Wine Chromatic Profile Prediction Project

Farhan Faisal, Daria Khon, Adrian Leung, Zhiwei Zhang

2024-12-07

Table of contents

Summary . . . 1
Introduction . . . 2
Methods . . . 2
Results . . . 3
Discussion . . . 5
References . . . 8

Summary

In this project, we developed a machine learning model to predict the color of wine (red or white) using its physiochemical properties such as acidity, pH, sugar content, and alcohol level. This work emphasizes the potential for data-driven tools to optimize wine classification processes, offering a scalable and efficient approach for the wine industry. 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. multicolinearity and linearly separable data) can possibly 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 on the importance of investigating the anomalous

results that could potentially bring us new insights on 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 subjective evaluations and extensive testings 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

The logistic regression algorithm was used 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) reveals significant differences between red and white wines across multiple features. For instance, red wines generally exhibit higher levels of volatile acidity, citric acid, and chlorides, while white wines tend to have higher residual sugar, free sulfur dioxide, and total sulfur dioxide. Alcohol content shows 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 is positively correlated with wine quality, which, along with pH (lower for red wines), might plays a key role in determining the color of wines. Sulphates shows left skewed shape, with red wines showing slightly higher concentrations.

While most predictors showed some overlap, notable differences were observed in their central tendencies and spreads. Features like wine quality and residual sugar exhibited less distinction between classes, but we retained them, anticipating that 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 moderate positive correlation, potentially due to their combined impact 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, all features were retained in the model to leverage potential interactions and maximize predictive insights. By including all predictors, the model is better positioned to capture both individual and joint effects of features, leading to a more comprehensive understanding of the factors influencing wine quality.

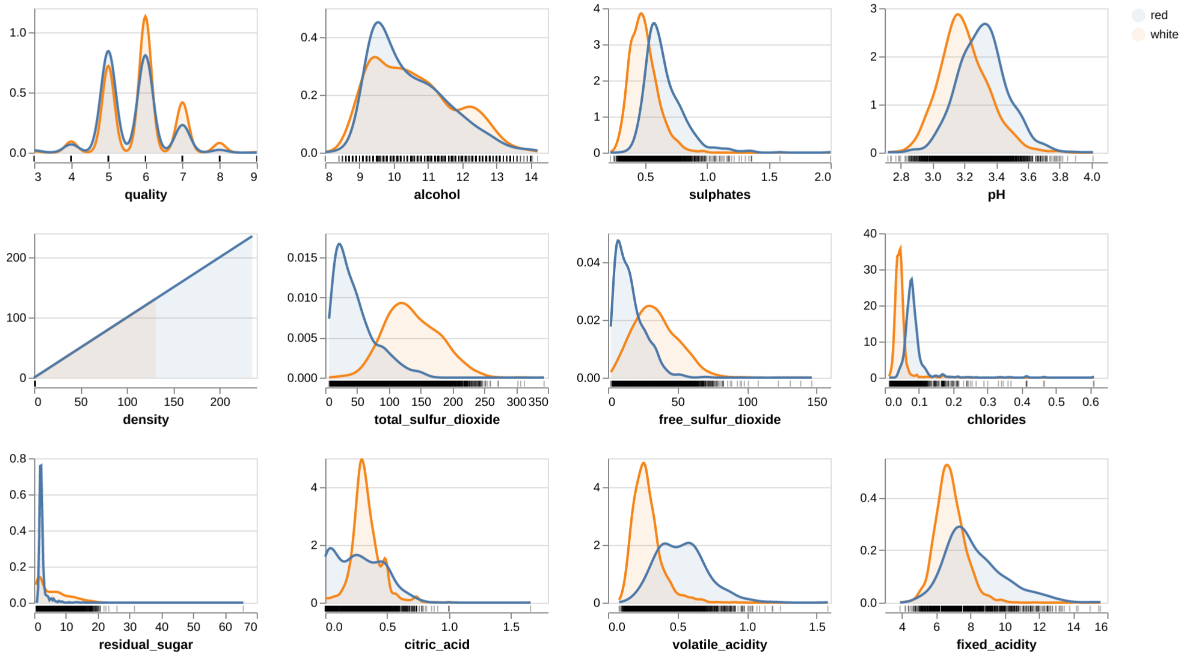

Figure 1: Distribution of Features per Target Class

We employed a logistic regression model for our classification task, utilizing 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 is shown in Table 1.

