Analysis of Student Result using Machine Learning in Python

DR. BONTHU KOTAIAH
ASSITANT PROFESSOR, Dept. of CS and IT, MAULANA AZAD NATIONAL URDU UNIVERSITY,
GACHIBOWLI, HYDERABAD, TELANGANA
EMAIL: KOTAIAH.BONTHU@MANUU.EDU.IN

DR. SYED MOHD FAZAL UL HAQUE
ASSITANT PROFESSOR, Dept. of CSE, POLYTECHNIC, MAULANA AZAD NATIONAL URDU UNIVERSITY,
GACHIBOWLI, HYDERABAD, TELANGANA
EMAIL: FAZAL@MANUU.EDU.IN

ABSTRACT

The higher education system suffers from various drawbacks and the main reason behind this is difficulty in its evaluation and hindrances in improvement opportunities. analysis of Institutional data is a new evolving concept which involves the development of different approaches to examining the novel type of dataset coming from the institutional environment and utilizing those approaches to recognize the students in a more appropriate way. In order to reach this objective, the organizations are required to acquire in-depth knowledge to adequately perform the task of assessing, forecasting, planning, evaluating, and decision making. A major part of the information can be achieved through the organization’s old records and databases. Data analyzing is simply a method of digging out unknown or in-depth knowledge from a set of a huge dataset. This paper is basically produced to highlight the role of data analysis or mining and its implications in the performance of higher education.

Keywords and terms: machine learning, student performance, regression, decision trees naïve Bayes Classification

.

1. INTRODUCTION

With the extensive use of internet and computers, there has freshly been a large improvement in the openly accessible dataset that can be explained. Be its website traffic, user habits, or online sales data, data is created every second. Such a huge quantity of information available both an opportunity and a difficulty. The difficulty is, it is tough for people to interpret such massive dataset. The opportunity is that this kind of information is perfect for computers to processing, because it is saved digitally in a suitably formatted manner, and computers can treat dataset extremely faster than individuals.

The particular paper is basically focused on education. The purpose of this analysis is to predict student results based on their performance.

The model predicts the result of student's performance based on other properties which are created by using student's dataset. First, the training dataset is used for input data. The data sets, including various types of information. This data set is in form of the table, each variable, or column, includes specific information data about a scholar, such as a gender, health, family background, age, or study time, sports, free time and internet access and where each row represents a student. The algorithm generates a model, the function is that outcomes failure or success of the student, utilizing other attributes as input.

This research estimates the effect of various machine learning methods and algorithms. Algorithms are applied in generating predictive rules is various, this research concentrates on three of them, which are decision trees, linear regression, and K-neighbour regression. This analysis also estimates the increase caused by feature engineering, which leads to transforming the dataset to present it more fitting for machine learning.

2. RELATED WORK

[1] (E.Venkatesan 2017) proposed a method using classification and clustering algorithm and said that no one can say that a particular algorithm is the best algorithm for the prediction purposes using any kind of real-world data set for some applications. Merely, it is possible, suggesting the public presentation of algorithms for the selected date. Grounded on this notion, the execution of clustering algorithms, named as k-Means and EM were compared. The outcomes were broken down by several executions of the plans. Normally, student performance dataset clustering best clustering, k-means algorithm.

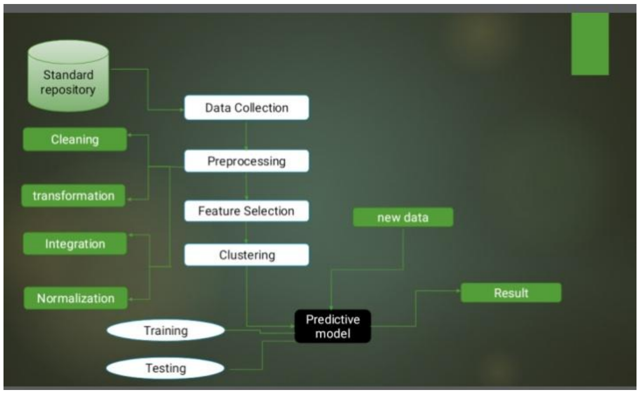
[2] (CH.M.H.Sai Baba 2017) proposed a classification method and said accordingly we have to increase the number of attributes taken because we can not predict the result of the student just by his previous year's marks. We have to take a big amount of data set in order to get accurate results. By considering all the attributes related to the student and later applying the classification techniques then we can predict even the student will be placed or not

