Volume 6, Issue 4, October 2019 ISSN: 2348-8565 (Online)

International Journal of Modern Engineering and Research Technology

Website: http://www.ijmert.org Email: editor.ijmert@gmail.com

Analysis of Student Performance Factors and Prediction using Machine Learning Techniques

Vamshi Paili
Computer Science & Engineering,
CVR College of Engineering,
Ibrahimpatnam, (T.S.) India
Email: vamshipaili@outlook.com

CH Vishal Reddy
Computer Science & Engineering,
CVR College of Engineering,
Ibrahimpatnam, (T.S.) India
Email: reddyvishal0@gmail.com

A. Venkata Krishna
Computer Science & Engineering,
CVR College of Engineering,
Ibrahimpatnam, (T.S.) India
Email: venkatabburi@outlook.com

ABSTRACT

Higher education institutions are often interested in whether students will be fruitful or not during their study. Academic institutions attempt to gauge the rate of successful students utilizing a few strategies such as physical examinations, statistical methods and recently data mining techniques. Right now in India, the need of existing framework to examine and monitor student advance and performance at an earlier stage is not being addressed. The interrelationship between factors and components for predicting performance take an interest in complicated nonlinear ways. The prime objective is to analyze students' performance to decide a correlation between the performance of students and past grades, demographic, social and school related features. In this study we create a classification demonstrates to foresee understudy execution utilizing Machine Learning which naturally learns numerous levels of representation. We train model on a moderately expansive real world students' dataset, and the exploratory comes about appear the viability of the proposed strategy which can be connected into academic pre-warning mechanism.

Keywords:— *Logistic Regression, Exploratory Data Analysis, UCI Student Dataset, Feature selection, Support Vector Machine*

I. INTRODUCTION

The ability to predict a student's performance could be useful in a great number of different ways associated with university-level distance learning. Students' key demographic characteristics and their marks on a few written assignments can constitute the training set for a supervised machine learning algorithm. The learning algorithm could then be able to predict the performance of new students, thus becoming a useful tool for identifying predicted poor performers. With the wide utilization of computers and web, there has recently been a gigantic increment in freely accessible information that can be analysed. Such a expansive sum of information show both a issue and an opportunity. The issue is that it is troublesome for people to dissect such expansive data. The concept of machine learning is something born out of this environment. The prime objective is to analyze students' performance to decide a correlation between the performance of

students and past grades, demographic, social and school related features and predicting whether the student likely to pass or fail using machine learning algorithms – Decision tree, Random forest classifier, Logistic Regression, SVC and ADA Boost.

2.1 Data Preparation

Data is the main and crucial part in machine learning. The data is collected and prepared accordingly. Then a model is chosen and this data is used for training and testing the model. Keen evaluation is done for the selected model and parameter tuning is done if required and finally the model is ready for the predicting the real time data.

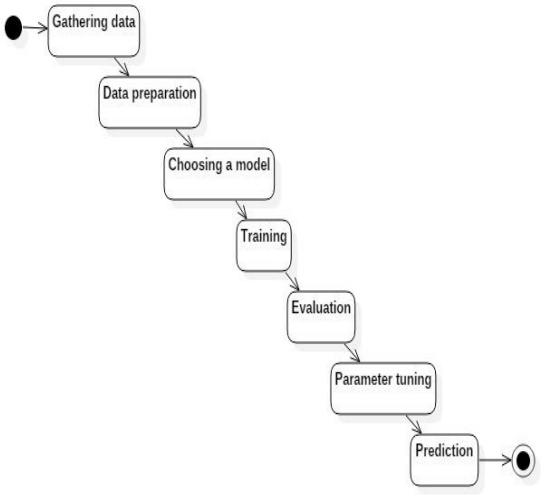

Figure 1: Activity Flow for Performance Prediction

Two datasets are being used regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). Students are classified into three categories, “good”, “fair”, and “poor”, according to their final exam performance. At that point the data is analyzed on few features that have critical impact on students' final performance, including Romantic Status, Alcohol Consumption, Parents Education Level, Frequency of going out, Desire of higher education and living area. At last, leveraging available features, created different machine learning models to foresee students' last performance classification and have compared models performance based on one-out test precision score.

2.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

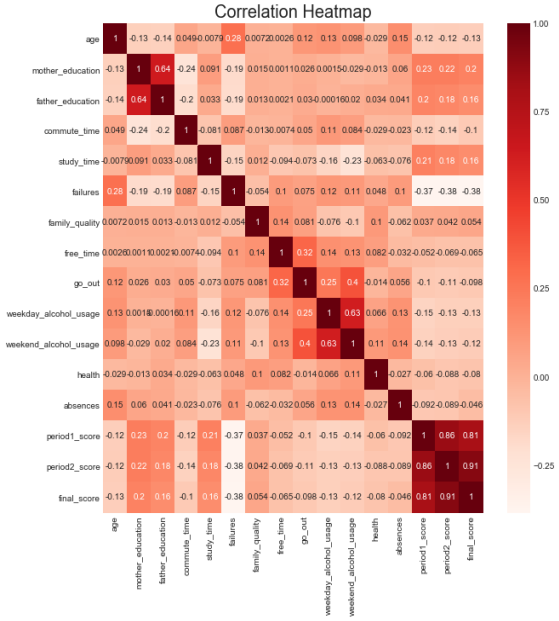

Figure 2: Correlation Heatmap

If the dataset constitutes large number of columns it's arduous to establish the relation between them, a good way to quickly check correlations among columns is by visualizing the correlation matrix as a heatmap in Figure 2.

