Coding Club - Machine Learning Recruitment
Task 1 Report

Name : Ruchita Satish Jadhav

Roll no. : 240103092

For doing this task, I used Kaggle. I loaded the dataset, imported the required libraries, and fetched the dataset from the csv file as a starter to proceed ahead.

In the end, felt very happy and satisfied with my work. This was the first recruitment task among all the clubs, which I honestly enjoyed and could do things by myself :)

Level 1

It was about identifying the hidden features. However, before moving onto this, I felt that the data should be preprocessed first and started replacing the empty spaces with medians and modes, until I realised that it was supposed to be done for level 2. But nvm, I copied all of the content into a fresh notebook and continued further.

Firstly, I drew the Histograms for all the three features, to know the count of each data. This was done to know what kind of data is present in the column and it falls in what range.

Clearly, I could see that Feaure_1 contained integer data values ranging from 15 to 22, Feature_2 data ranged from integers 1 to 4 and from 1 to 5 in Feature_3.

Next, I drew a correlation heatmap, to find the correlation between various attributes, to guess the behavior of the unknown features.

Following were the observations for the three features:

Feature_1 : Couldn’t conclude anything as such from the correlation heatmap, but looking at the range of values in the histogram, I could infer that this feature could be **age**. The reason for this was that the values were integers ranging from 15 to 22, that can describe the age of a student.

Feature_2: From the heatmap, I was able to infer that this feature has a positive correlation with grades and negative correlation with freetime, traveltime, goout etc. So this could possibly be the **studytime** of the student with values ranging from 1: very less studytime to 4: very high studytime.

Feature_3: From the heatmap, I could see a strong correlation with goout and Dalc. Since, both of these can be associated with partying, I estimated this feature to be **partytime** with values ranging from 1: very less partytime to 5: very high partytime.

I renamed the column names accordingly and proceeded towards level 2.

Level 2

Firstly, found out the datatypes of all the columns. They were either object or float or int.

Next, I found out the number of missing values in each column.

There were two categorical columns with missing values: famsize and higher. I replaced the missing values in famsize with mode of the data. I avoided doing this with higher column as pursuing higher education is a matter of opinion and it would be wrong if I replace it with the mode. So to be on a safer side, I preferred replacing with NULL.

Replaced the missing values in all numerical columns with median of the data.

Level 3

Chose some insightful questions and obtained conclusion by visualising using bar charts, scatter charts and violin charts.

Following were the inferences drawn from the observations:

Q1 : Does the address(urban or rural) of the student affect the tendency of higher education?

-> : Students residing in urban areas are most likely to pursue higher education.

Q2 : Does weekday alcohol consumption affect the grades of the students?

-> : Higher weekday alcohol consumption is associated with lower and more varied final grades.

Q3 : Does number of school absences affect grades?

-> : Students with less or no absences have relatively higher grades than those with more number of absences.

Q4 : Is there a correlation between studytime and grades?

-> : There is no regular trend, but students with studytime 3 and 4 have higher grades than those with 2 and 1.

Q5 : Are students with higher grades romantic?

-> : Students with higher grades are not romantic.

Level 4

It was about analysing whether or not a person was likely to be in a romantic relationship. Firsty, encoded the data in categorical columns.

The target column (romantic) was split from the dataset as its behavior was to be estimated from the remaining features.

Dataset was split into train set and test set.

Classifiers such as decision trees, random forests and logidtic regression were used.

Best accuracy of 68% was obtained with logistic regression predicting "not in relationship".

Level 5

Used SHAP for two students, one predicted "NO" and one "YES".

a. for student who predicted "NO": Features such as health, absences, famsup, nursery, Dalc, etc. contribute positively to the prediction. Partytime and other 23 features contributed negatively to the prediction. Model output is positive, so the prediction is correct (model favours the decision).
b. for student who predicted "YES" : Most of the features contribute negatively to the prediction. Only sex, G3 and freetime contribute positively. Model output is negative, so the prediction is incorrect (model does'nt favour the decision).