.[3] (Mrs. M.S. Mythili 2014) proposed a classification method and said that the work examines the power of algorithms of machine learning in choosing the impact of the result, gender, parental education, locality, and economy within the education system and analysis performance of high school students. It is found that the Random Forest performed better than that of another algorithm applied in the student's data information. This analysis is conforming to be extremely useful for the educational system. In the future, it is committable to enhance the interpretation by utilizing various association rule mining and clustering methods for the information dataset of students.

[4] (Havan Agrawal 2015) proposed a machine learning method and he said that Present education point that educational achievements of the scholars fundamentally depend on their prior achievements. This research validates that previous achievements have surely got vital control over the appearance of students. More, we established that the representation of neural networks improves with an addition in the size of the dataset.

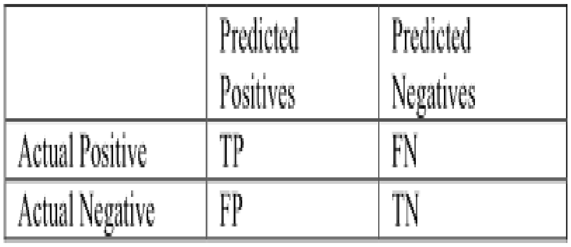
Machine learning has become far away from its crescent phase and can show to be an important intermediary in the educational system. In future, applications identical to the one advanced, as well as any alterations thereof may become a combined portion of every educational system.

3. Proposed Methodology



4. EVALUATION METRIX:

In order to estimate the performance of the model, foretold values need to be compared with the exact values. There are various standards for evaluating the accuracy of the prediction system.



	Predicted Positives	Predicted Negatives
Actual Positive	TP	FN
Actual Negative	FP	TN

Fig1: confusion matrix

Precision = TP/(FP+TP)
Accuracy=(TN+TP)/(TN+TP+FN+FP)
Recall = TP/(FN+TP)

Recall and precision are generally applied collectively to obtain a reliable computation. The main purpose is that correctly predicting the favorable outputs is not sufficient. A stable predictive system must have a genuine mixture of successful favorable outcomes and successful unfavorable outcomes.

Where, TN, FP, FN, and TP designate sequentially to the no of true negative occurrences, the no. of false positive instances, to the no. of false negative cases and the no. of true positive occurrences.

5. IMPLEMENTATION ,AND RESULT ANALYSIS.

The purpose of the analysis was to analyze various machine learning algorithms in the student results in the performance prediction system. The prediction system was designed using Python language. It is a programming language generally applied for machine learning. It has developed functions for the three chosen methods for this analysis, which are decision tree, linear regression, and K-neighbor classification. It also generates the required outcomes for estimating and improving the outcomes of predictions. The code is written in Python.The data set carries knowledge regarding students and have 31 several parameters.

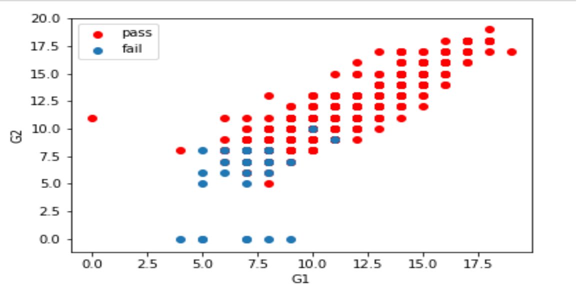


Figure 5.1 relationship between g1 and g2

As figure shows based on G1 and G2 marks a student will perform better and score higher marks if he scored G1 and G2 more than 10 marks. As we can see the result the figure shows positive incrissing relation with G1 and G2. A student will always pass if he score marks more 12 in both G1 and G2

