Wine Chromatic Profile Prediction Project

Farhan Faisal, Daria Khon, Adrian Leung, Zhiwei Zhang

2024-12-07

Table of contents

Summary

In this project, we developed a machine learning model to predict the color of wine (red or white) based on its physiochemical properties, such as acidity, pH, sugar content, and alcohol level. By accurately identifying a wine's color, the model can help wineries and retailers maintain better-organized inventories—ensuring that red and white wines are stored, shelved, and marketed correctly. It can also serve as a quality-control checkpoint, quickly detecting any misclassification or labeling errors before the product reaches consumers. Over time, integrating such data-driven tools can streamline the supply chain, improve quality assurance, and ultimately offer a more efficient, scalable way for the wine industry to manage its products.

As such, we implemented a logistic regression model with balanced class weights and optimized it through hyperparameter tuning. The final model performed exceptionally well, achieving an accuracy of 0.98 on the test data. The precision-recall analysis indicated high precision and recall scores (above 0.98), further corroborated by the confusion matrix which showed minimal misclassifications.

While the model demonstrated strong predictive accuracy, the near-perfect results raised potential concerns about possible overfitting on the existing data, suggesting further evaluation

on truly unseen data is necessary. Moreover, the assumptions that the logistic regression model made (e.g. multicollinearity and linearly separable data) can hinder the future performance of the model if the assumptions do not hold for any new unseen data. With the limitations of our model, we further stressed the importance of investigating the abnormal results that could potentially bring us new insights into how the physiochemical properties of a wine affect its chromatic profile.

Introduction

Wine classification plays a crucial role in both production and quality assessment. It is a key step to enforce quality assurance and maintain the excellence of the product. However, traditional methods often rely on extensive research and subjective evaluations by experts. To ease the wine quality assessment process, this project seeks to answer the following question:

How do we accurately predict the chromatic profile of wine using its physiochemical properties?

Developing a machine learning model for wine classification has several advantages. For winemakers, it could provide a scalable method for analyzing large datasets, identifying trends, and optimizing production processes. For consumers and retailers, it could serve as a tool to verify wine characteristics without requiring advanced laboratory equipment. Through this project, we aim to contribute to the industry's adoption of data-driven approaches, enabling efficient, reproducible, and cost-effective methods for wine analysis.

Methods

Data

The dataset for this project is sourced from the UCI Machine Learning Repository (Dua and Graff 2017) and focuses on wines from the Vinho Verde region in Portugal. It includes 11 physiochemical attributes, such as fixed acidity, volatile acidity, pH, and alcohol content, collected from 1,599 red wine samples and 4,898 white wine samples.

Validation

We conducted several validation checks on our dataset, including assessments for duplicates, correct data types, and missing values, most of which passed successfully. However, the outlier check flagged a few variables with potential outliers. To keep the analysis straightforward, we chose not to remove these outliers for this iteration. Future iterations could explore handling these outliers more thoroughly to refine the analysis.

Analysis

We leveraged the logistic regression algorithm to build a classification model to predict whether a wine sample is red or white (as defined by the `color` column in the dataset). All 11 physiochemical features in the dataset, including fixed acidity, volatile acidity, pH, and alcohol content, were utilized for model training. The dataset was split into 70% for the training set and 30% for the test set.

Preprocessing steps included removing duplicate entries, ordinal encoding for the `quality` feature, and standardizing all numeric features to ensure uniform scaling. A randomized search with 10-fold cross-validation was conducted, using F1 as the scoring metric, to fine-tune the regularization parameter (`C`). This process helped minimize classification bias and maximize accuracy while identifying the optimal model. Balanced class weights were employed to address potential class imbalances in the dataset.

The Python programming language (Van Rossum and Drake 2009) and the following libraries were utilized for the analysis: NumPy (Harris et al. 2020) for numerical computations, Pandas (McKinney 2010) for data manipulation, Altair (VanderPlas 2018) for visualization, and scikit-learn (Pedregosa et al. 2011) for model development and evaluation. The complete analysis code is available on GitHub: https://github.com/UBC-MDS/DSCI522-2425-22-wine-quality.git.

