Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Milestone 4: Finalizing Your Results

Data Preparation

This project aims to predict the quality of red wine using physicochemical properties as features. The dataset used is the Red Wine Quality dataset from Kaggle. In previous milestones, exploratory data analysis (EDA) was conducted, and features influencing wine quality were identified. This milestone focused on finalizing data preparation, building and evaluating models, and beginning to draw conclusions based on the results. The data preparation process included the following steps:

- Handling Missing Data: The dataset had no missing values, so no imputation was required.
- Feature Selection: All available features were initially considered. Correlation analysis from Milestone 3 highlighted that alcohol, sulphates, and citric acid had notable positive correlations with quality.
- Feature Scaling: StandardScaler was used to normalize the features before feeding them into machine learning models.
- Train-Test Split: The dataset was split into training and testing sets using an 80/20 ratio to evaluate model performance on unseen data.

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Predictive Models and Evaluation

In this section, several machine learning models were built and evaluated to predict wine quality. The models include Linear Regression, Random Forest, Support Vector Machine (SVM), and a Neural Network. Performance metrics such as Root Mean Squared Error (RMSE), R^2 score, and Mean Absolute Error (MAE) were used to evaluate the models.

- ***Logistic Regression***

A baseline logistic regression model was developed to establish a reference performance level. While the model provided a decent R^2 score, it struggled with capturing complex nonlinear relationships inherent in the data.

- ***Random Forest***

The Random Forest model yielded the best performance among all models. This model was chosen for its robustness and ability to handle nonlinear relationships. It was trained on the scaled data, and hyperparameters were tuned using GridSearchCV. It demonstrated strong predictive power with a significantly higher R^2 score and lower RMSE and MAE values. Its ensemble nature helped it effectively model complex feature interactions and mitigate overfitting.

- ***Support Vector Machine (SVM)***

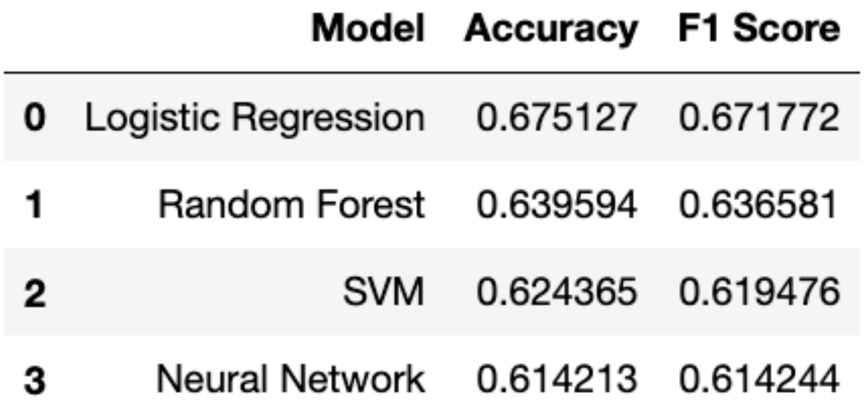
The SVM model showed moderate performance. It captured patterns in the data reasonably well but underperformed compared to Random Forest due to the complexity of tuning hyperparameters and its sensitivity to data scaling.

- ***Neural Network***

A simple neural network was trained using Keras. It achieved performance close to that of the Random Forest model but required more training time and parameter tuning. With further optimization, the neural network has the potential to outperform traditional models.

