COSC480 - COMPUTER PROGRAMMING
Project

Wine Quality Analysis 🍷

INTRODUCTION

Wine is intriguing because of the complexity and layers of its flavour. This project tries to go deeper and understand the attributes and the basic factors that determine the quality of white wine and make it taste a certain way.

The Wine Quality Dataset has been taken from UCI Machine Learning Repository for white wine analyses.

This project aims to:

1. Determine factors that significantly affect the quality of white wine,
2. Interpret the effects of such factors on the quality of white wine,
3. Use the determined factors to predict the quality of white wines,

DATA SET DESCRIPTION:

[Taken from the website itself]
“The dataset is related to the white variant of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.”

Attribute Information:

Input variables (based on physicochemical tests):

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

The dataset is from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

The term "physiochemical" refers to the combination of physical and chemical properties or processes related to a substance or system. It represents the integration of both the physical and chemical aspects of a material or phenomenon.

Physio-chemical properties of wine are characteristics that can be measured and analyzed (using physical and chemical methods) to assess the quality of wine. These properties provide information about the composition, structure, behavior, and interactions of the substance.

Upon exploring the data for white white, these were the important observations made-

- The dataset contains 4898 rows and 12 columns. 100 rows have been randomly removed and put into a 'test set' which will be used to test the model later.
- The explained variable is discrete and not continuous-valued. The corresponding domain is {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, i.e., a score indicating the quality of wine.
- The explanatory variables are all continuous-valued and positive and are related to the physio-chemical properties of the wine.

HOW TO USE THE PROGRAM

The code file provided is in the format ".ipynb". This file can be loaded into Jupyter Notebook and run on the wine dataset.

The dataset must be uploaded to Jupyter Notebook before loading it into the program.

The file name can be entered in the variable 'filename'. This analysis has been performed on the white wine dataset, but the dataset for red wine can also be loaded. The program will then perform the same analysis on the red wine dataset and present a classification model.

THE DEVELOPMENT PROCESS

Initially I tried writing my python code in Rstudio because I wanted to export my final file into a PDF. It's not really built for python, so I shifted to Jupyter Notebook instead, so that I could also learn a new tool in the process. I installed Anaconda and Jupyter Notebook and then began my analysis.

I first outlined the steps of my approach. This was done after reading several blogs and asking CatGPT how I can analyse a wine dataset.

Here are the steps I took-

1. **Importing the libraries –** I used numpy, matplotlib, and statistics. These were taught in our course. The project specification mentioned that we had to use one library that wasn't taught in the course, so I used pandas because it has 'Series' (1-D labelled array) and 'DataFrames' (2-D labelled array with columns and rows). These features make it easy to handle tabular data such as CSV. Pandas also offers functions for data cleaning and transformation, indexing and selection and for computing descriptive statistics. It also integrates well with matplotlib.
   Although numpy can also handle arrays, numpy assumes that each column has the same datatype. Interestingly, Pandas internally uses NumPy for certain operations, so if you're using Pandas, you already have access to the underlying NumPy functionalities. Pandas is preferred for advanced analysis.
   I also used scipy and sklearn, but I only used specific functions from these to build my model.

2. **Importing the dataset –** I imported my dataset and looked at its shape. I set aside 100 observations as a 'test set' for assessing the accuracy of the model.

3. **Exploratory Data Analysis** – EDA involves examining and summarizing the main characteristics and patterns within a dataset. It helps to gain insights, identify relationships, detect outliers, and understand the overall structure of the data.
   I used `duplicated()` to check for duplicate values, and removed them from the dataset. I also used `head()`, `describe()`, and `shape`. I created a summary table to see if there were any missing values.

   Then I plotted the distribution of the explained variable, 'quality'. Its value ranged from 0 to 10. I found the mean of quality and decided to transform this discrete variable into a binary categorical variable with classes 0 and 1 so that I could perform logistic regression on it. I created a new column in my dataframe called 'class'. If the quality of the wine was above the mean, its class became '1' or 'good', and if the quality was below the mean, its class became '0' or 'bad'.

   Then I looked at the distribution of class to find out how many good and bad wines were in the dataset.

4. **Data cleaning and outlier removal** – Removing outliers can help provide a more accurate representation of the underlying data distribution and improve the validity of our statistical inferences. I looked at each of my predictors to see if there were any outliers. I did not want

to remove too much data from my dataset so I chose a threshold of ‘5 standard deviations away from the mean’. Any data point beyond this threshold (either above the upper bound or below the lower bound) was treated as an outlier and was removed. This threshold removes data points that deviate significantly from the majority of the data, which means it removes a relatively small number of data points. I also plotted the distributions for each of the predictors and also plotted lines to indicate the threshold for the outliers.
I found 41 outliers in the white wine dataset. This is not a big number so I decided to drop entire rows containing these observations.

