Milk Quality Prediction Report

Nijiati Abulizi

Kulaphong Jitareerat

February 8, 2024

Abstract

This project aims to predict milk quality using supervised machine-learning techniques. We explored various models (Logistic Regression (LR), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Naive Bayes (NB), K-Nearest Neighbors (KNN), Random Forest (RF), AdaBoost (AdaBoost), Support Vector Machine (SVM)) to classify milk based on quantitative and qualitative attributes, including pH, temperature, color, taste, odor, fat, and turbidity. Our methods involved data preprocessing, model selection, hyperparameter tuning, and performance evaluation, emphasizing model robustness and interpretability. The results indicate that certain models, particularly Random Forest, show promising accuracy and insights into feature importance for milk quality prediction.

Introduction

Supervised machine learning offers powerful tools for classification tasks, where models learn to categorize new data based on labeled examples. This has applications in diverse fields, from spam filtering to medical diagnosis. In the dairy industry, traditional milk quality assessment methods can be subjective and time-consuming. Machine learning models trained on datasets containing pH, temperature, fat content, and other quality indicators could dramatically improve this process.

Algorithms like Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest excel at finding complex patterns in data to determine appropriate classifications. This project applies supervised machine learning to the prediction of milk quality grades ("Low," "Medium," or "High"). Successful implementation could lead to faster, more consistent, and data-driven decision-making in the dairy industry. Given the complexity of milk as a biological fluid, accurate prediction of its quality is crucial for consumer safety and industry standards.

Methods

The "Milk Quality Prediction" dataset from Kaggle (1) features seven predictors (pH, temperature, taste, odor, fat, turbidity, and color) and a target variable for milk grade (Low, Medium, High), used to assess milk quality.

In our exploratory data analysis (EDA), we first examined each feature's distribution and their interrelationships, crucial for understanding the underlying patterns and informing our modeling strategy. We found no missing data in the dataset, ensuring a smooth analysis process without the need for imputation techniques.

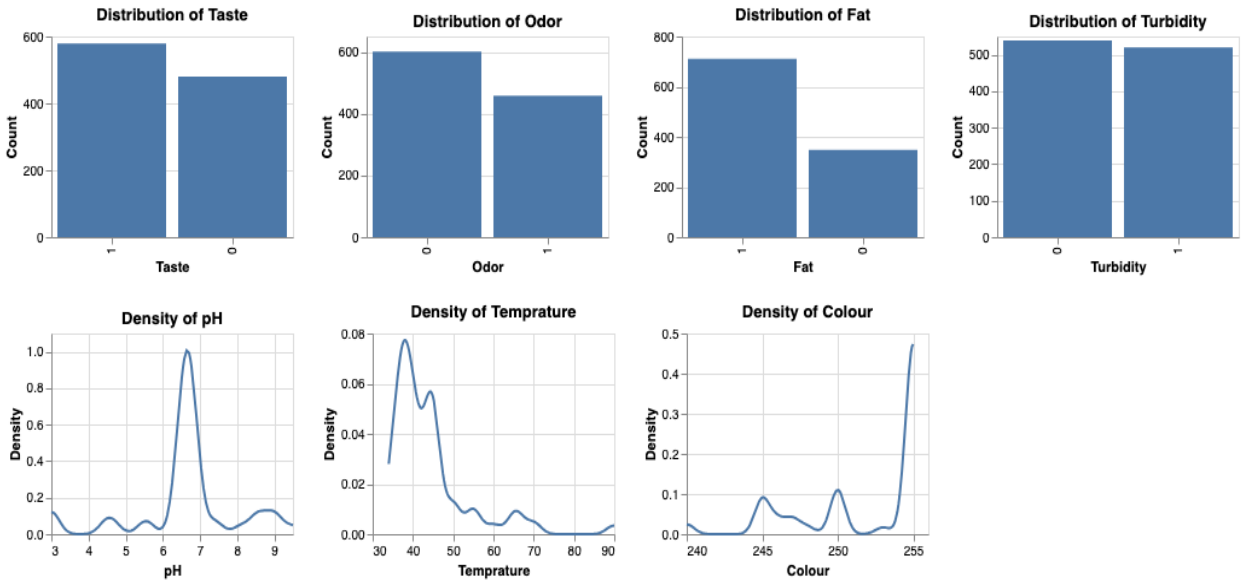

Figure 1 Histograms and density plots illustrating the distribution of taste, odor, fat, turbidity, pH, temperature, and color in a dataset, highlighting key trends and outliers.