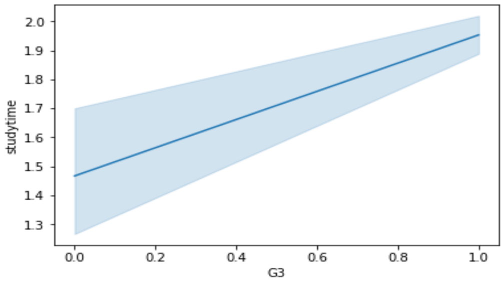


Figure 5.2 relationship between study time and g2

The study time and G3 shows incrissing order graph. If a student study more houre definitly he will score higher marks. This also represent postive incrissing relation with each other

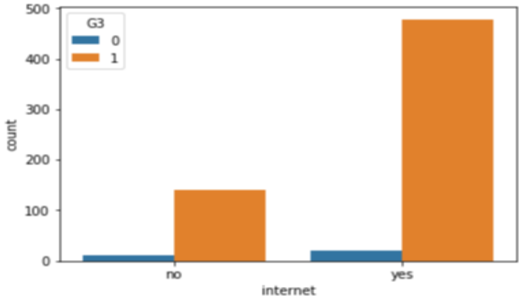


Figure 5.3. bar graph using internet attribute

In the figure no , By comparing usage of internet access we came to the conclusion if student use internet more time then he will score higher marks. It may be possible that student use internet as a study platform. Not for the other purpose.

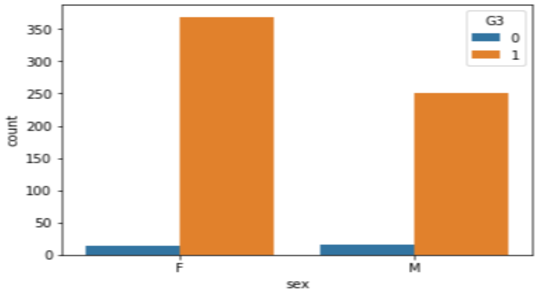


Figure 5.4. bar graph using male and female attribute

The figure shows that female student performance is better that male student. Female student scores higher marks tha male student hence chances of getting passed of female student is higher.

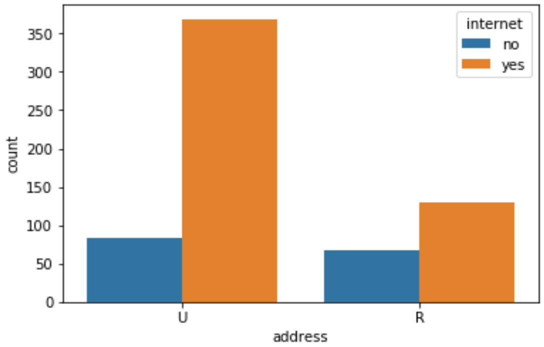


Figure 5.5. bar graph using address attribute

By comparing addresses of urban and rural with respect to internet usage. The urban student is performing better because it may possible that urban student has more facility to use internet hence still the rural area is lack of internet connection and above figure no 5 , we saw that if student use internet more then he score higher marks. So here also we may came to the conclusion that urban student will perform better.

6. DATA SETS

The data set contains 395 students. In the dataset, the test and training datasets were generated, and models were developed applying the machine learning techniques. Then, models were examined by using test dataset and confusion matrix was created as the appropriate results. Figure 1 can be exhibit the confusion matrix for the decision tree, naïve Bayes classifier, and linear regression algorithms.