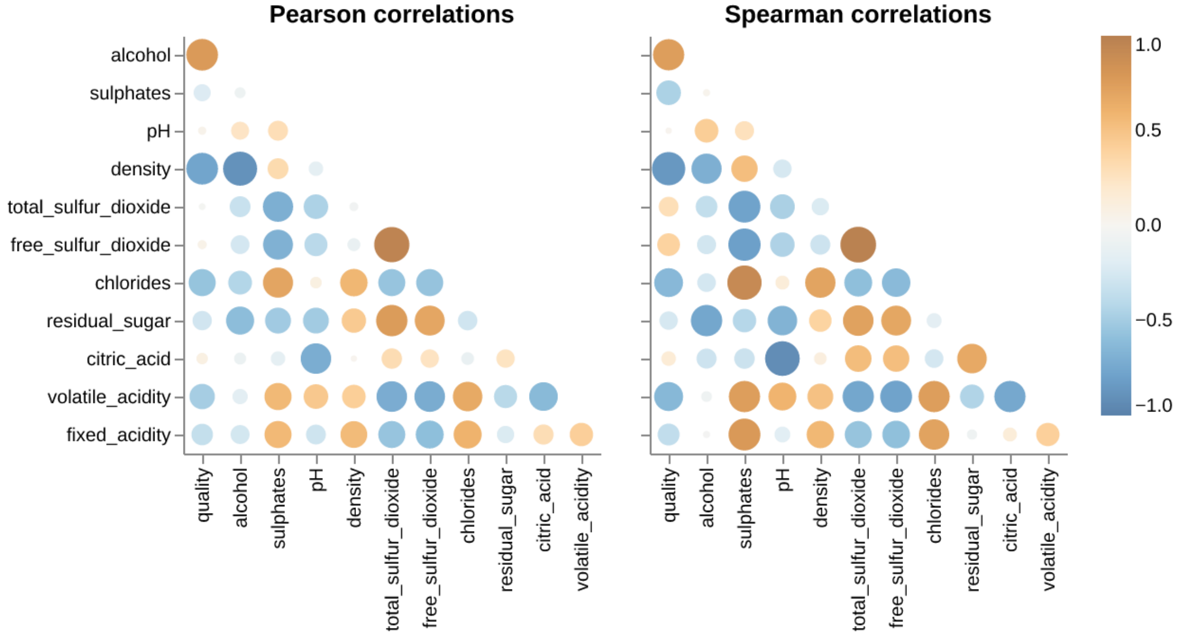

Figure 2: Correlation between Wine Color Prediction Features

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 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 model's predictions on both datasets are consistent and align 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 summarize 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 performs exceptionally well in predicting the colors of the wines as the model achieved an impressive test accuracy of 0.991. It fits our goal of accurately classifying a wine color, whether it be red or white, without human judgement. This can serve as a reliable and efficient algorithm for wine analysis as long as we have sufficient information of the wine.

Based on the information from (Figure 3) and (Figure 4), the added insights include that the model achieves excellent classification performance for wine color prediction. From the