5. **Model Building: Feature selection –** After obtaining my clean data, I started building my model. For that, I first needed to determine which predictors/features were important and actually had an impact on the quality of the wine. To determine this, I used independent samples t-test. For each predictor (using a for loop), I compared the mean values of the continuous predictor variables based on whether they were in class 0 or class 1. If there was a significant difference in the means, it indicated that this particular predictor did have an effect on the quality of the wine.

   H_0: the variable is not significant, i.e., the presence of that variable doesn't have a significant effect on the wine.
   H_a: the variable is significant, i.e., the presence of that variable has a significant effect on the wine.

   The p-value indicates the probability of observing such a difference by chance. If the p-value is below a chosen significance level (e.g., 0.05), we can reject the null hypothesis and conclude that the presence of that variable has a significant effect on the wine class.

   I stored my final selected variables in a list.

6. **Model Building: Fitting a model –** I fit a logistic regression model on my training data and obtained the coefficients.

7. **Evaluating the Model –** using the created model, I predicted the class labels for the training dataset and found the training MSE.
   Then I ran this model on my test data and calculated the test MSE. I also computed a confusion matrix and calculated the error rate.

FINDINGS AND RESULTS

The Training MSE was 0.2542 and the test MSE was 0.28. Both these values are quite low. The error rate was also 0.28. This indicates that the model performs decently.

I found that the main physiochemical attributes that affect the quality of a white wine are - 'fixed acidity', 'volatile acidity', 'residual sugar', 'chlorides', 'total sulfur dioxide', 'density', 'pH', 'sulphates', and 'alcohol'.

THINGS THAT WENT WELL

My outlier detection went well and was easy to implement.
It was also fairly easy to work with the Pandas library. I found it quite similar to NumPy.
I enjoyed this project and learned a lot from it, and it felt good to be able to apply my python skills on a practical challenge.

CHALLENGES FACED

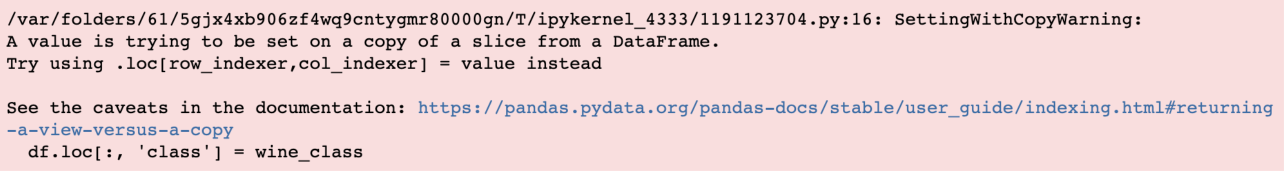
It was my first time performing an analysis like this and during my research I found out that there was so much I could do to analyze this dataset.

For example, there were so many different methods suggested for feature selection, like multicollinearity, tree-based selection, dimensionality reduction, etc. However, at this stage I don’t understand these concepts very well, so I refrained from using them in my analysis. I stuck to t-test because it has been taught in my course. I’m not sure if I did it correctly, because the t-test assumes normality and equal variances in the two groups and I didn’t know how to check for that.

Further, I got some feedback on my Statement of Intent saying that I could use seasonal information like amount of rain, sunshine etc. to predict wine quality, so that we might be able to tell if a particular set of wine, grown in a particular year will be better or worse. However, my dataset did not have those features so I could not perform that analysis.

I realized how vast data science is and how many things we have to look at, which can affect the goodness of a model.

I also got a warning while trying to create the ‘class’ column in my dataframe and I couldn’t figure out how to remove it.



```
/var/folders/61/5gjx4xb906zf4wq9cntygmr80000gn/T/ipykernel_4333/1191123704.py:16: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning
-a-view-versus-a-copy
  df.loc[:, 'class'] = wine_class
```

POSSIBLE FUTURE DEVELOPMENTS

Logistic Regressions learns a linear boundary for category separation, which may not be sufficient in our case. Using other stronger models like SVM, Neural Networks and Decision Trees, can help in overcoming this challenge.
More advanced techniques can be used for feature selection for a more accurate model.

Perhaps it is also possible to obtain a lower error rate and MSE rate in the future.

REFERENCES

1. https://www.youtube.com/watch?v=YaKMeAlHgqQ [“How do I select features for Machine Learning?”]
2. ChatGPT [“What are the steps in Exploratory Data Analysis?”]
3. ChatGPT was also used for debugging the code.

— END OF FILE —