KDE Plot described as Kernel Density Estimate is used for visualizing the Probability Density of a continuous variable. It depicts the probability density at different values in a continuous variable. This type of plot can help to quickly identify the most correlated variables like

period1_score, period2_score, mother_education and father_education. The influence of parents education on the student performance is analysed by plotting performance against good_student parent_education, poor_student_parent_ education for both mother and father.

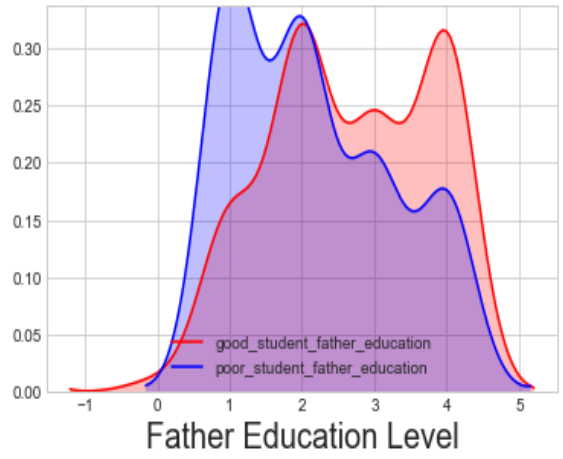

Figure 3: KDE plot for Father Education

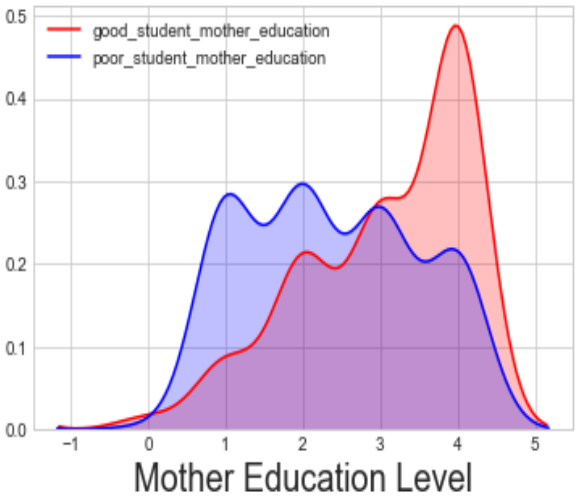

Figure 4: KDE plot for Mother Education

Here the ordinary least squares (OLS) statistical graph tells that parents' education level has a positive correlation with students' final score. Comparatively, mother's education level has bigger influence than father's education level. Furthermore, analysing the impact of the parents' education it is determined that the higher the education levels of the parent more probable high scores for the students. Specifically, it is observed that mothers with healthcare profession and fathers with teacher profession likely to have the student with high scores.

