A Knowledge Discovery in Databases (KDD) Approach to Wine Quality Classification

Siri Batchu
Department of Software Engineering
San Jose State University

October 21, 2024

Abstract

This paper presents an application of the Knowledge Discovery in Databases (KDD) methodology on the Wine Quality dataset, which contains chemical properties of various wines and their corresponding quality ratings. The KDD process involves data selection, preprocessing, transformation, data mining, and evaluation. The goal is to predict wine quality using machine learning models, specifically a Random Forest Classifier. The model achieved high accuracy, confirming its effectiveness in predicting wine quality based on chemical attributes.

1 Introduction

The field of data mining is concerned with extracting useful patterns from large datasets. The Knowledge Discovery in Databases (KDD) process is a structured framework that includes several key stages: data selection, preprocessing, transformation, data mining, and interpretation. In this study, we apply the KDD process to the Wine Quality dataset, which is widely used to classify and predict the quality of wine based on a set of chemical attributes.

Wine quality is subjective and traditionally assessed by human experts; however, machine learning offers an objective and repeatable way to predict wine quality. This paper explores the application of the Random Forest algorithm for classifying wine based on its chemical properties, following the KDD framework.

2 Methodology

2.1 Data Selection

The dataset contains 12 variables, including fixed acidity, volatile acidity, citric acid, and the target variable extitquality, which rates wines on a scale from 0 to 10. This study used the entire dataset without sampling because of its manageable size.

2.2 Data Preprocessing

Data preprocessing involved checking for missing values and removing irrelevant features. The dataset had no missing values, ensuring that no imputation was necessary. An irrelevant column, extitId, was dropped as it did not contribute to the predictive analysis.

2.3 Data Transformation

The data was divided into features (X) and the target variable (y). A train-test split was performed, using 70% of the data for training and 30% for testing. The features were standardized using extitStandardScaler to ensure that they had the same scale, which helps in improving the performance of machine learning models.

2.4 Data Mining

The data mining step involved building a classification model using the Random Forest Classifier. This algorithm was chosen for its ability to handle complex relationships between features and its robustness against overfitting. The model was trained on the training set, and predictions were made on the test set.

2.5 Interpretation and Evaluation

The model's performance was evaluated using accuracy, precision, recall, and the F1-score. Additionally, a confusion matrix was generated to assess the number of correct and incorrect classifications made by the model across different wine quality categories.

3 Results and Discussion

3.1 Data Exploration

A preliminary exploration of the dataset revealed some interesting patterns. Descriptive statistics showed that most wines were rated between 5 and 6 on the quality scale, with relatively few wines rated as either very poor (0-3) or excellent (8-10). The dataset's chemical features, such as extitalcohol, extitsulphates, and extitvolatile acidity, showed significant variation across wines, suggesting their potential impact on wine quality.

3.2 Model Performance

The Random Forest Classifier achieved an accuracy of 73% in predicting wine quality on the test set. The classification report in Table 1 shows the precision, recall, and F1-score for each quality category.

Table 1: Classification Report for Wine Quality Prediction

Quality Rating	Precision	Recall	F1-Score
3	0.80	0.25	0.38
4	0.60	0.29	0.39
5	0.73	0.91	0.81
6	0.74	0.78	0.76
7	0.65	0.43	0.52
8	1.00	0.25	0.40

The confusion matrix (Figure **??**) shows that the model performed well in predicting wines with quality ratings of 5 and 6, which constitute the majority of the dataset. However, the model struggled to classify wines rated 3, 4, and 8, likely due to the smaller number of examples of these ratings in the dataset.

3.3 Discussion

The high accuracy and F1-scores for the majority wine quality categories (5 and 6) indicate that the Random Forest Classifier can successfully predict common wine qualities. However, the model's performance drops for less frequent categories, such as 3 and 8, which highlights the challenge of class imbalance in classification tasks.

The Random Forest's ability to handle both continuous and categorical variables made it an appropriate choice for this dataset. The relatively low precision and recall for rarer classes suggest that further tuning or sampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) could help improve the model's performance in these cases.

4 Conclusion

In this study, the KDD process was applied to the Wine Quality dataset to build a predictive model using the Random Forest Classifier. The model achieved a strong performance for the most common quality ratings, while performance on rarer categories was weaker. This suggests that while the Random Forest algorithm is effective for wine quality prediction, handling class imbalance should be a focus for future work.

Future research could explore other classification models such as Gradient Boosting or XGBoost, as well as applying techniques for balancing the dataset, such as oversampling or undersampling. Additionally, incorporating domain knowledge from oenology (the science of wine) could help improve the feature selection process and result in more accurate predictions.

References

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4), 547-553.

[2] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

[3] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.