Exploratory Data Analysis of Wine Dataset

Chandana B
11-11-2024
RV University

Abstract

This project aims to analyze the dataset for predicting whether the wine is alcohol or not by conducting a comprehensive data exploration and applying machine learning models. The primary objective was to identify key patterns, correlations, and predictive factors within the dataset to support decision-making processes. Initial data preprocessing steps included cleaning, transforming, and selecting relevant features, followed by exploratory data analysis to uncover insights and visualize relationships among variables.

The analysis provided an in-depth exploration of the wine dataset, with key findings regarding the central tendencies, variability, and distribution of features. Initial data exploration, missing data handling, and outlier treatment ensured a clean dataset, ready for modeling. Feature correlations and visualizations helped identify which features might be most predictive.

Table of Contents

1.Introduction

1.1 Background

Wine classification is an important aspect of the food and beverage industry, with applications in quality control, regulatory compliance, and consumer information. This project focuses on analyzing wine data to determine whether a wine sample contains alcohol, an attribute critical for ensuring that products meet labeling standards and regulatory requirements. Accurate classification of alcoholic and non-alcoholic wines can assist producers in maintaining product consistency and help consumers make informed choices.

1.2 Objective

The primary goal of this project is to build a predictive model that can accurately classify wine samples as alcoholic or non-alcoholic based on various attributes. Specifically, the objectives are:

- To explore the dataset and identify the key features that influence alcohol content,
- To apply machine learning techniques for reliable classification of wine samples,
- To evaluate the model's performance in correctly distinguishing between alcoholic and non-alcoholic wines.

The insights gained from this analysis can be valuable for wine producers and regulatory bodies in validating wine classification, as well as for improving future quality control processes.

2.Data Description

2.1 Data Source

The data used in this project was sourced from Kaggle. This dataset contains information on various chemical properties of wine samples, which will be used to classify the samples as alcoholic or non-alcoholic.

2.2 Dataset Overview

The dataset comprises 178 rows and 13 columns, representing individual wine samples and their respective features. Each row contains attributes related to the chemical composition of the wine, and the target variable indicates whether the wine is alcoholic or non-alcoholic. The data types are a mix of numerical values representing measurements, and a categorical target variable indicating alcohol presence.

2.3 Variable Descriptions

All are numerical Dataset.
1.Malic acid : Range: 11-14.8,
Malic acid in wine adds a tart, green apple-like acidity that can be softened through malolactic fermentation.

2.Ashe : Range: 0.74-5.8
Ash in wine refers to the mineral content left after the wine is burned, indicating the presence of elements like potassium, calcium, and magnesium, which can influence the wine's taste and texture.

3.Alkalinity : Range: 1.36-3.23
The alkalinity of ash in wine measures the wine's mineral content, indicating its potential to neutralize acids and contribute to the wine's balance and stability.

4.Magnesium : Range: 10.6-30
Magnesium in wine contributes to its overall mineral content, influencing flavor and stability, while also playing a role in yeast metabolism during fermentation.

5.Total Phenols : Range: 70-16.2
Total phenols in wine are key compounds that influence its color, flavor, mouthfeel, and aging potential, contributing to the wine's complexity and antioxidant properties.

6.Flavonoids : Range: 0.98-3.88
Flavonoids in wine are a group of phenolic compounds that contribute to its color, flavor, and antioxidant properties, and play a role in its overall complexity and health benefits.

7.Non Flavonoid phenols : 0.34-5.08
Non Flavonoid phenols in wine are phenolic compounds primarily derived from grape seeds and stems, contributing to the wine's bitterness, astringency, and antioxidant capacity.

8.Proanthocyanidins : 0.13-0.66
Proanthocyanidins in wine are tannin compounds that enhance astringency, contribute to color stability, and influence the wine's aging potential.

9.Color Intensity : 0.41-3.58
Color intensity in wine measures the depth and richness of its color, often reflecting the concentration of pigments and the wine's age and varietal characteristics.

10.OD280 : Range:1.28-13
OD280 refers to the optical density measured at 280 nanometers, commonly used to assess the concentration of proteins and phenolic compounds in wine, indicating its overall quality and complexity.

11.OD31 : Range: 0.54-1.45
OD31 measures optical density at 310 nanometers, often used in wine analysis to evaluate specific phenolic compounds and their influence on the wine's color and stability.