Results

To evaluate the usefulness of each feature in predicting wine color, we visualized their distributions in the training dataset, color-coded by class (Figure 1).

The feature distributions (Figure 1) reveal significant differences between red and white wines across multiple features. For instance, red wines generally exhibited higher levels of volatile acidity, citric acid, and chlorides, whereas white wines tended to have higher residual sugar, free sulfur dioxide, and total sulfur dioxide. Alcohol content showed a near-normal distribution for both wine types, with white wines displaying a broader range and slightly higher peaks at lower alcohol levels. Interestingly, alcohol was positively correlated with wine quality, which, along with pH (lower for red wines), might play a key role in determining the color of wines. The distribution of sulphates was left-skewed, with red wines showing slightly higher concentrations.

While most predictors showed some overlaps, notable differences were observed in their central tendencies and spreads. Features like wine quality and residual sugar exhibited less distinction between classes, but we still retained them since their interactions with other features might enhance predictive power.

We also examined multicollinearity among predictors (Figure 2), initially identifying high correlations between `alcohol and quality`, as well as `total sulfur dioxide and free sulfur dioxide`. A closer inspection of other features revealed moderate correlations between `density and residual sugar`, reflecting the physical influence of sugar content on the density of wine. Similarly, `sulphates and chlorides` displayed a moderately positive correlation, which might be due to their combined impacts on wine preservation and stability. Furthermore, `volatile acidity and quality` showed a negative correlation, aligning with the perception that high levels of volatile acidity can degrade the sensory appeal of wine.

However, further validation using Deepchecks confirmed that none of these correlations exceeded the threshold of 0.8. As a result, we retained all features in the model to leverage potential interactions and maximize predictive insights. By including all predictors, the model was better positioned to capture both individual and joint effects of features, leading to a more comprehensive understanding of the factors influencing wine quality.