During the EDA, as depicted in Figure 1, we observed a wide range of values across the predictors. Continuous distributions were evident for numerical variables like pH and temperature, while categorical variables, including taste, odor, fat, and turbidity, demonstrated an even distribution, suggesting a well-balanced dataset. Further Spearman (2) correlation analysis, illustrated in Figure 2, revealed significant relationships between certain features, particularly a notable correlation between turbidity and odor. This insight indicated a potential interaction affecting milk quality, underscoring the importance of careful feature selection in our predictive models that were evenly distributed, indicating a balanced dataset.

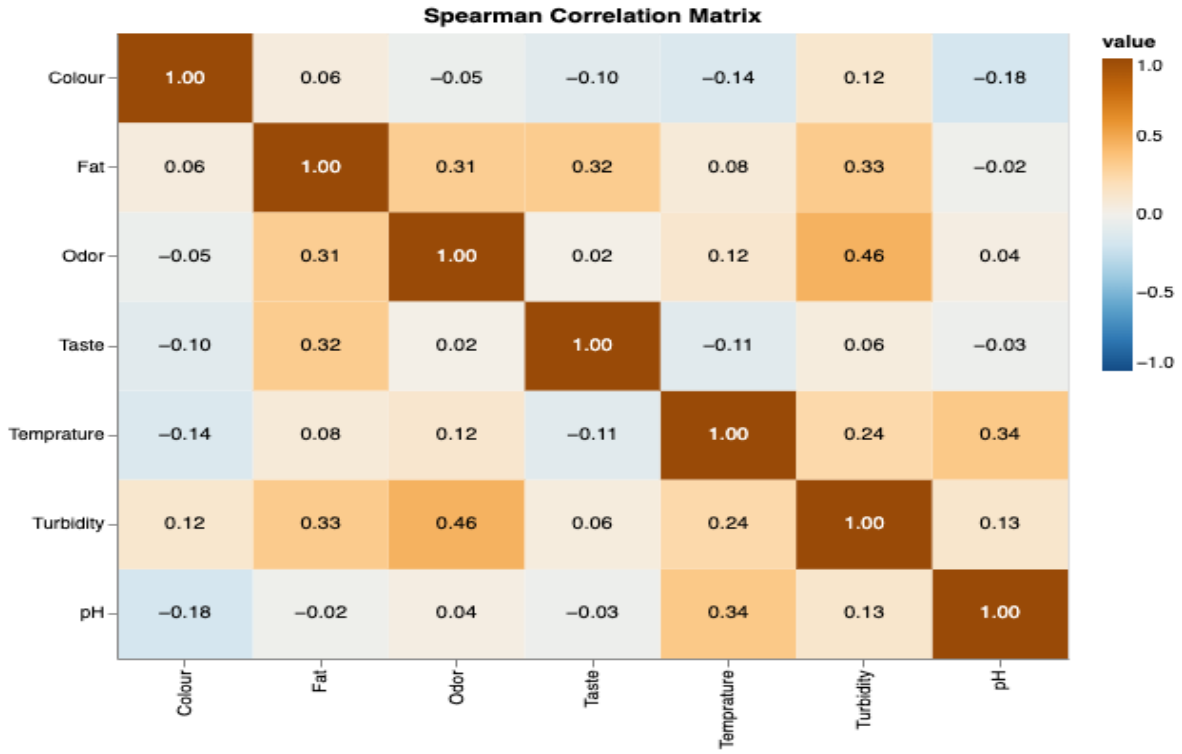

Figure 2 Spearman Correlation Matrix visualizing the relationship between sensory attributes and physicochemical properties of the dataset.

After the EDA, we split the dataset into 80% for training and 20% for testing, then standardized features with the Standard Scaler for models like KNN and SVM. We trained various models (LR, LDA, QDA, Naive Bayes, KNN, Random Forest, AdaBoost, SVM), optimizing them with 5-Fold Cross-Validation and Grid Search, focusing on the F1-Score for evaluation due to our dataset's imbalance. Following hyperparameter optimization, we re-trained the models on the entire training dataset.