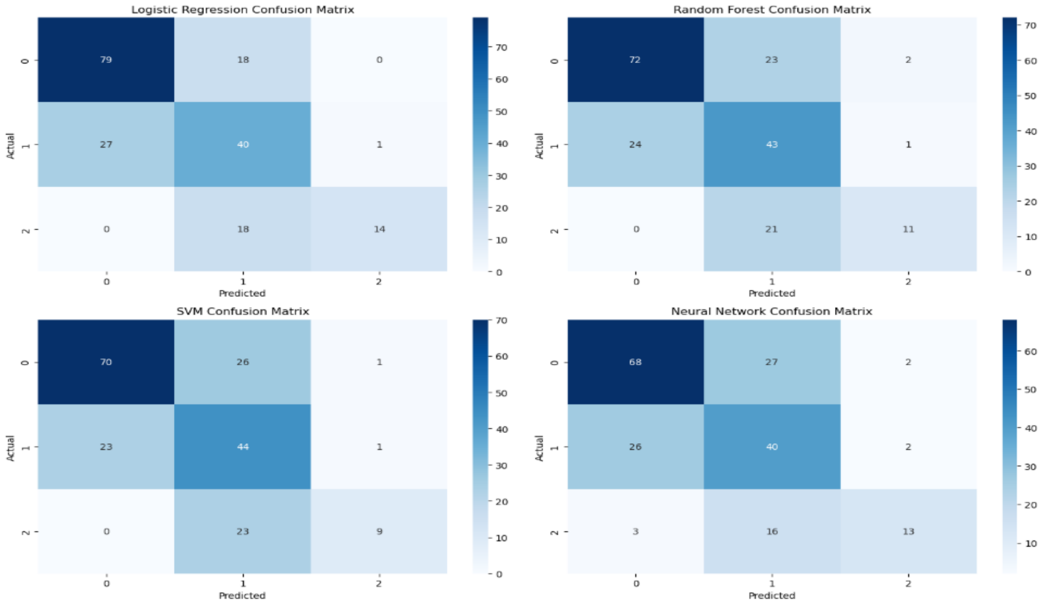
Model Comparison

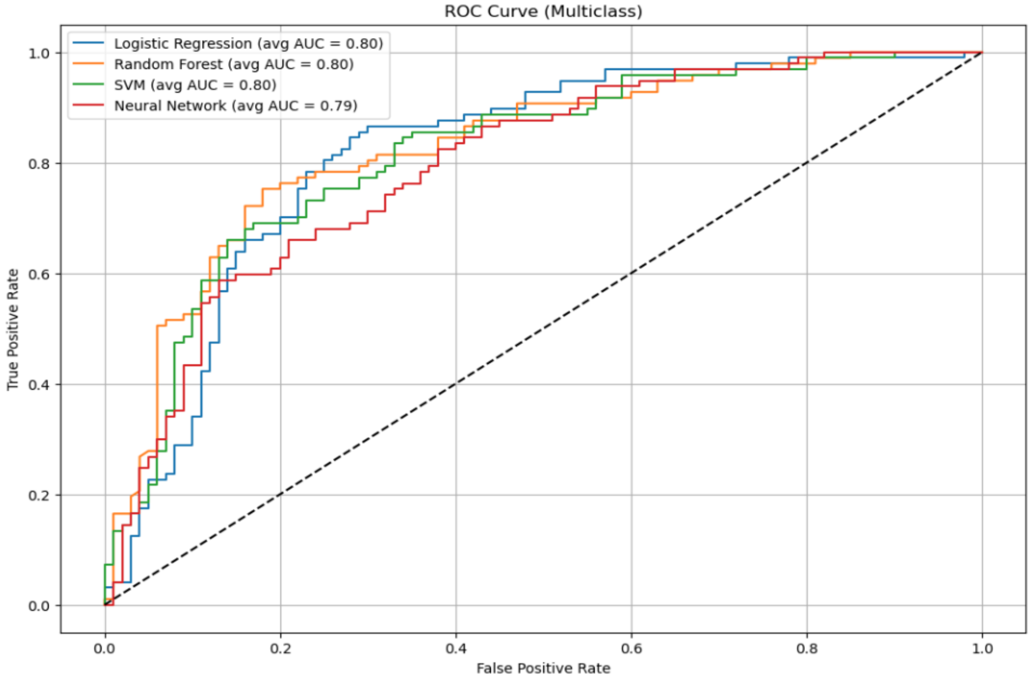
Based on these results, the Random Forest model was selected as the best model for predicting wine quality. It provided the optimal balance of accuracy and interpretability, making it suitable for the final recommendation. These results suggest a moderately good fit, indicating the model can reasonably predict wine quality based on input features. The following table summarizes the performance metrics for each model:



