L(f_m) \approx \sum_{i=1}^{n}[g_m(x_i)f_m(x_i)+\frac{1}{2}h_m(x_i)f_m(x_i)^2]+const. \tag{4}$$
$$\propto \sum_{j=1}^{T_m}\sum_{i\in R_{jm}}[g_m(x_i)w_{jm}+\frac{1}{2}h_m(x_i)w_{jm}^2].$$

Gradient Boosting Classifier

Gradient Boosting also works similarly to AdaBoost, but instead of modifying the weight values for every wrongly classified observation at every turn, it attempts to fit the newly-built predictor to residual errors made by the predictor constructed earlier. This algorithm uses Gradient Descent to find the errors and faults in the previous learner's predictions.

C. Neural Networks

Neural networks are a set of algorithms that work similar to the human brain, and are designed to recognize patterns. They act as a clustering and classification layer on top of the data and help to cluster the unlabelled data according to similitude among the example inputs, and they classify data when they have labelled dataset to train on. Deep neural networks or "Deep Learning" can basically be thought of as a component of larger machine-learning applications involving algorithms for reinforcement learning, classification and regression.

ANN operates on a Hidden State. These hidden states are alike the neurons. Each of the hidden state is a temporary form which has a probabilistic behaviour. A grid of such hidden state functions as a link between the input states which brings the initial data into the system for further processing by subsequent layers of artificial neurons and the output state [9].

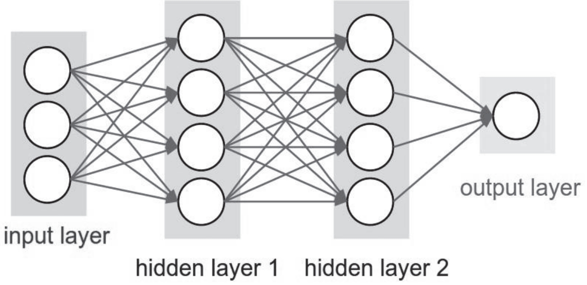

Fig. 8. Architecture of Multi-layer ANN

ANN uses the activation functions of a node which determines the output of that node given the inputs. Sigmoid is one particular kind of activation function having a typical sigmoid curve which is "S" in shape. It is regarded as a unique case of the logistic function which produces a set of probability outputs between 0 and 1 when fed with inputs. [10].

$$Sigmoid(x) = \frac{1}{1+e^{(-x)}} = \frac{e^x}{1+e^x} \tag{5}$$

Authorized licensed use limited to: Dalhousie University. Downloaded on July 12,2024 at 03:43:10 UTC from IEEE Xplore. Restrictions apply.

RNN or Recurrent Neural Networks are a potent and resilient type of neural networks which have an ability to recall key characteristics of the input they received because of their internal memory, which permits them to predict the following outputs. They produce predictive results on sequential data, a feature which other algorithms lack. In a RNN, the information flows circularly through a loop, so when a decision is taken, the current input and learnings from the previous inputs are taken into consideration.

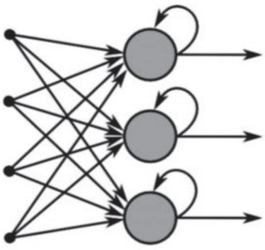

Fig. 9. Information flow in RNN

RNN functions as a chain or series of Neural Networks that are successively trained with backpropagation. Backpropagation is essentially moving backwards through the neural network to learn the partial derivatives of the error vis-a-vis the weights, allowing them to deduct this value from the weights. Those derivatives are then manoeuvred by Gradient Descent, an algorithm which recapitulates minimizing the given function. Then it alters the weights, in order to prune the error [11].

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t) \tag{6}$$

$$y_t = W_{hy}h_t$$

V. Results

Table 1 tabulates the accuracy of all the used models in this paper. This table shows that Gradient Boosting algorithm enhances the accuracy up to 93.8%, which is best in all results. Boosting algorithm is performing better because it uses hyper parameters which enhances the performance by fine tune. On the other hand, logistic regression maps the values between 0 and 1 but never exactly at those limits and it leads to imprecision.

TABLE I. Accuracies of models

Models	Accuracy
Decision Tree	88.2%
SVM	89.8%
Random Forest	91.7%
Logistic Regression	59.8%
Gradient Boosting	93.80%
XGBoosting	88.21%
AdaBoosting	85.13%
ANN	88.89%
RNN	69.65%

VI. Conclusion

Our plan for the future is initially to collect more data and training the models on that data. By feeding more data, we can improve the accuracy of Neural Networks and delve more into the Deep Learning side of our project.

Next, we are going to work towards making an end-to-end platform solution for colleges. This would not only help in predicting the performance of students, but it would also help in identifying their weak points.

Our research would also work as a useful tool for teachers, as predicting the students' performance will help them to take necessary actions and help to determine the sections of class where the teachers have to pay extra attention.

References

[1] https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning, accessed at 14.05.2019.

[2] Pauziah Mohd Arsad, Norlida Buniyamin and Jamalul-lail Ab Manan. "A Neural Network Students' Performance Prediction Model (NNSPPM)" 2013 IEEE International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA), Malaysia.

[3] Midhun Mohan M G, Siju K Augustin and Dr. Kumari Roshni V S "A BigData Approach for Classification and Prediction of Student Result Using MapReduce", 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), India, 2015, pp. 145-150.

[4] Madhav S. Vyas and Reshma Gulwani. "Predicting Student's Performance using CART approach in Data Science", 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), India, 2017, pp. 58-61.

[5] Huda Al-Shehri, Amani Al-Qarni, Leena Al-Saati, Arwa Batoaq, Haifa Badukhen, Saleh Alrashed, Jamal Alhiyafi and Sunday O. Olatunji. "Student Performance Prediction Using Support Vector Machine and K-Nearest Neighbors", 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Canada, 2017.

[6] https://archive.ics.uci.edu/ml/datasets/student+performance, accessed at 14.05.2019.

[7] https://chemicalstatistician.wordpress.com/2013/06/09/exploratory-data-analysis-kernel-density-estimation-in-r-on-ozone-pollution-data-in-new-york-and-ozonopolis/, accessed at 14.05.2019.

[8] https://en.wikipedia.org/wiki/Box_plot, accessed at 14.05.2019.

[9] https://www.analyticsvidhya.com/blog/2014/10/introduction-neural-network-simplified/, accessed at 14.05.2019.

[10] https://blog.goodaudience.com/artificial-neural-networks-explained-436fcf36e75, accessed at 14.05.2019.

[11] https://towardsdatascience.com/recurrent-neural-networks-and-lstm-4b601dd822a5, accessed at 14.05.2019.

Authorized licensed use limited to: Dalhousie University. Downloaded on July 12,2024 at 03:43:10 UTC from IEEE Xplore. Restrictions apply.