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) using its physiochemical properties such as acidity, pH, sugar content, and alcohol level. A logistic regression model with balanced class weights was implemented and optimized through hyperparameter tuning. The final model performed exceptionally well, achieving an accuracy of 0.98 on unseen 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 raise potential concerns about overfitting, suggesting further evaluation on truly unseen data is necessary. This work emphasizes the potential for data-driven tools to optimize wine classification processes, offering a scalable and efficient approach for the wine industry.

Introduction

Wine classification plays a crucial role in both production and quality assessment, yet traditional methods often rely on subjective evaluations by experts. Therefore, this project seeks to answer **whether we accurately predict the color 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). 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`. 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.

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.

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

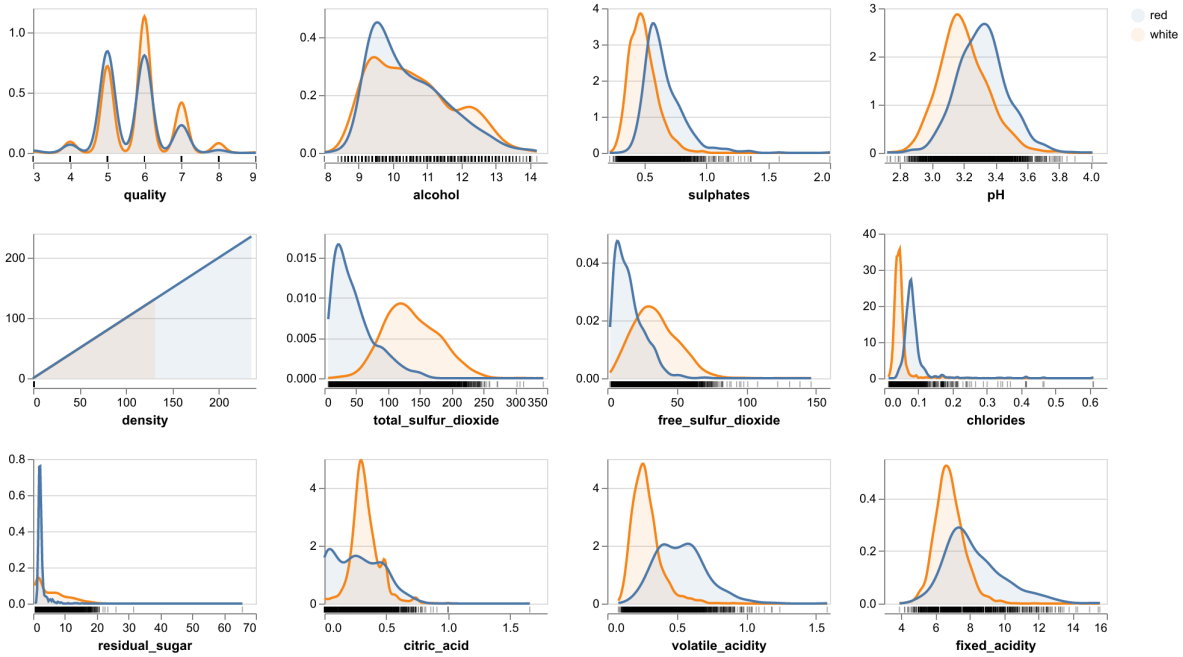

Figure 1: Distribution of Features per Target Class

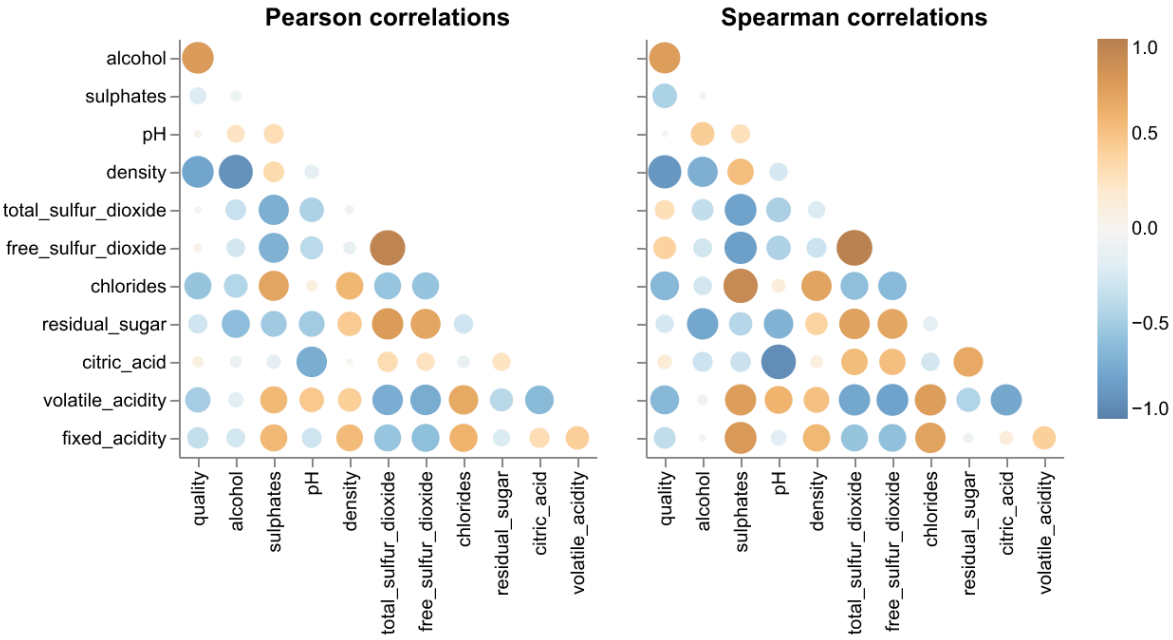

Figure 2: Correlation between Wine Color Prediction Features

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

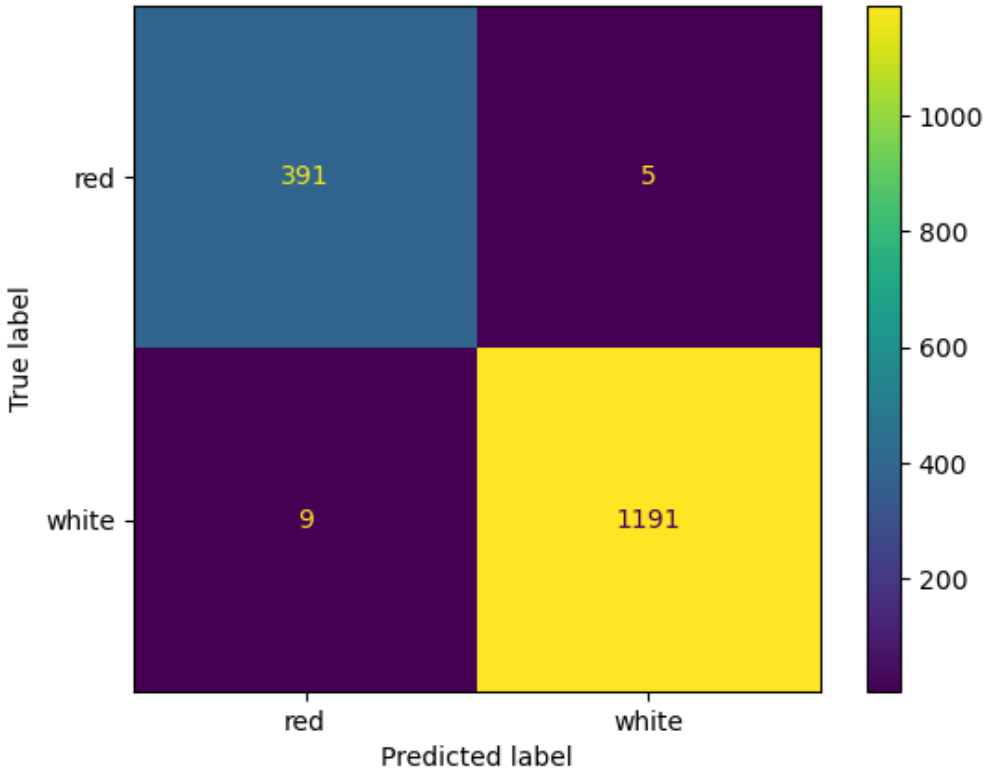

Figure 3: Confusion Matric

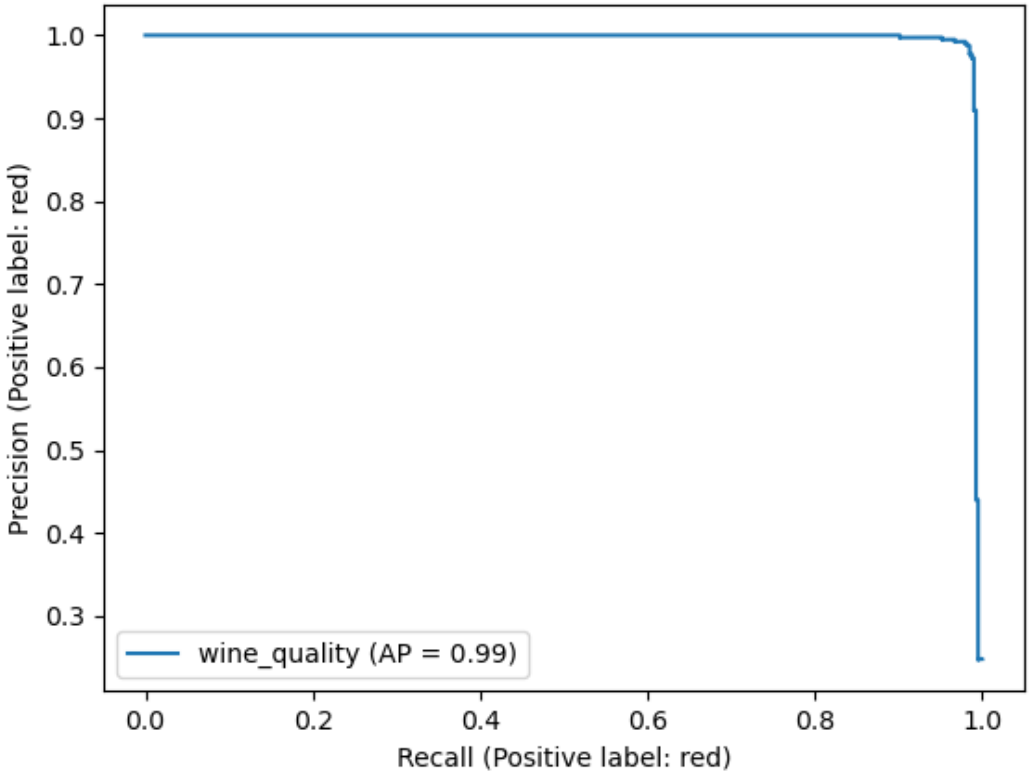

Figure 4: Precision Recall Curve

Discussion

The confusion matrix (Figure 3) confirms that the number of false positives and false negatives is low, and the model achieved an impressive test accuracy of 0.991. The precision-recall (PR) curve (Figure 4) further demonstrates the model's robust performance across various thresholds (AP = 0.99), supported by a high test precision of 0.978 and test recall of 0.987. These metrics collectively indicate that the model is highly effective at differentiating between wine colors.

The near-perfect results on the test data exceeded our expectations, as we anticipated more variability in performance, particularly with recall and precision both exceeding 0.98.

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 exceptional performance on both the training and test sets may indicate limited generalizability to truly unseen data, warranting further evaluation to ensure robustness.

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."