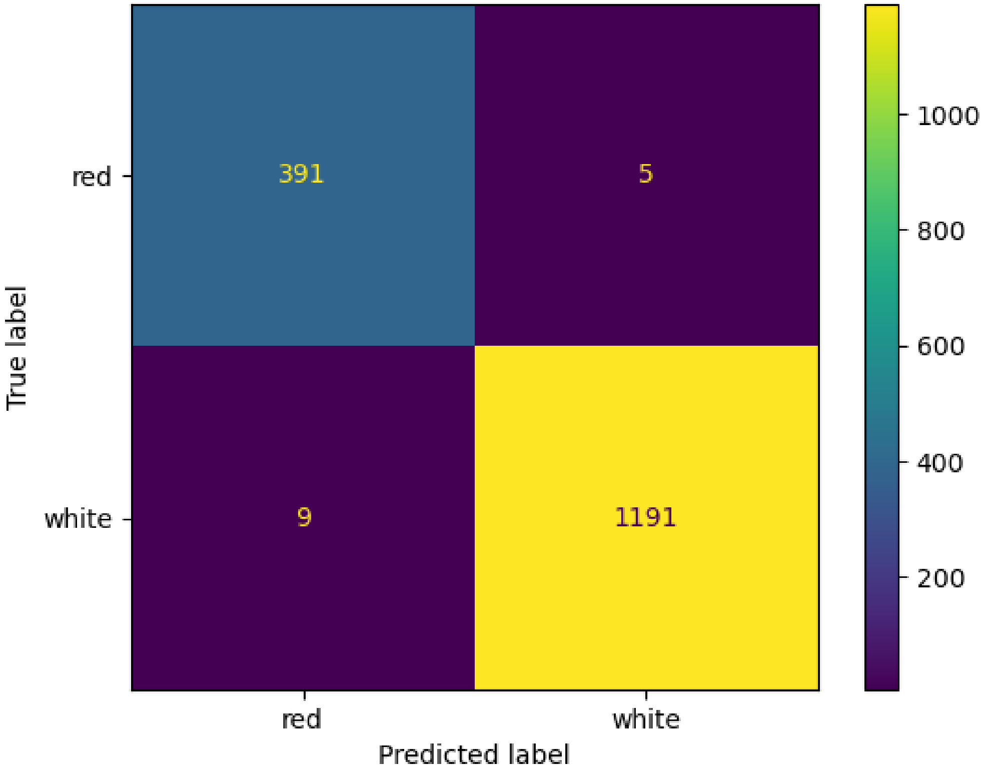

Figure 3: Confusion Matric

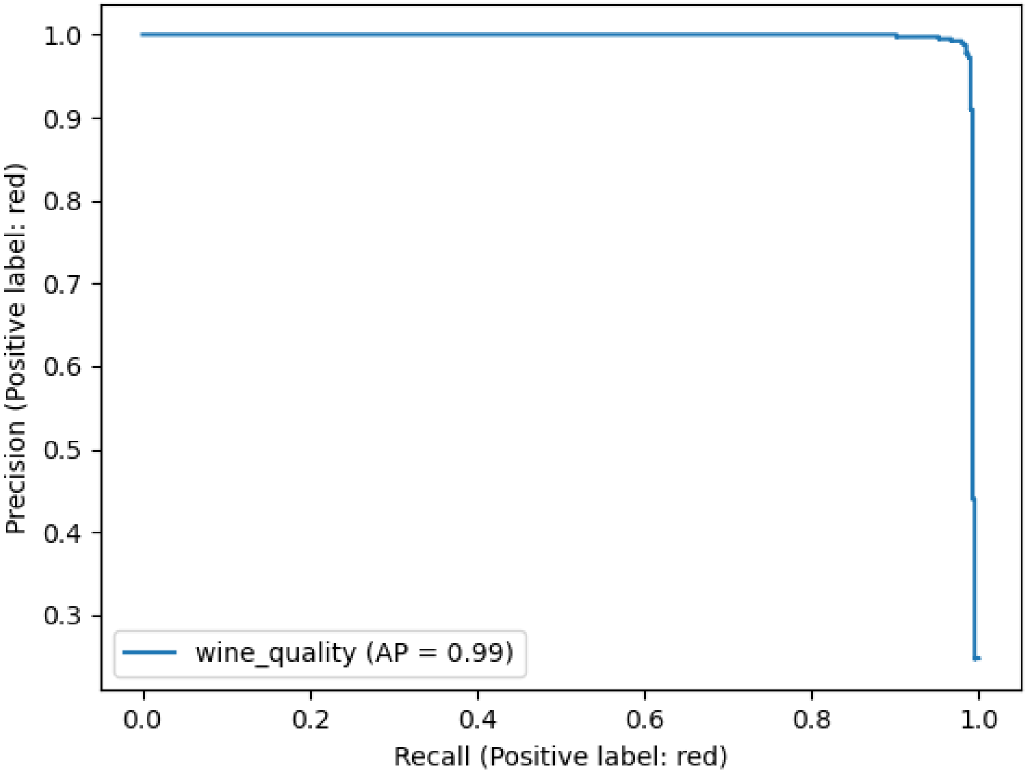

Figure 4: Precision Recall Curve

These near-perfect results on the test data exceeded our expectations, as we anticipated more variability in performance, particularly with recall and precision as we would expect a larger trade-off. With both recall and precision scores exceeding 0.98, we can conclude that the trade-off between both scores are surprisingly minimal as the model is very reliable on predicting the correct color of the wine.

Limitations

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

Moreover, the logistic regression model assumes that there is no multicolinearity within the features and the data is linearly separable. By using this model, we are ignoring the possibility that features of a wine can be dependent with one another and its subsequent effect on the wine color prediction. Also, we cannot be completely certain that all newly introduced wine data can be well separated by the linear decision rule set by the model. In fact, the existing errors from the confusion matrix (Figure 3) show that there are possible outliers that are not linearly separable by our model. Thus, human supervision over the model performance on wine analysis is still necessary since our model can still make mistakes over time.

It is crucial to comprehend how and where our model fails on the prediction. As the model has exceptional scores on the test data, we should pay even more attention to the actual errors since there might be hidden physiochemical rules in those anomalous wine data that the model fails to figure out. Thus, when we spot an error in predicting the wine color from the model in the future, we should try interpreting the feature importances by using tools like SHAP plots to investigate which properties of the wine cause the anomaly. This can help us understand what makes the wine an outlier and lead us to new insights on 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."