Data Wrangling

Data wrangling is about cleaning and reading the data so that it can be used for further analytics.

About the Data-

The data contains two CSV files, one having scores of Mathematics and another have the score of the Portuguese language. The data was obtained in a survey of student’s math and Portuguese language courses in secondary school. It contains a lot of interesting social, gender and study information about students [1].

Steps were taken in Data Wrangling process-

1. Read the CSV files into the two separate data frames named as ‘math\_df’ and ‘port\_df’.
2. Make a join between these two data frames and see how many students are common, the common data frame is named as the ‘common\_df.’
3. Now a correlation analysis is done for the mathematics and Portuguese final scores for the 382 common students. It is found that the correlation coefficient is – 0.48. This shows a strong relationship between the scores.
4. In the next step a graph is plotted to visualize the relationship, The first scatter plot is confusing and does not the convey the required information to the viewer.
5. To make the visualization clear, the slice of the data frame is plotted as a line plot. It can be easily inferred now that students with high mathematics score have high Portuguese scores and vice versa. The probability of this trend is almost ~50%.
6. To catch the outliers, it is checked if any final grade is greater than the 20. The final grades are on the scale of 1 to 20, so no grade should be greater than 20. There were no outliers in the code.
7. Similarly, both the data frames are checked for any null values.
8. Most of the features in the data set are categorical. They need to be converted into one hot encoding. Most of the features like ‘traveltime’, ‘studytime’, ‘famreal’ and ‘famtime’ are label encoded. These features represent the quantities like how much a student travels to reach school? How much is the average study time? How much average time is spent with the family? All these features with higher value represents the higher amount of these features, which make sense. Machine learning algorithms should give more weightage to these features if they have higher values.
9. However, on the other hand, the features like- ‘Mjob, ‘Fjob’, ‘reason’ and ‘guradian’ have categories which are independent of each other. Hence, it makes sense to one hot encoding for these features. It increases the number of features in the data set, but this is a necessary step to avoid misinterpretation by Machine Learning Algorithms.
10. After one hot encoding, the Data is clean and clear for further processing.

Reference-

[1] Data and source can be accessed at- https://www.kaggle.com/uciml/student-alcohol-consumption