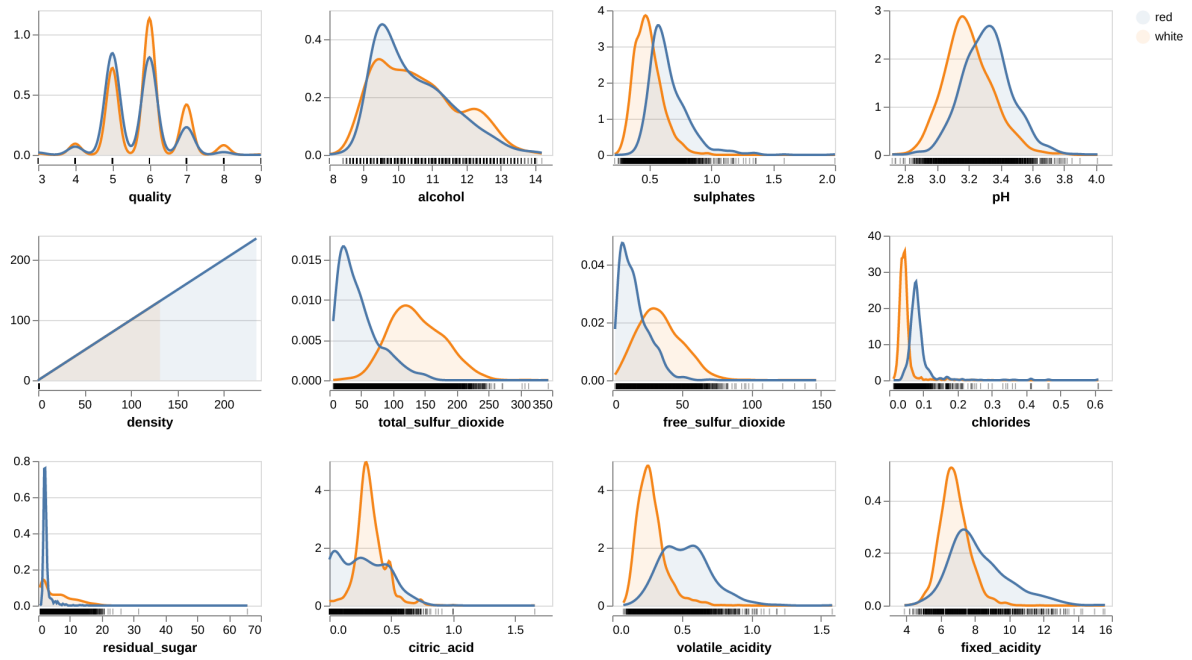

Figure 1: Distribution of Features per Target Class

We employed a logistic regression model for our classification task, utilizing a randomized search with 10 iterations for hyperparameter optimization. The primary goal was to determine the best regularization parameter (`C`), which was found to be 1.27, to maximize predictive performance. To evaluate model performance during the search, we used the F1 score (with "red" as the positive class) as our scoring metric. Cross-validation results using the best model were shown in Table 1.

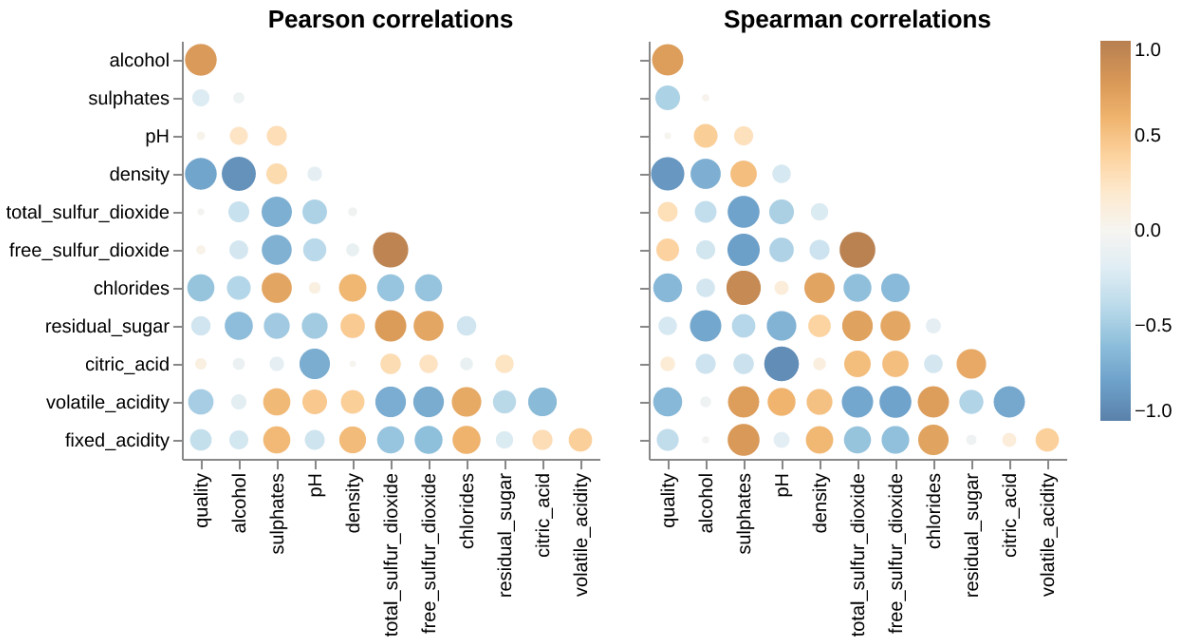

Figure 2: Correlation between Wine Color Prediction Features

Table 1: Random Search Best Model Cross-Validation Results

	test_accuracy	test_precision	test_recall	test_f1
0	0.994	0.987	0.991	0.989

Before evaluating the model performance on the test set, we validated that the target distributions between the training and test sets were comparable using a prediction drift check. The validation was successful, with a low Prediction Drift Score of 0, indicating that the predictions on both datasets were consistent and aligned with the expected target distribution.