In the final evaluation of the test dataset, we assessed model accuracy in predicting milk quality. To tackle multicollinearity in models like LR, LDA and QDA, we applied PCA to reduce feature dimensions while preserving 93% of the variance, enhancing model performance and ensuring accurate milk quality predictions.

Experiment

Following the methods outlined, we started the experimental phase, where we applied the data preprocessing steps to both the training and testing datasets. We explored the hyperparameter spaces for each model, using the strategies developed during the methodology phase. We conducted hyperparameter tuning through Grid Search (3), informed by the cross-validation (4) results to select the best parameters for each model, and the results are shown in Table 1 below.

Table 1 Comparative performance metrics of machine learning models with best hyperparameters on training and test datasets.

model	score_cv	score_train	score_test	best_params
K-Nearest Neighbors	0.998962	1.000000	0.994492	{'n_neighbors': 21, 'p': 1, 'weights': 'distance'}
Random Forest	0.998962	1.000000	0.994492	{'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}
Support Vector Machine	0.996550	1.000000	0.994492	{'C': 1, 'gamma': 1, 'kernel': 'poly'}
Quadratic Discriminant Analysis (PCA 93%)	0.983481	0.993745	0.986466	{'reg_param': 0.0}
NaiveBayes	0.925870	0.931992	0.916258	{'var_smoothing': 0.1}
AdaBoost	0.917615	0.917543	0.903386	{'learning_rate': 0.01, 'n_estimators': 100}
Logistic Regression	0.851933	0.848336	0.839155	{'C': 10, 'max_iter': 50, 'penalty': 'l1', 'solver': 'liblinear'}
Linear Discriminant Analysis	0.750690	0.728673	0.701511	{'shrinkage': None, 'solver': 'svd'}

After training, we conducted an evaluation using the testing set, emphasizing the F1-Score to assess model performance. The evaluation shown in Figure 3 demonstrated that Random Forest and K-Nearest Neighbors (KNN) achieved the highest F1-Scores, signifying their good performance in accurately predicting milk quality grades. Both models scored close to 1 on the testing set, indicating near-perfect precision and recall. Support Vector Machine (SVM) and Quadratic Discriminant Analysis (QDA) also showed strong performance, with F1-Scores just above 0.9 on testing.

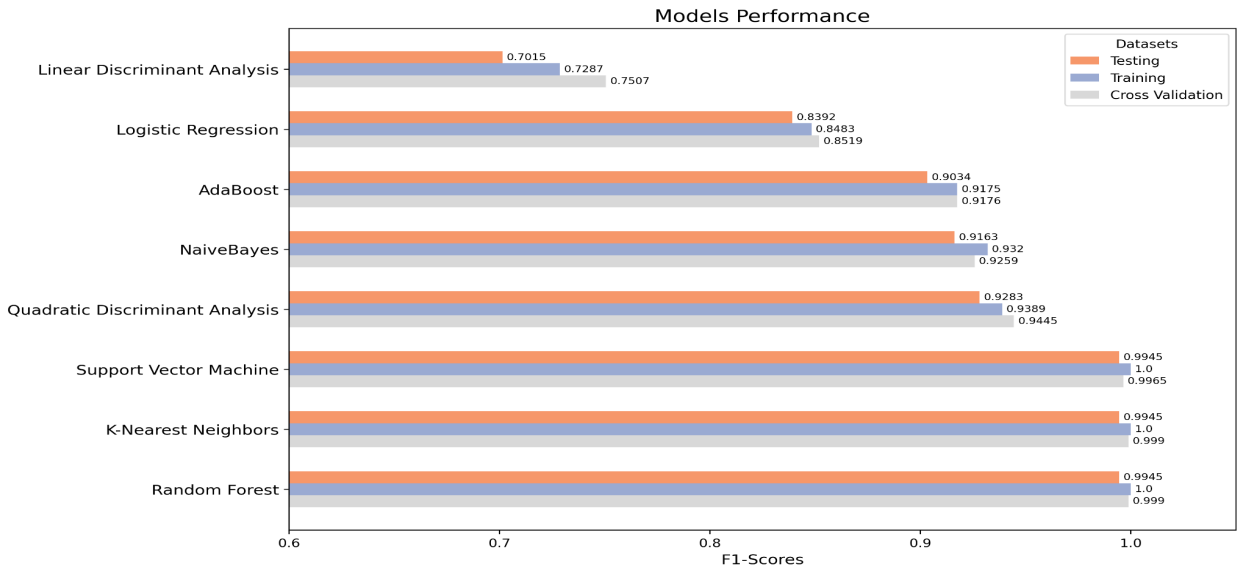

Figure 3 Bar chart comparison of F1-Scores for various machine learning models across training, testing, and cross-validation datasets.