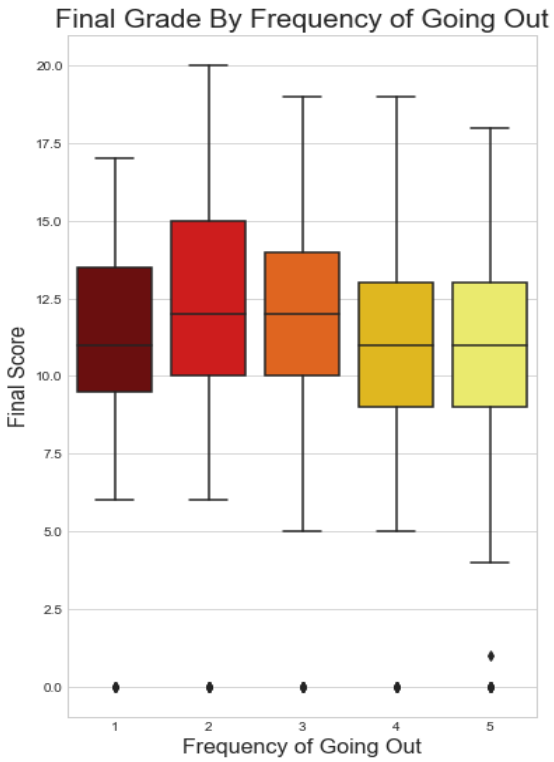

Figure 5: Boxplot for Frequency of going out

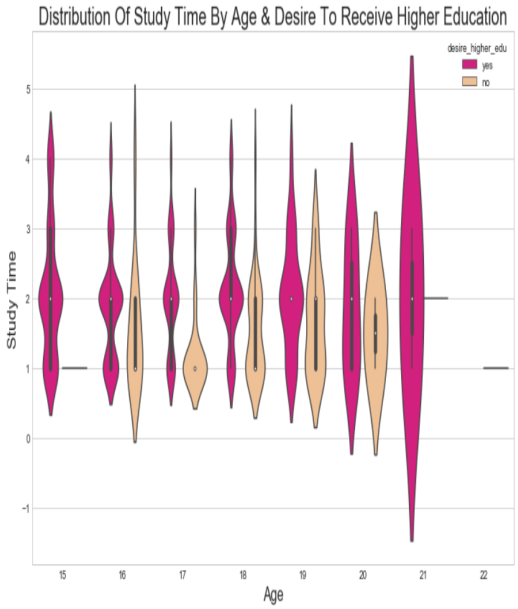

Figure 6: Violin plot for study time by age and desire for Higher education

Hypothesis Testing confirmed, the frequency of going out with friends has a significant correlation with students' final performance. The boxplot demonstrates the frequency of going out against final score. Further, distribution of study time by age & desire to receive higher education depicted in the violinplot illustrates significant correlation with students' final performance.

2.3 Classification

Logistic Regression

It is the Supervised Machine Learning classification used to predict binary outcomes for a given set of independent variables. The dependent variable's outcome is discrete value.

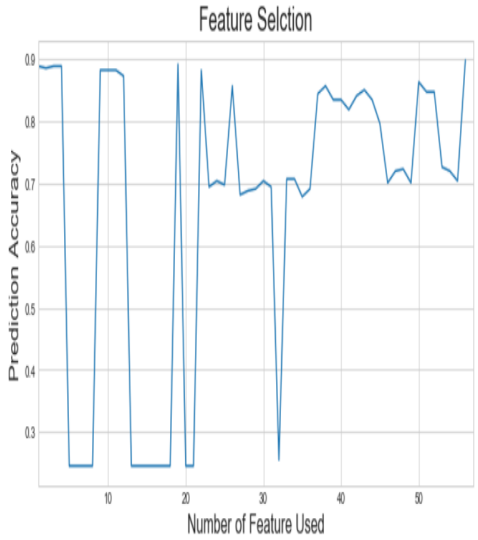

Figure 7: Feature selection

Feature selection is the process of choosing variables that are useful in predicting the response. Having irrelevant features in the data can decrease the accuracy of many models, especially linear algorithms like logistic regression in this case. Three benefits of performing feature selection before modelling the data includes improved accuracy, reduced over fitting and training time.

The number of optimal features to be used is determined thereby increasing the model accuracy and cross validation score is determined.

III. CONCLUSIONS

The logistic regression model provides a better accuracy with the score of 90%. By tuning the parameters through superior selection of features the model precision has increased. The valedictorian of the college is likely to have this profile:

- Doesn't go with friends frequently.
- Having strong desire for higher education.
- Parents both received higher education.
- Study more than ten hours weekly.

REFERENCES:

[1] Hany M. Harb and Malaka A. Moustafa, "Selecting Optimal Subset of Features of Student Performance Model", IJCSI International Journal of Computer Science Issue, Vol. 9, Issue 5, No, September 2012, pp. 253-262

[2] CarlosMarquez, Cristobal Romero Morales and Sebastian Ventura Soto "Predicting School Failure and Dropout by Using Data Mining Techniques" IEEE Journal of Latin-American Learning Technologies, Vol. 8, No. 1, February, 2013, pp. 7-14.

[3] Kiri Wagstaff and Claire Cardie "Constrained K-means Clustering with Background Knowledge" Proceedings of eighteenth international conference on machine learning, 2001, pp. 577-584.

[4] Grigorios F. Tzortzis and Aristidis C. Likas, Senior Member, IEEE "The

Global Kernel K-Means Algorithm for Clustering in Feature Space" IEEE transactions on neural networks, Vol. 20, No. 7, July 2009, pp. 1181-1194

[5] K.A Abdul Nazeer and M.P Singh "Improving the accuracy and efficiency of k means, kohenon self organizing map and hierarchical agglomerative clustering". Proceedings of world congress on engineering. Volume 1, London u.k, (2002).

[6] Saadat Naziova "Survey on Spam Filtering Techniques", Communication and Network, August 2011, pp. 153-160.

[7] P. Moniza and P. Asha "An Assortment of Spam Detection System", International Conference on Computing, Electronics and Electrical Technologies [ICCEET] 2012, pp.77-83

[8] Patricia Bellin Ribeiro, Luis Alexandre da Silva and Kelton Augusto Pontara da Costa "Spam Intrusion Detection in Computer Networks Using Intelligent Techniques", IFIP IEEE IM Workshop: 1st International Workshop on security for Emerging Distributed Network Technologies (DISSECT), 2015, pp. 304-311.

[9] Yen-Liang Chen, Hsiao-Wei Hu and Kwei Tang, "A Novel Decision-Tree Method for Structured Continuous-Label Classification" IEEE Transactions on Cybernetics, 2013, pp. 1734–1746.

[10] Qiang Yang, Senior Member, IEEE, Jie Yin, Charles Ling, and Rong Pan, "Extracting Actionable Knowledge from Decision Trees" IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 1, January 2007.

* * * * *