Timothy Bender

CSCI 347 – Data Mining

Final Project – Wine Classification

April 15, 2021

Wine Quality Data Analysis and Keras Classification

## Links:

*Dataset:* <https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv>

*TensorFlow:* [*https://www.tensorflow.org/*](https://www.tensorflow.org/)

## Introduction

This project will attempt to perform data exploration and analysis on the Wine Quality dataset for red wine available from the UCI Machine learning repository. Using datamining techniques, I will attempt to ascertain whether certain attributes of wine have a high correlation with the quality of the wine. The question of whether certain attributes affect others and thus have a high correlation with each other will also be investigated.

After performing data exploration and any necessary pre-processing a TensorFlow-Keras model will be created to predict wine quality based on a wine’s attributes. For our purposes the quality attribute of the dataset will be categorized using a sigmoid function. The Keras model will be evaluated based on its accuracy and then compared against a different model to generate a model with the highest accuracy possible.

# Into the Data

The dataset contains 1599 instances and is made up of twelve attributes. The first eleven attributes are floating point attributes pertaining to various aspects of the wine such as alcohol content, and citric acid levels. The twelfth attribute is an integer-categorical attribute which denotes the quality of the wine measured on a 0-10 scale. The dataset contains no missing or null values, and therefore no forward filling methods will need to be implemented. The quality attribute will be treated as a categorical attribute and converted threshold label encoding for use in binary classification.

# Data Exploration

Text

Description automatically generated Before determining what pre-processing, steps need to implement, we must perform some data exploration. To begin a pair plot for all attributes was generated using Seaborn.

By examining this plot, it appears that several attributes may have a high correlation with each other. While the most interest lies in attributes with high correlations with the quality of the wine, if multiple attributes have very high correlations our dataset may prove to be biased and an accurate model may be impossible to generate.

The correlations between attributes was displayed using a Seaborn heatmap. Several attributes are indeed displaying very high correlations with each other. Citric acid appears to have a very high correlation with fixed acidity. Density appears as well to feature a high correlation with fixed acidity. Interestingly the alcohol level of a wine could possibly correlate strongly with its quality.

Chart, histogram

Description automatically generated

Since alcohol content seems to be the only attribute with a reasonably high correlation with the quality of wine, a closer examination into their connection will need to be done. To examine this connection a scatterplot was created, and a line of best fit overlayed on top of this plot. A clear linear relation between the two attributes can be seen, implying that a higher alcohol content correlates strongly to a higher wine quality.

While a higher alcohol content seems to indicate a higher wine quality, this is not a suitable parameter from which to entirely gauge the quality of a wine. There are numerous examples of high-quality wines which contain low alcohol percentages. However, it does appear that most quality wines contain a higher alcohol percentage.

Chart, line chart, scatter chart

Description automatically generated

A picture containing chart

Description automatically generated To better understand the relations between attributes, the covariance was now inspected using another Seaborn heatmap. As expected, free sulfur dioxide and total sulfur dioxide have a very high covariance, undoubtedly due to their intrinsically linked nature. Notably residual sugar and free/total sulfur dioxide seem to share an extremely high covariance. Once again, the alcohol content seems to be the only attribute with a high covariance with quality.

Through data exploration the connection between the alcohol content of wine and the quality has now been made. However, the complexity of determining a wine’s quality has also been made apparent from the number of outlier’s who do not conform to this linear approach to classification. It is therefore clear as expected that a wine’s quality may only be accurately measured using numerous attributes. For this reason, no dimensionality reduction will likely be done before training the neural network.

Before creating the Keras model the variance of the quality attribute should be inspected to ensure that this dataset is not biased or incomplete.

Chart

Description automatically generated

As shown in the histogram created above, the quality attribute seems especially clustered between five and seven. Based on this result, the dataset is unvaried, and any training model created will likely not be able to account accurately for any outliers. Accuracy for this specific dataset will likely be unhindered.

# Pre-Processing

To begin, z-score filtering was implemented to remove outlying instances. The z-score of each attribute was calculated and then each instance was evaluated. Any instance which contained a z-score variance from the attribute variance of three or greater was removed from the dataset. This had no measurable impact however on the accuracy of the model when tested, leading me to believe that each attribute is very uniform in nature.

