COSC 5557
Exploratory Data Analysis

Almountassir Bellah Aljazwe

February 22, 2024

1 Data Exploration

1.1 Introduction

To provide some needed context, our predictive machine learning classifier model, from our first warm-up exercise, performed very poorly. It predicted the correct 'wine quality' class 16.68% of the time, and that is indicative of a problem; that problem could be with the dataset the model trains on, or the problem could be with the model itself.

Therefore, the goal of this section aims to explore and study the dataset used for the mentioned classifier model. We also incorporate a second dataset - the "Primary Tumor" dataset, in an attempt to compare our findings between our "Red Wine Quality" dataset and our second dataset.

The main areas we wish to explore relate to the characteristics of the raw dataset; they are the following:

- Shape : how many features and samples does the dataset contain?
- Target Feature :
  - Type : numeric or categorical?
  - What does it represent?
  - The distribution of samples to each target feature value (if categorical).
- Initial issues : any missing values?
- Correlation : any relationship between the features and the target feature?

1.2 Results

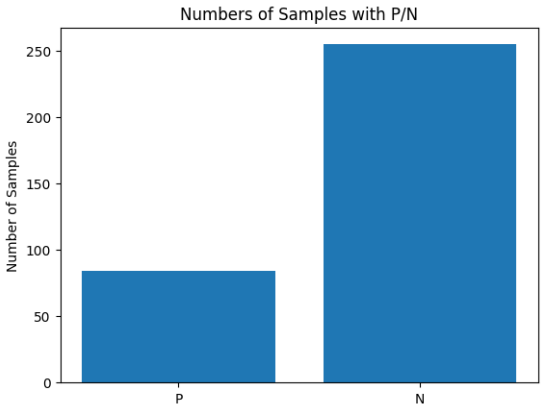
1.2.1 Shapes

With regards to the shapes of the dataset, such as the number of samples and features, both of the datasets differ from one another. The "Primary Tumor" dataset contains 18 features, including the target feature, and 339 samples. On the other hand, the "Red Wine Quality" dataset contains 12 features, including the target feature, and 1599 samples.

1.2.2 Target Features

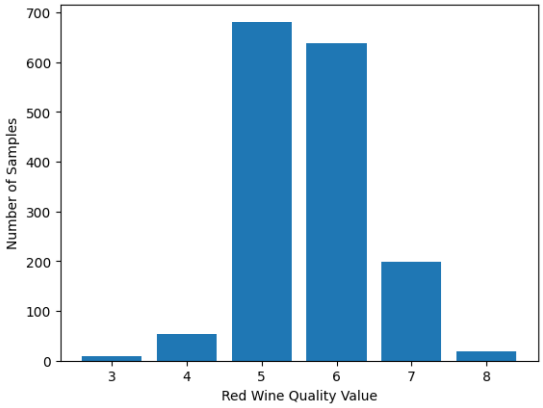
Two of the datasets have categorical target features. They differ, however, in the amount of categories a target value can be. To explain further, the "Primary Tumor" dataset contains only binary values for its target feature; these binary values are 'P' and 'N'. From the context of the dataset, these binary letters most likely represent the positive (P) presence of a primary tumor, or the negative (N) presence of a primary tumor in a single sample. On the other hand, the "Red Wine Quality" dataset contains 6 different categorical values for its target feature; these values are discrete integers, starting at '3' and ending at '8'. From the context of the dataset, these discrete integers most likely represent the number rating, with '3' being the lowest and '8' being the highest. There could be more values if more samples were to be added, such as '1' or '9', as the target feature values seem to be a number scale from one to ten.