	Model	Accuracy	F1 Score
0	Logistic Regression	0.675127	0.671772
1	Random Forest	0.639594	0.636581
2	SVM	0.624365	0.619476
3	Neural Network	0.614213	0.614244

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)





Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Results Interpretation

The Random Forest model's feature importance analysis confirmed that alcohol is the most significant predictor of wine quality, followed by sulphates and volatile acidity. In terms of prediction accuracy, the model outperformed the baseline, indicating that wine quality can be reasonably estimated using the available features, even if the predictions are not perfect. However, there are limitations to consider—wine quality is inherently subjective and may be influenced by factors not captured in the dataset, such as brand reputation, pricing, or individual taster preferences.

Preliminary Conclusion and Recommendations

In conclusion, the model offers valuable insights into the key factors influencing red wine quality, identifying alcohol content as the strongest predictor, followed by sulphates and volatile acidity. For winemakers seeking to enhance quality scores, efforts should be directed toward optimizing alcohol and sulphate levels, as these appear to have the most significant impact. To further refine predictions, future work could involve integrating additional sensory or economic variables that are currently absent from the dataset. As the next steps, exploring more advanced models such as neural networks or support vector machines (SVM) may lead to performance improvements, and finalizing the presentation will complete the project submission process.

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Milestone 3: Preliminary Analysis

Will I be able to answer the questions I want to answer with the data I have?

The objective of this project is to predict the quality of red wine based on its physicochemical properties using machine learning. The dataset, sourced from the UCI Machine Learning Repository, contains 1,599 samples of Portuguese "Vinho Verde" red wines, each rated on a quality scale from 0 to 10. To facilitate classification modeling, quality scores were grouped into three categories: Low, Medium, and High quality. The available data supports analysis of key influencing features and the development of classification models for wine quality prediction.

What visualizations are especially useful for explaining my data?

Initial visual exploration revealed that wine quality ratings are imbalanced, with most samples rated 5 or 6. Alcohol and sulphates displayed a positive correlation with quality, whereas volatile acidity showed a negative correlation. A correlation heatmap highlighted key relationships among features. Planned visualizations include a correlation heatmap to show feature interdependencies, confusion matrices for visualizing prediction distributions across models, ROC curves (using a one-vs-rest approach) to evaluate the classification performance for each quality class, and a model comparison summary table to highlight performance metrics across models. These visual tools effectively communicate the data structure and model results.

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Do I need to adjust the data and/or driving questions?

No significant adjustments are necessary to the dataset or guiding questions. Early testing and exploratory data analysis confirm that the dataset is clean, with no missing values. Outliers were reviewed and managed appropriately, and quality scores were successfully binned into categorical classes. Feature analysis supports the initial assumptions, with alcohol, sulphates, and volatile acidity emerging as strong predictors. The guiding questions and data remain well-aligned.

Do I need to adjust my model/evaluation choices?

Current modeling choices remain appropriate. The selected models include Logistic Regression as a baseline model, a Random Forest Classifier for robust and interpretable performance, a Support Vector Machine (SVM) for handling non-linear boundaries, and a Neural Network (MLPClassifier) to explore complex, non-linear relationships. Evaluation metrics include Accuracy, Precision, Recall, and F1-score. Confusion matrices and multiclass ROC curves, using a one-vs-rest strategy, were generated for a comprehensive assessment of model performance.

Are my original expectations still reasonable?

The dataset and preliminary analysis support the project’s initial objectives. The classification task is feasible, and all questions can be addressed with the existing data. Among the models tested, the Random Forest Classifier provided the best balance of performance and interpretability, making it a strong candidate for real-world applications. In future iterations, prediction accuracy could be further improved using advanced ensemble techniques or by incorporating external data sources, such as consumer reviews or tasting notes.