Finally, we evaluated the model on the test set (Table 2). We also generated a confusion matrix (Figure 3) and a precision-recall (PR) (Figure 4) curve to further substantiate the results.

Table 2: Test Set Results

	accuracy	precision	recall	F1 score
0	0.991228	0.9775	0.987374	0.982412

Discussion

Stengths

Our model performed exceptionally well in predicting the colors of the wines as the model achieved an impressive test accuracy of 0.991. Moreover, both recall and precision scores exceeded 0.98. It fit our goal of accurately classifying a wine color, whether it be red or white, without human judgment. This can serve as a reliable and efficient algorithm for future wine analysis as long as we have sufficient information on the wine.

Based on the information from (Figure 3) and (Figure 4), the added insights indicated that

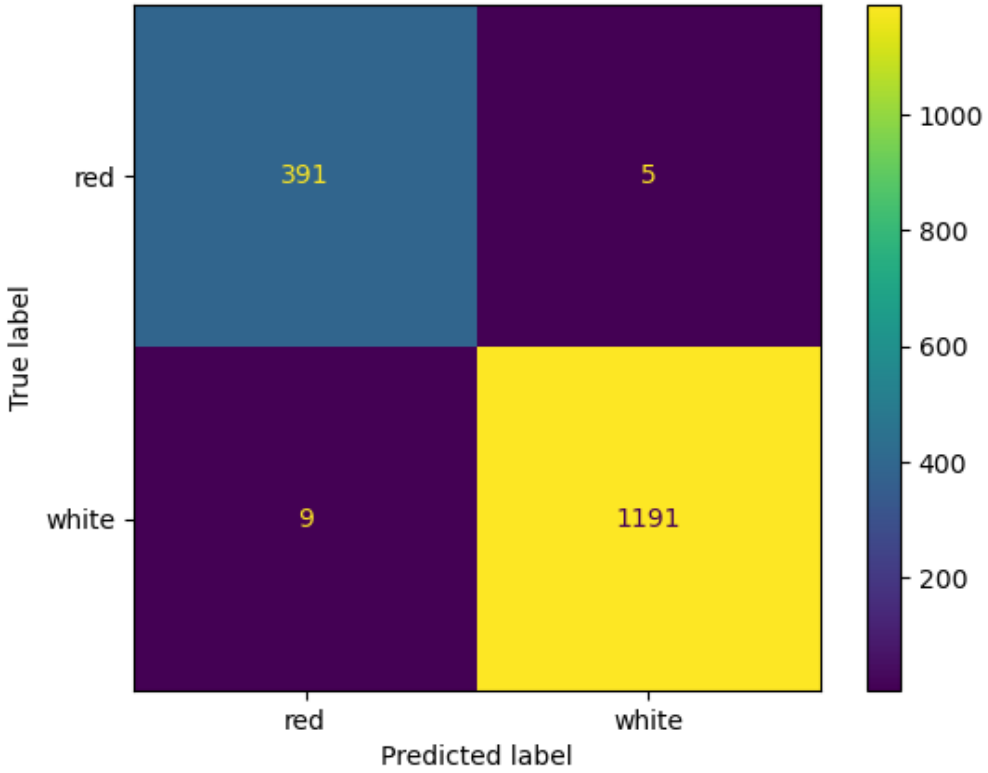

Figure 3: Confusion Matric

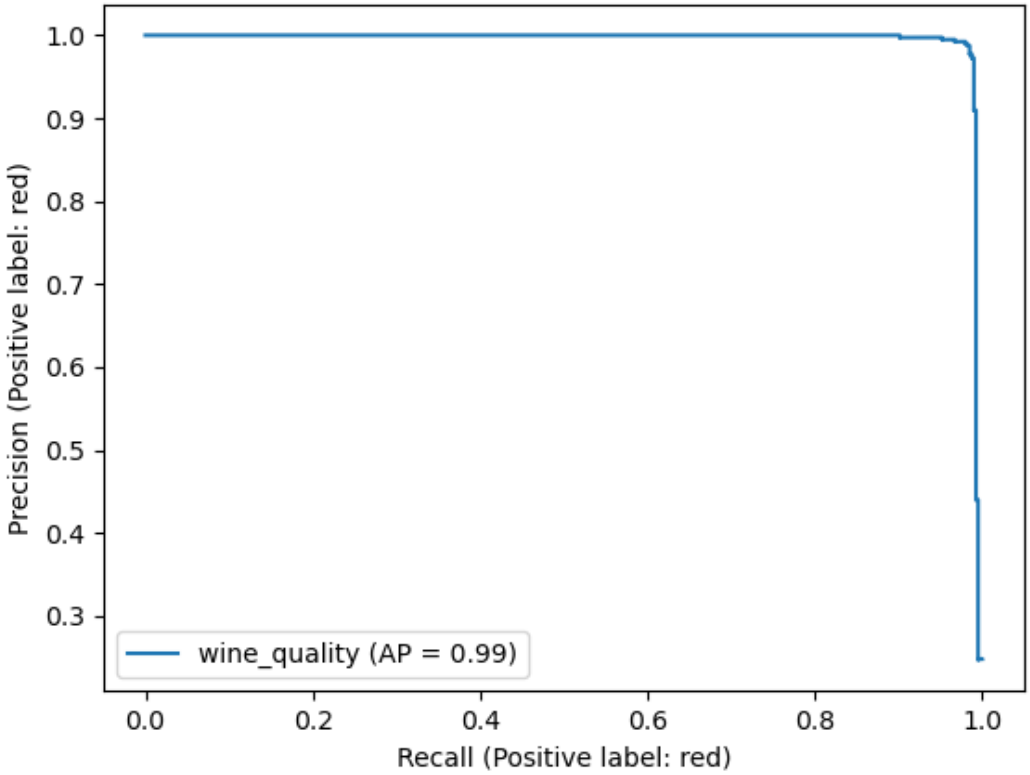

Figure 4: Precision Recall Curve

score of 0.99, the model exhibited near-perfect precision and recall balance. This indicated that at most thresholds, the model maintained high precision while effectively capturing almost all true positive cases, minimizing false negatives. These metrics jointly implied that the model is highly effective at differentiating between wine colors.

Limitations

However, while these high scores suggested the model was likely to perform well on new data, they also raised concerns about potential overfitting. The model's unusually exceptional performance on both the training and test sets might hint at a possible limited generalizability on truly unseen data. This will warrant further investigation and evaluation of new data to ensure robustness.