In contrast, models like LDA and LR exhibited relatively lower F1 scores, suggesting that they may be less effective for this task. The performance of AdaBoost and Naive Bayes was moderate, placing them between the higher- and lower-performing models.

These results allow us to conclude that, for the milk quality prediction task, ensemble methods like Random Forest and distance-based methods like KNN are most effective, while other models might require further parameter tuning or may inherently be less suited to the dataset's characteristics. The high performance of the leading models on both the training and testing sets also suggests good generalization without overfitting.

Additionally, we addressed the challenge of multicollinearity in models such as LR, LDA and QDA by implementing PCA as shown in Figure 4. This step was crucial for reducing feature dimensions while retaining significant variance (93%), thereby enhancing QDA model performance by mitigating the effects of collinearity among features. We also observed that PCA

successfully improved the performance of our QDA model as the missing in multicollinearity. However, LR and LDA experienced slightly decreased performance, suggesting that certain information relevant to these models may have been lost in the variance reduction process.

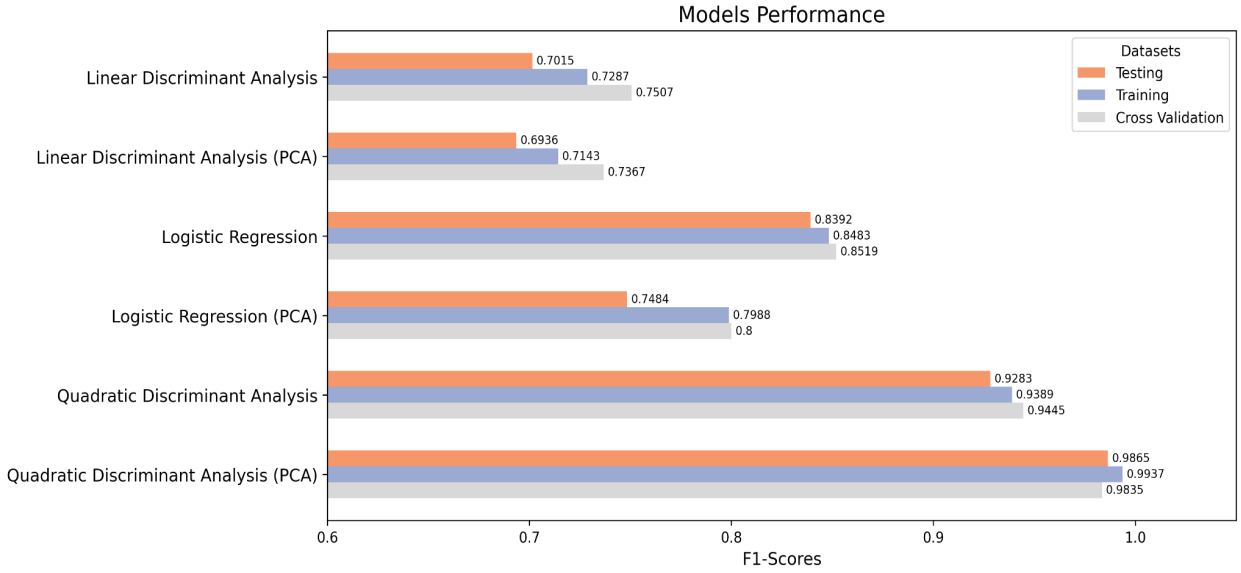

Figure 4 Horizontal bar chart showing F1-Scores for machine learning models with and without PCA, evaluated on testing, training, and cross-validation datasets.