7. CONCLUSION AND FUTURE SCOPE

The benefit of machine learning in foretelling student achievement relies on the reliable usage of dataset and machine learning techniques. Picking the best learning technique for accurate queries is important to obtain excellent outcomes. Still, the algorithm simply can not afford the healthiest prediction outcomes. Feature engineering, the method of transforming dataset for machine learning, is including an essential factor in propagating reliable prediction outcomes. The purpose of this article was to analyze feature engineering and selected method, in terms of their capacity to enhance the prediction outcomes. Two diverse datasets were examined with three various machine learning techniques, and their outcomes were analyzed applying four evaluation criteria. Techniques applied were naïve Bayes classification, decision trees, and linear regression. For the computation of machine learning techniques, feature engineering was employed to the modified and raw version of the dataset individually. The central process of feature selection was feature engineering. In the state of regression and classification trees, supplementary feature engineering was prepared in the manner of custom feature production. Feature engineering was performed both with automatic interpretation and manual functionality of the dataset. In summation, features tuning was made with an error and trial strategy. The outcome of both datasets shows differences and relationships with their employment in fundamental education. In the first dataset, the relationship is that recall value was frequently greater than the precision value. The variation was in the efficiency value. This can be connected to the various individualistic attributes. In the original analysis, the dependent attribute was not changed into binary valued attributes. Concluding the dependent attribute might have addressed the predictions carefully in this research. The second dataset, the primary analysis adopted an additional attribute that symbolizes the prior exam ranks and delivered better performance than from this article. Though, once those attributes are neglected, accuracy rates were comparable to those paper.

8. REFERENCES

[1] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. “Data clustering: a review.” ACM Computing Surveys (CSUR) 31.3 (1999): 264-323.

[2] The Apache Mahout project’s goal is to build a scalable machine learninglibrary.http://mahout.apache.org/

[3] Apache Hadoop. http://hadoop.apache.org/ Last accessed: 02/19/2015

[4] Wiederhold, G., Foreword. In: Fayyad U., Shapiro G.P., Smyth P., Uthurusamy R., editors, Advances inKnowledge Discovery in Databases. California: AAAI/MIT Press, 1996;2.

[5] Han, J. and Kamber, K., Data mining: Concept andTechniques. San Francisco: Morgan Kaufman Publisher (2001).

[6] Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, New Jersey: Printice Hall (1988).

[7] Kaufman, L. and Rousseeuw, P. J., Finding Groups inData: An Introduction to Cluster Analysis, New York:John Wiley & Sons (1990).

[8] Ng, R. and Han, J., Efficient and Effective ClusteringMethod for Spatial Data Mining, Proc. of the 20th
VLDB Conf. 1994 September. Santiago, Chile (1994).

[9] Zhang, T., Ramakrishnan, R., and Livny, M., BIRCH: an Efficient Data Clustering Method for Very Large Databases, Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, 1996 June. Montreal, Canada
(1996).

[10]. Sun Hongjie, “Research on Student Learning Result System based on Data Mining”, IJCSNS International Journal of Computer Science and Network Security, Vol.10, No. 4, April 2010

[11]. Academy Connection – Training Resources In html, http://www.cisco.com/web/learning/netacad/index, December 28th, 2005.

[12]. Wayne Smith, "Applying Data Mining to Scheduling Courses at a University", Communications of the Association for Information Systems, Vol. 16, Article 23, 2005.

[13] Erkan Er. "Identifying At-Risk Students Using Machine Learning Techniques", International Journal of Machine Learning and Computing, Vol. 2, No. 4, pp. August 2012.

[14] S. Kotsiantis, I.D. Zaharakis, and P. Pintelas, "Assessing Supervised Machine Learning Techniques for Predicting Student Learning Preferences"

[15]. Wayne Smith, "Applying Data Mining to Scheduling Courses at a University", Communications of the Association for Information Systems, Vol. 16, Article 23, 2005.

[16]. Baradwaj, B. and Pal, S. (2011) „Mining Educational Data to Analyze Student s‟‟ Performance‟‟, International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69.