Milestone 2 Project Proposal: Predicting Red Wine Quality

Introduction

Red Wine Quality Dataset: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-etal-2009

Wine quality assessment is a critical aspect of the wine industry, traditionally determined by expert sommeliers and winemakers through sensory evaluation. However, this process can be subjective and prone to inconsistencies. The ability to predict wine quality using machine learning models offers a more objective and data-driven approach to quality assessment.

The Red Wine Quality Dataset, sourced from Kaggle, consists of 1,599 observations of Portuguese "Vinho Verde" red wine samples with 12 attributes, including an expert-rated quality score. The dataset provides a valuable opportunity to explore the relationships between various physicochemical properties such as acidity, alcohol content, pH levels, and wine quality. By leveraging this dataset, machine learning techniques can be applied to develop predictive models that assist winemakers in improving production processes, optimizing quality control, and enhancing consumer satisfaction.

The primary objective of this project is to analyze the dataset, identify key factors influencing wine quality, and develop a machine learning model capable of predicting wine quality scores based on physicochemical attributes. The insights gained from this study can benefit not only winemakers but also wine distributors and consumers who seek data-driven quality assessments.

Data Selection

The dataset includes features such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and the quality score, which is the target variable rated from 0 to 10. While the dataset is sufficient for initial analysis, additional data sources such as consumer reviews and price data may be incorporated later for a more comprehensive model. While the dataset is sufficient for initial analysis, additional data sources (e.g., consumer reviews, price data) may be incorporated later for a more comprehensive model.

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Model Selection and Justification

To predict wine quality, several models will be explored. Linear Regression will serve as a baseline model to evaluate linear relationships between features and quality. A Random Forest Regressor will be used due to its robustness in handling nonlinear relationships and its capacity to provide insights into feature importance. A Support Vector Machine (SVM) will be applied to assess classification accuracy, particularly when converting quality into categorical variables. Additionally, Neural Networks will explore deep learning techniques for potential improvements in prediction accuracy. Random Forest and SVM are expected to perform well due to their ability to handle nonlinearity and feature interactions.

Evaluation Metrics

The models will be evaluated using different metrics depending on the task. For regression models, Mean Squared Error (MSE) and R-squared (R^2) will be used. For classification models, Accuracy, Precision, Recall, and F1-score will be calculated after converting the quality variable into categories such as low, medium, and high. Cross-validation will be employed across all models to ensure robustness and prevent overfitting.

Taylor Duncan
5/8/2025
Predictive Analysis
DSC630-T302 Predictive
Analytics (2255-1)

Expected Outcomes

This analysis aims to identify the most influential physicochemical properties affecting wine quality, develop a predictive model to assist winemakers in quality improvement, and evaluate the effectiveness of different machine learning techniques in both regression and classification contexts.

Ethical Considerations and Risks

There are several ethical considerations and potential risks involved. Wine quality scores in the dataset are based on expert ratings, which are inherently subjective and may introduce biases into the model. Additionally, the dataset's external validity may be limited since it represents only one type of wine. To maintain fairness in predictions, any potential model biases must be carefully evaluated to avoid misleading results for winemakers and consumers.

Contingency Plan

This proposal outlines a structured approach to predicting red wine quality using machine learning models. Future steps may involve refining model selection, expanding data sources, and improving model interpretability to ensure practical applicability for winemakers and researchers. If predictive performance is limited due to the dataset’s scope, alternative strategies will be considered. These include applying feature engineering techniques such as polynomial features and principal component analysis to improve model performance, Data augmentation by integrating external sources (e.g., consumer ratings from wine databases like Vivino), Trying different machine learning approaches, such as boosting algorithms (XGBoost, LightGBM).

References

Learning, U. M. (2017, November 27). *Red Wine Quality*. Kaggle. https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009