Normalization was not performed upon the dataset. Upon completion of the model, should the model be used to generate predictions for newly created data instances the same normalization would need to be performed up on the instance, making it overly complicated if not impossible for a second-hand user.

The quality of a wine must be categorized in order to make binary classification possible. For this reason, label encoding was used on the quality attribute, with any quality above seven being treated as a good wine and anything below seven a sub-par wine. A threshold of six was tested but was shown to drop the accuracy of the model from 92% to 62%. Label encoding was preferred over one-hot-encoding to preserve the dimensionality of the data and ensure that the neural network could generate a one-dimensional prediction as required by the sigmoid activation function.

Finally, the data was split into training and test segments with a test size of 10%. A random state of three was used to ensure that every compilation resulted in the same test and training data sets. Test sizes higher than 10% proved to reduce the accuracy of the model.

## Creating the Keras Model

To begin the Keras model the GlorotNormal kernel initializer was selected as it is the default initializer for most classification problems and proved to be the most accurate. Afterwards a Sequential model containing five dense layers was created.

|  |  |  |
| --- | --- | --- |
| Layer Type | Neuron count | Activation function |
| Dense | 48 | relu |
| Dense | 24 | relu |
| Dense | 12 | relu |
| Dense | 6 | relu |
| Dense | 1 | sigmoid |

Diagram

Description automatically generated

1. – A visualization of a binary classification neural network

Each Dense layer contains a relu activation function, as it is again the standard function for classification problems. Neuron counts between each layer were assigned to arbitrary numbers, discovered through trial and error. The final Dense layer features a sigmoid activation function to map our prediction to the range (0,1). A tanh activation function could also have been used, but sigmoid functions tend to be more reliable and are considered standard for binary classification.

While compiling the model, the adam optimizer was used as is standard. The loss function used was categorical cross entropy, the default for binary classification models. The model was now trained using 300 epochs upon the training subsets. Validation was run with the accuracy parameter being tracked against the testing subsets.

Upon examining the accuracy of the model, it was found to have achieved an accuracy of 91.25%, an impressive accuracy for such a basic model. Factors which affected the accuracy greatly included the relative size of the training/testing data subsets as well as the shape of the neural network. Trial and error were the only method used to create an accurate binary classification network.

While the model proved to be 91% accurate, it is worth mentioning that with such a skewed and biased dataset it is unclear how the model would perform on newly created data instances who do not fall within the same parameter standards that are seen in this dataset.

# Random Forest Classification

To visualize the Euclidean distance between different clusters a dendrogram was created, which showed high clustering into two primary clusters. This graph demonstrates that a tree based classification algorithm could very accurately predict the cluster assignments of data instances. Chart, histogram

Description automatically generated

To compare the binary classification neural network previously discussed, the Random Forest Classifier from sklearn was implemented and trained upon the same data subsets that the neural network was. The accuracy of this Classifier was found to be 93.1%, significantly improving upon the neural network.

# Summary

Through my analysis, I have determined that the alcohol content of a wine has a high correlation with the wine’s quality. However, to accurately determine the quality of a data instance all factors must be considered. I have also determined that this dataset is very biased regarding quality classifications, and a broader dataset with more varied qualities must be created before a universally accurate model can be created.

My work has resulted in two relatively accurate classification models, which others may use to predict the quality of wine. I have also determined the relative covariance and correlation between several attributes. Should this work continue, a more precise Keras model would need to be created in parallel with a more varied and less biased dataset. This would ultimately result in a model which would be accurate for any type of red wine across the range of qualities. The same model could likely also be used for white wine classification.

## References:

1. Goldberger, Ariel. “Training Neural Networks for Binary Classification: Identifying Types of Breast Cancer (Keras in R).” *Medium*, Duke AI Society Blog, 12 Mar. 2019, medium.com/duke-ai-society-blog/training-neural-networks-for-binary-classification-identifying-types-of-breast-cancer-keras-in-r-b38fb26a500c.