Another important aspect, in exploring the raw dataset, is to look at the distribution of the samples with regards to the target values. The following figure displays the distribution of the "Primary Tumor" dataset :



P	N
84/339	255/339
$\approx 25\%$	$\approx 75\%$

Moving on to the "Red Wine Quality" dataset, the following figure displays the distribution of the "Red Wine Quality" dataset :



3	4	5	6	7	8
10/1559	53/1559	681/1559	638/1559	199/1559	18/1559
$\approx 0.6\%$	$\approx 3\%$	$\approx 44\%$	$\approx 41\%$	$\approx 13\%$	$\approx 1\%$

What we see from our results may offer some valuable information for us when we attempt to train our machine learning model on these datasets. In particular, the "Red Wine Quality" dataset is the most interesting; we can observe that the data distribution is cluttered around the middle quality values. This, in reality, is reasonable as it may be rare to encounter wine with extremely low quality or extremely high quality; it is reasonable to expect wine with average quality. With regards to our model, the imbalanced nature of this dataset may present obstacles in getting the best model performance; this is due to the low number of samples for quality values other than '5' and '6'. The "Primary Tumor" dataset, offers more samples for each categorical value, relative to the "Red Wine Quality" dataset. It could, however, be considered more imbalanced compared to anther dataset as there is an approximate 1:4 ratio of 'P' to 'N' target values.

1.2.3 Initial Issues

A clear issue when first viewing one of the raw datasets is the missing values. This issue is not present in the "Red Wine Quality" dataset, but it is present in the "Primary Tumor" dataset. For the "Primary Tumor" dataset, the following table describes the missing values, per column :

Sex	Histologic-Type	Degree-of-Diffe	Skin	Axillar
1	67	155	1	1

We can see that the "Histologic-Type" and "Degree-of-Diffe" columns contain the highest percentage of missing values : $67/339 \approx 20\%$ and $155/339 \approx 46\%$, respectively.

In terms of missing values, per row, the dataset contains 207 rows with missing values; that is $207/339 \approx 61\%$. However, most of the rows have only a single value missing. In addition, the maximum amount of missing values, per row, is only two.

1.2.4 Correlation

Exploring the correlation of individual features, with the target feature, is a potentially crucial piece of information; it may provide us clues as to the important features, when deciding the target feature value.

Beginning with the "Primary Tumor" Dataset, we can observe the correlation between each feature and the target feature. From what is observed, the features that offer the highest "Spearman Correlation" values are the following:

Histologic-Type	Degree-of-Diffe	Brain	Mediastinum
0.42	-0.46	0.22	0.45

These relatively high values can be supported through the plots below. For each relationship, a bar plot is used to count the number of 'P' and 'N' for each feature value. This can be seen in the following :

