Stat 435 project

**Project: Wine Quality Prediction**

**1. Problem Setup:**

* Chose the Wine Quality dataset, both red and white wines
* Define a problem: Predict the quality of wine based on physicochemical attributes.
* Statistical Question: What statistical model can accurately predict the wine quality score?

**2. Implementation:**

* Used SVM, linear regression, and logistic regression
* Evaluated model performance using metrics like mean squared error, R-squared, F1 score, and MSE. Then used cross validation and ROC curves to evaluate models.

**3. Depth:**

* Explored the importance of different physicochemical attributes in predicting wine quality.
* Considered feature selection methods to identify relevant input variables.
* Implemented and compared various algorithms.

**4. Model Evaluations and Comparisons:**

**I** chose 3 models: linear regression, logistic regression, and SVM. I chose linear regression because it is often used when the target variable is continuous, which made it a good choice for wine quality scores. Linear regression gives coefficients for each predictor, which lets you interpret the impact of each feature on the target variable. This is valuable to understand the relationships between individual physicochemical attributes and wine quality.

Logistic regression is a binary classification algorithm, usually used when the target variable is categorical with two classes. In this project, I transformed the problem into a binary classification task, and predicted whether the wine quality was greater than 5 or not. Logistic regression coefficients represent the log odds of the target variable being in a particular class. The interpretation of coefficients helps identify which features contribute positively or negatively to the likelihood of the wine quality being greater than 5.

SVM is a versatile algorithm that can be used for both regression and classification. It’s especially useful when dealing with complex relationships and high dimensional data. SVM can capture nonlinear relationships between features and the target variable by using different kernels. This is beneficial when the relationships between physicochemical attributes and whine quality isn’t linear.

**Linear Regression Model Summary Red Wine:**

* **Residual Standard Error**: This is an estimate of the standard deviation of the residuals, approximately 0.6441. Smaller values indicate a better fit of the model to the data.
* **Multiple R-squared:** The model explains around 36.62% of the variability in red wine quality.
* **Adjusted R-squared:** A modified R-squared considering the number of predictors in the model, standing at 35.99.
* **F-statistic:** With a p-value < 2.2e-16, the model is statistically significant.
* **Red Wine MSE:** 0.43545

**Logistic Regression Model Summary Red Wine:**

* **Accuracy:** The model correctly predicts whether the quality is greater than 5 about 72.86% of the time.
* **Precision**: About 75.82% of the instances predicted as quality greater than 5 are true positives.
* **Recall:** Approximately 72.27% of actual instances with quality greater than 5 are correctly identified by the model**.**
* **F1 Score:** The harmonic mean of precisioßn and recall is 74%.
* **Red Wine MSE:** 25.345167

**Linear Regression Interpretation Red Wine:**

* The negative t-value and low p-value for "Volatile Acidity" suggest that higher volatile acidity is associated with lower wine quality.
* Other predictors with low p-values (e.g., "Residual Sugar," "Free Sulfur Dioxide," "Total Sulfur Dioxide," "Alcohol") are likely to be significant predictors of red wine quality.

**Logistic Regression Red Wine Interpretation:**

* **Volatile Acidity:** A negative coefficient indicates that higher volatile acidity is associated with a lower likelihood of red wine quality being greater than 5.
* **Free Sulfur Dioxide, Total Sulfur Dioxide, Alcohol:** Positive coefficients suggest that higher values of these features are associated with a higher likelihood of red wine quality being greater than 5**.**
* **Chlorides, pH, Density:** Negative coefficients suggest that higher values of these features are associated with a lower likelihood of red wine quality being greater than 5.

**Linear Regression Model Summary White Wine:**

* **Residual Standard Error**: This estimate of the standard deviation of the residuals is approximately 0.7488, indicating a reasonable fit of the model to the data.
* **Multiple R-squared:** The model explains around 28.55% of the variability in white wine quality.
* **Adjusted R-squared:** The adjusted R-squared, accounting for predictors, is 28.32.
* **F-statistic:** The F-statistic with a p-value < 2.2e-16 underscores the model's statistical significance.
* **White Wine MSE:** 0.576736

**Logistic Regression Model Summary White Wine:**

* **Accuracy:** The model correctly predicts whether the quality is greater than 5 about 73.32% of the time.
* **Precision**: About 76.62% of the instances predicted as quality greater than 5 are true positives.
* **Recall:** Approximately 86.18% of actual instances with quality greater than 5 are correctly identified by the model**.**
* **F1 Score:** The harmonic mean of precision and recall is 81%.
* **White Wine MSE:** 27.870901

**Linear Regression Interpretation White Wine:**

* The negative t-value and low p-value for "Volatile Acidity" suggest that higher volatile acidity is associated with lower white wine quality.
* Other predictors with low p-values, such as "Residual Sugar," "Free Sulfur Dioxide," "Total Sulfur Dioxide," and "Alcohol," are likely significant predictors.

**Logistic Regression White Wine Interpretation:**

* **Volatile Acidity:** A negative coefficient indicates that higher volatile acidity is associated with a lower likelihood of white wine quality being greater than 5.
* **Free Sulfur Dioxide, Total Sulfur Dioxide, Alcohol:** Positive coefficients suggest that higher values of these features are associated with a higher likelihood of white wine quality being greater than 5**.**
* **Chlorides, pH, Density:** Negative coefficients suggest that higher values of these features are associated with a lower likelihood of white wine quality being greater than 5.

**Support Vector Machine (SVM) Model:** For both red and white wines, a Support Vector Machine model was trained. The mean squared error was used to evaluate the model’s performance. The MSE for the red wine SVM model was .381 and for the white wine SVM model was .482.

**Cross-Validation:** Cross-validation was used to get a more robust estimate of model performance. The data was divided into 10 folds and the model was trained and evaluated 10 times, with each fold serving as the test set once. For the red wine model, the results of the cross-validation were as follows:

* **Root Mean Squared Error (RMSE):** 0.6519379
* **R-Squared**: 0.3466255
* **Mean Absolute Error (MAE**): 0.5081582

For the white wine model, the results of the cross-validation were as follows:

* **Root Mean Squared Error (RMSE):** 0.7483681
* **R-Squared:** 0.2868018
* **Mean Absolute Error (MAE):** 0.574295

These results provide a more reliable estimate of the model’s performance because they average the performance over multiple different splits of the data. This helps to ensure that the performance estimate is not overly optimistic due to a particularly favorable split of the data.

**ROC Curve:** The Receiver Operating Characteristic (ROC) curve was computed for the logistic regression models. The ROC curve is a plot of the true positive rate against the false positive rate, and it provides a useful way to evaluate the performance of a binary classifier. The area under the ROC curve (AUC) was also calculated, with a value closer to 1 indicating a better model. The ROC curve for the red wine model has an Area Under the Curve (AUC) of 0.857, indicating a very good model. The ROC curve for the white wine model has an AUC of 0.755, indicating a reasonably good model. These values suggest that both models are capable of distinguishing between wines with quality greater than 5 and those with quality less than or equal to 5.

**Data Visualization:** Scatter plots of volatile.acidity vs quality and boxplots of alcohol by quality were created for both red and white wines. These visualizations provide insights into the relationships between these features and wine quality.

**Feature Engineering:** Interaction terms were created between volatile.acidity and alcohol, and a log transformation was applied to residual.sugar. The interaction term captures the combined effect of volatile.acidity and alcohol on wine quality. Log transformation of residual.sugar is applied to handle skewness in the data. It’s helpful when the distribution of the variable is right skewed to make it more symmetric.