Data Preprocessing on Student Performance Data

Shubham Barudwale
School of Computing Sciences and Engineering, VIT Chennai, Tamilnadu, India 600127
Email: barudwaleshubham.dinesh2015@vit.ac.in

Abstract-

Machine learning is a field of data science where we make the machine learn the information from the data available. We take training data to learn and test the result on the testing data. But when we get the data from the outer real world it might not be perfect each time. So we perform some operations in order to make the data compatible to process algorithms and to take care of non available or noisy data. This process is called as Data Preprocessing.

1. Introduction

Machine learning is the process in which the algorithm will take the data input and will try to find out the decision boundary to classify the given input into appropriate classes and will make some predictions. When we do all this data available plays a major role in the learning process. Some data can be missing from the dataset or could be noted with different units. Text might be available in the entries. So due to all this noise the data will be distrubing the proper classification of data. So to improve the learning and make the predictions proper we clean the data before processing it through various algorithms.

2. Methadology

Noisy data can contain duplicate data, improper data, missing values, text data. To make data clean we can either ignore the dirty data or can assign some values based on other values of the same field. While preprocessin the data we will check for duplicate entries and we will delete them. For missing data we can ignore them or assign the zero values. The better method to deal with such data is to take average of the data in the same field and put the value in the missing places so that the generalization can done. We cannot deal with text while drawing the decision boundries so to reinforce that we use the temporary aasignment values for each unique text field.

By plotting the scatter graphs, seeing the mean and mode plots and values we can easily identify the outliers in the data. Graphical methods are very useful in such data preprocessing techniques.

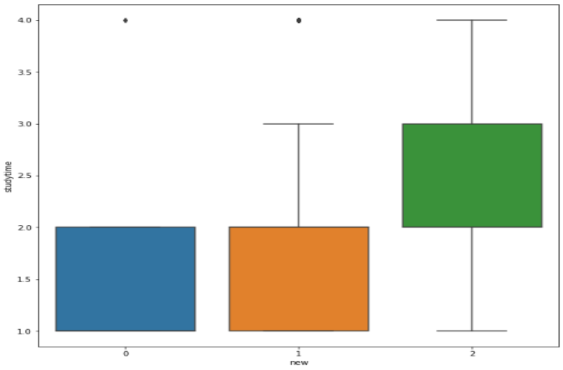
3. Dataset – Student Performance Dataset

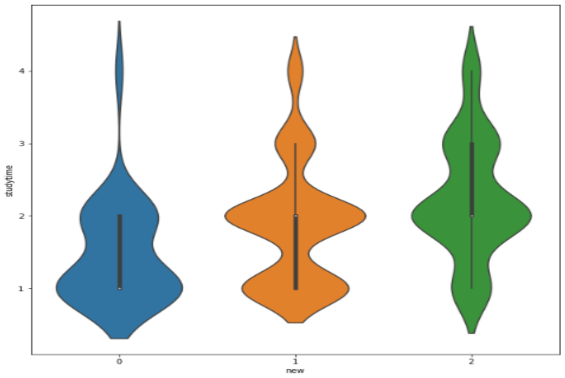
Student performance dataset is the dataset which has data about 650 students in two schools and their academic performance. It has thirty three attributes for each student e.g. School, sex, age, Family Size, study Time, health and the target academic marks. The school GP has 423 records and rest are of MS. The many attributes of data are represented as strings. e.g. Father's job, school, sex etc. And rest are as integer entries.

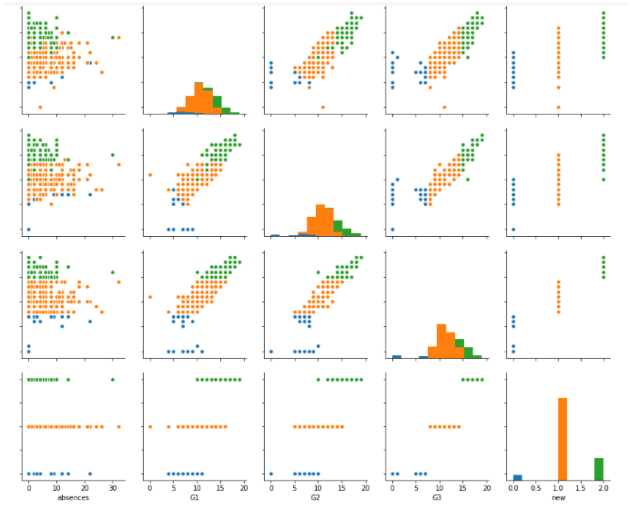
5. Experiments

The student information was gicen as sring with repeted fixed entries. So after finding out the fixed inputs to the fields, I changed the data to integer values so we can use them easily. Two schools were labeled as 0 and 1, sex was categorised in 0 and 1. The data which had multiple values were ranged from 0 to n, i.e. number of unique entries. e.g. Family size, Mother's job, Father's jobetc.

I also drew and visualized the various interrelation between data.







6. Conclusion

The data in the real world always contains noise. To make learning more precise we need to clean the data with appropriate techniques in order to get desired output as learning. The data can be cleaned using two major techniques i.e. by replacing the data with other valu or by ignoring the particular value.

Here in our data we have replaced the string with numbers in order to make data usable for leaning and applying algorithms on data. I have also found some outliers in the data.

7. Referances

1)http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-3.html

2) http://scikit-learn.org/