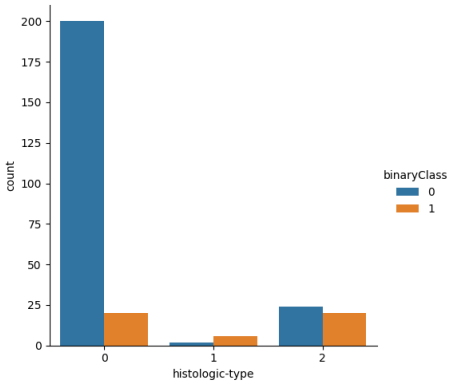


Figure 1: Histologic-Type Correlation

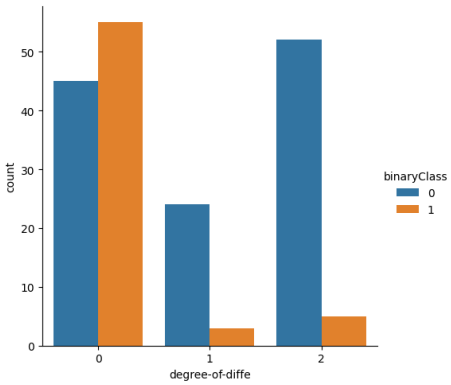


Figure 2: Degree-of-Diffe Correlation

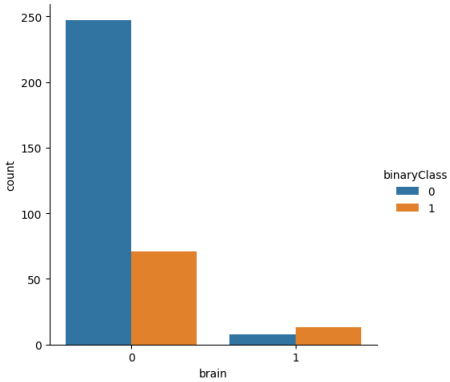


Figure 3: Brain Correlation

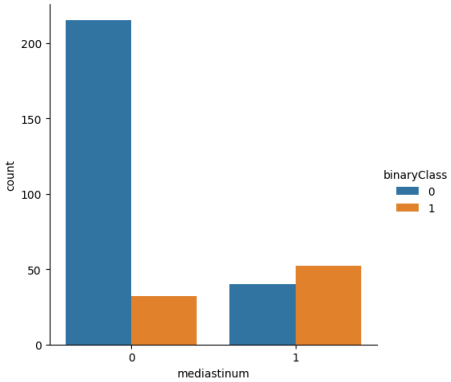


Figure 4: Mediastinum Correlation

Lastly, with the "Red Wine Quality" Dataset, we can also observe the correlation between each feature and the target feature. From what is observed, the features that offer the highest "Pearson Correlation" values are the following :

Volatile Acidity	Citric Acid	Sulphates	Alcohol
-0.39	0.23	0.25	0.48

These relatively high values can be supported through the plots below. For each relationship, a box plot is used to summarize the data corresponding to each target feature value. This can be seen in the following :

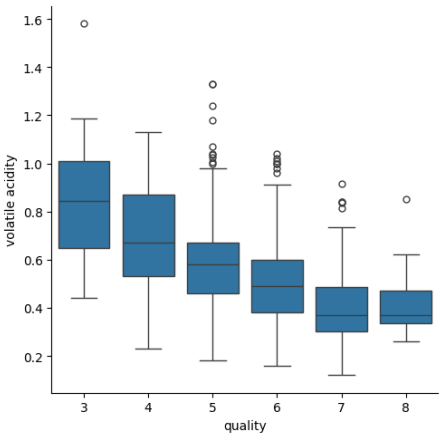


Figure 5: Volatile Acidity Correlation

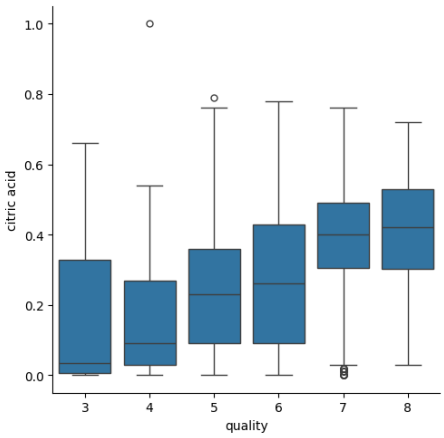


Figure 6: Citric Acid Correlation

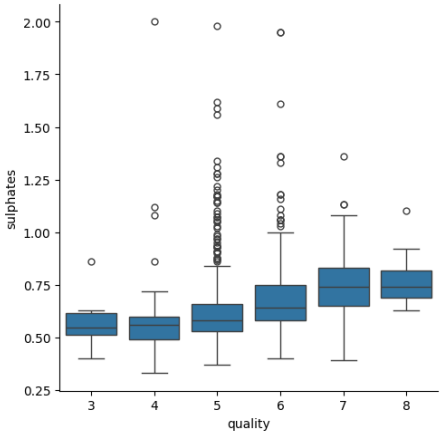


Figure 7: Sulphates Correlation

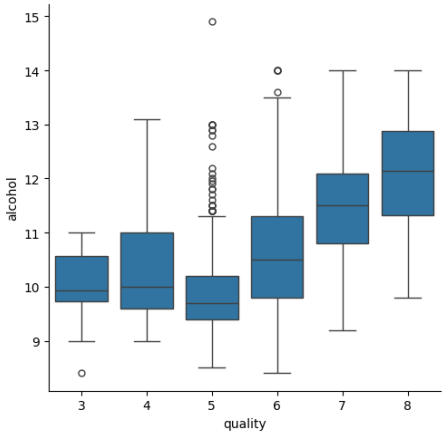


Figure 8: Alcohol Correlation