**Wine Quality (Red)**

**Find a relevant dataset from Kaggle- On classification**

Jalin Roberts, jroberts12@bellarmine.edu

**ABSTRACT**

Up to 150 word summary of your project.

1. **INTRODUCTION**

I chose to use a dataset on Kaggle called “Red Wine Quality.” I decided to use the logistic regression classification model to analysis this dataset. The reason I chose this model was because it gives a fairly interesting perspective on overall wine quality and we can find different tests to give us different quality outcome(s). With this dataset, I explored the data with EDA in Python to analyze the overall structure and fit of the data. He purpose of this is to discover how other measurable factors can lead to poor or good wine quality.

1. **BACKGROUND**
   1. *Data Set Description*

This dataset can be found at <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>. I chose to work with this dataset because it gives an interesting perspective on something I enjoy, wine! UCI Machine Learning is the author of this dataset, and I decided to pursue looking into it because it is a simple and clean dataset that is good for modeling and practice.

* 1. *Machine Learning Model*

Logistic regression is a machine learning model that uses statistics to model and analyze the relationship between a dependent variable and one( or typically more) independent variables. This typically works through splitting the data into a ratio (typically 80/20) that we will use for training and testing, respectively. Once we have this, we can use the information now to predict the probability of our dependent variable taking on a certain value or range of value(s).

1. **EXPLORATORY ANALYSIS**

This section will be similar to your exploratory analysis project. First, provide a summary of the data set similar to your first exploratory analysis: *e.g. this data set contains 398 samples with 7 columns with various data types*. In this summary, provide the data types of your columns (in a table) and then rather than providing tabular statistics and plots for each variable, provide only statistics and plots that seem unusual. For example, if one or two variables have significant missing values or the distribution of the variable is skewed or looks unusual note that. Provide the unusual statistics or plots in this section. Provide any appropriate plots (e.g. correlation matrix, heatmaps, bar charts, etc.) that you deem necessary.

**Table 1: Data Types**

|  |  |
| --- | --- |
| *Variable Name* | *Data Type* |
| Fixed Acidity | Float64 |
| Volatile Acidity | Float64 |
| Citric Acid | Float64 |
| Residual Sugar | Float64 |
| Chlorides | Float64 |
| Free Sulfur Dioxide | Float64 |
| Total Sulfur Dioxide | Float64 |
| Density | Float64 |
| pH | Float64 |
| Sulphates | Float64 |
| Alcohol | Float64 |
| Quality | Int64 |

1. **METHODS**
   1. *Data Preparation*

This dataset was already prepared for us when I downloaded it. Because it is a pretty simple dataset, there is not a ton of manipulation that needs to be performed. There was nothing that really needed to be dropped from the dataset, because everything was relevant. There are no known null values, so there was almost no additional work needed to prepare the dataset as a whole.

* 1. *Experimental Design*

I ran the model several times with different parameters to find the best fit for the data with the least error. We tried different ratios of splitting the data into training, testing, and validation sets. The table below shows all the different settings we tried for the parameters.

Table X: Experiment Parameters

|  |  |
| --- | --- |
| **Experiment Number** | **Parameters** |
| 1 | All eleven (11) raw features with 80/10/10 split for train, validate, and test |
| 2 | All eleven (11) normalized features with 80/10/10 split for train, validate, and test |
| 3 | All eleven (11) raw features with 70/15/15 split for train, validate, and test |
| 4 | All eleven (11) normalized features with 70/15/15 split for train, validate, and test |

* 1. *Tools Used*

Describe all of the software tools you used to perform your data preparation and model implementation. For example:

The following tools were used for this analysis: Python v3.5.2 running the Anaconda 4.3.22 environment for Apple Macintosh computer was used for all analysis and implementation. In addition to base Python, the following libraries were also used: Pandas 0.18.1, Numpy 1.11.3, Matplotlib 1.5.3, Seaborn 0.7.1, SKLearn 0.18.1.

1. **RESULTS**
   1. *Classification Measures*

Provide the classification measures for each experiment using a confusion matrix and classification report.

* 1. *Discussion of Results*

Explain your model finding using a confusion matrix, the accuracy score and classification report.

* 1. *Problems Encountered*

Building the confusion matrix was problematic because it has more squares in it than I was used to dealing with, so reading/understanding it was difficult.

* 1. *Limitations of Implementation*

I think for something like a wine dataset, your basic scatter plot/histogram may work out better in specific situations, not everything will be best explained using things like confusion matrix’s.

* 1. *Improvements/Future Work*

I think that having more variety in a dataset could prove to ultimately be more beneficial for dynamic analysis. While we have everything necessary to perform our analysis, there’s not a ton of overall information to allow for different problem-solving techniques, and while beneficial for beginner Logistic Regression, it does not able us to have a multi-step problem to work out.

1. **CONCLUSION**

Ultimately, I wanted to understand how wine quality can be affected by other factors like citric acid, density, and acidity. The purpose of this analysis is to better understand and estimate the relationship between our variables. By performing this regression, we have a closer understanding of what it happening, and how things are affected by specific independent variables.

**Other directions:**

1. 10-pt, Times New Roman, 1” margins all around (if you use this template you are already set).
2. Ensure all tables and figures are numbered appropriately and referenced in the text. See examples above and below.

|  |  |
| --- | --- |
| **Figure 1: Comparison of X/Y from dataset (single plot) (8 pt.)** | **Figure 2: (a) Function Output (b) A against B (multiple plots) (8 pt.)** |