Insights from the Random Forest Model

As depicted in Figure 5, pH emerged as the most dominant factor in milk quality prediction, aligning with its known impact on bacterial growth and spoilage. Temperature also proved significant, highlighting its influence on bacterial activity and chemical changes within milk. Conversely, sensory features like turbidity, fat, odor, and taste, as well as color, showed lower importance. This suggests a potentially weaker correlation between these factors and milk quality within the specific context of this dataset.

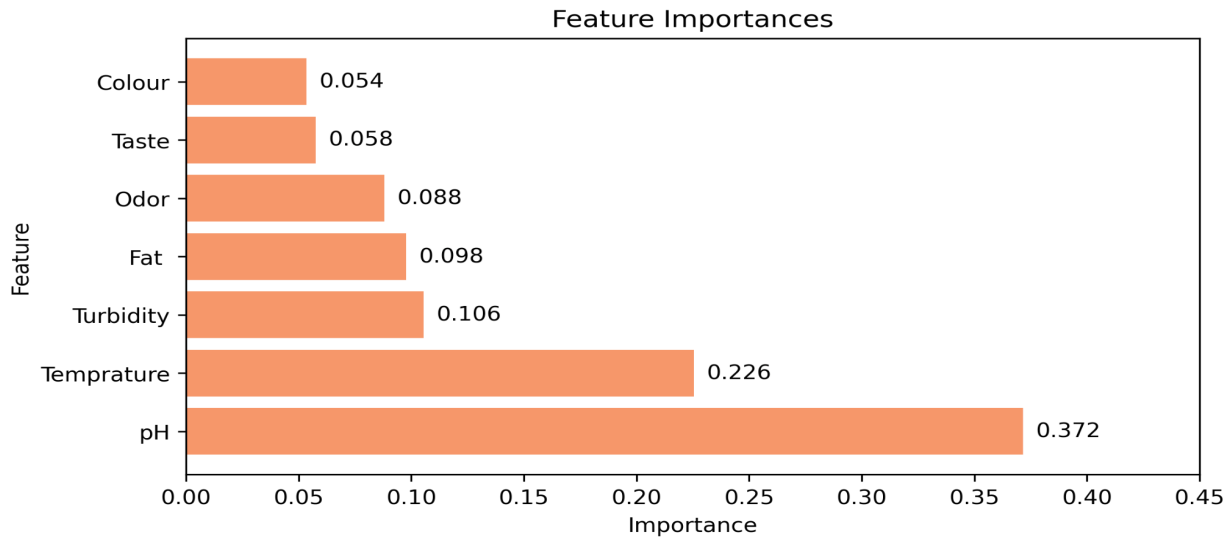

Figure 5 Bar chart of feature importance showing the relative influence of each variable, with pH and temperature being the most significant predictors.

Conclusion

This project highlights the potential of supervised machine learning for accurate milk quality classification. Random Forest and K-Nearest Neighbors proved exceptionally effective, achieving F1 scores near 0.99. Their interpretability and efficiency suggest suitability for real-world use. Random Forest's feature importance analysis underscores the significance of pH and temperature in milk quality, offering actionable insights for the dairy industry. While future work can address limitations, this study demonstrates the value of data-driven approaches for milk quality assessment. Additional research on model refinement and integration into practical workflows could significantly optimize efficiency within the dairy sector.

Reference

1. SHRIJAYAN RAJENDRAN (n.d.). Data source: Milk Quality Data. Kaggle. , from https://www.kaggle.com/datasets/cpluzshrijayan/milkquality

2. SciPy developers. (n.d.). scipy.stats.spearmanr: Spearman rank-order correlation coefficient. SciPy v1.8.0 Reference Guide. , from https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

3. Scikit-learn developers. (n.d.). GridSearchCV: Exhaustive search over specified parameter values for an estimator. scikit-learn. , from https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

4. Scikit-learn developers. (n.d.). Cross-validation: Evaluating estimator performance. scikit-learn. , from https://scikit-learn.org/stable/modules/cross_validation.html

Contribution of group members

1. Nijiati Abulizi
2. Kulaphong Jitareerat

All members of the group contributed equally to every stage of this research study. This effort included the initial exploratory data analysis, the brainstorming and execution of model hyperparameter tuning strategies, the training and testing of the predictive models, and the critical discussions leading to the final model selection. Further, the team engaged in the in-depth analysis of the results, the creation and delivery of the presentation materials, and the process of writing and revising the final draft.