Moreover, the logistic regression model assumed that there was no multicollinearity within the features and that the data was linearly separable. By using this model, we were ignoring the possibility of dependencies across features and their subsequent impact on the wine color prediction. Also, we could not be absolutely certain that all newly introduced wine data could be well separated by the linear decision rule set by the model. This was supported by the existing errors from the confusion matrix (Figure 3) as it showed that there were possible outliers that our model failed to classify. Thus, human supervision over the model performance on wine analysis is still necessary since our model can make mistakes over time.

It is crucial to comprehend how and where our model fails in the predictions. The fact that the model had exceptional scores on the test data raised more concerns and emphasis on the actual errors. These errors could allude to hidden physiochemical rules, which the model failed to uncover, in the constituents of the wines. Thus, when we spot an error in predicting the wine color from the model in the future, we should investigate the constituents of the wine that cause the anomaly by interpreting the feature importances using tools like SHAP plots. This can help us understand what makes the wine an outlier and lead us to new insights into how the physiochemical composition of a wine determines its chromatic profile.

References

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences.

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, et al. 2020. "Array Programming with NumPy." *Nature* 585: 357–62.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference*, 56–61. https://pandas.pydata.org/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30. https://scikit-learn.org/.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

VanderPlas, Jake. 2018. "Altair: Interactive Statistical Visualizations for Python."