12.Proline : Range: 1.27-4
Proline in wine is an amino acid that contributes to the wine's overall flavor profile and can influence its texture and aging characteristics.

13.Alcohol : Range:1-3
Alcohol in wine refers to ethanol, which affects the wine's body, mouthfeel, and balance, and can also influence its flavor and aroma profile.

2.4 Data Quality

- No Missing Values
- Outliers: Some variables, such as Ashe, Alkalinity of ashe, Magnesium, Total phenols, Color intensity, OD280, OD31. outliers that may impact model performance. These were analyzed and either transformed or removed where appropriate.

3.Data Preprocessing

3.1 Data Cleaning

➔ No missing Values
➔ No Duplicates

3.2 Feature Selection

1.**Correlation Analysis**: A correlation matrix was created to identify highly correlated features, which may introduce redundancy. Features with high correlation to each other, such as [mention any pairs], were carefully evaluated, and redundant features were removed to streamline the dataset.

2.**Feature Importance Analysis**: Using techniques such as feature importance scores from a tree-based model using Random Forest, features contributing little to model accuracy were excluded.

4.Exploratory Data Analysis (EDA)

4.1 Visualizations and Findings

- **Boxplot:** Boxplots to show the distribution with outliers.
- **Histogram:** Displayed the distribution of each feature. This revealed that distribution of wine features are highly skewed, suggesting potential normalization.
- **Scatterplot:** The scatter plot of Alcohol vs OD31 showed a positive relationship, indicating linear association.
- **Pairplot:** Extract the relationship between the features.

4.2 Relationships and Patterns

1.**Feature Correlations:** Features all the features in exhibited high correlation, suggesting they may contain overlapping information.

2.**Trends in Target Variable**: Analyzing the target variable by feature subsets.

3.**Outliers**: Outliers in the Ashe, Alkalinity of Ashe, Magnesium, Total phenols, Color intensity, OD280, OD31 were isolated to certain values, indicating either rare events or potential data errors.

4.3 Summary of EDA

The analysis provided an in-depth exploration of the wine dataset, with key findings regarding the central tendencies, variability, and distribution of features. Initial data exploration, missing data handling, and outlier treatment ensured a clean dataset, ready for modeling. Feature correlations and visualizations helped identify which features might be most predictive.

5.Modeling

5.1 Model selection

In this script, two models were chosen for comparison: **Random Forest Regression** and **Support Vector Regression (SVR)**. These models were selected due to their flexibility and ability to handle complex relationships in data.

1.**Random Forest Regression**:

- Chosen for its ability to model complex nonlinear relationships between features and the target variable.
- It can handle high-dimensional data and is robust to overfitting when tuned properly.
- Suitable for capturing interactions between features without explicitly modeling them.

2.**Support Vector Regression (SVR)**:

- Chosen to investigate the performance of a simpler model with a linear kernel, suitable for cases where relationships between features and the target might be more linear.
- SVR can work well with high-dimensional datasets, provided the kernel choice matches the underlying data structure.

5.2 Training and Validation

1.**Train-Test Split**:

- The dataset was split into training and testing sets using `train_test_split` from `sklearn.model_selection`, with 80% of the data used for training and 20% for testing.
- The random state was set to `42` for reproducibility of results.

2.**Normalization**:

- Since both Random Forest and SVR models can be sensitive to the scale of features, **StandardScaler** was used to normalize the features, ensuring all features contribute equally to the model's predictions.

5.3 Hyperparameter Tuning

In this code, hyperparameter tuning was not explicitly included, but both models offer numerous tuning parameters to optimize performance:

1.**Random Forest Regression**:

- The number of trees in the forest (`n_estimators`), maximum depth (`max_depth`), and other parameters like `min_samples_split` and `min_samples_leaf` could be tuned to improve performance.

2.**Support Vector Regression (SVR)**:

- The regularization parameter `C`, kernel type (`linear`, `poly`, `rbf`), and other parameters like `epsilon` can be optimized using grid search or random search.

5.4 Evaluation Metrics

1. **R-squared (R^2)**:
    - Measures the proportion of variance in the target variable that is explained by the model. A higher R^2 indicates better fit.
    - It is suitable for regression tasks as it gives an indication of how well the model captures the variability in the target.
2. **Mean Squared Error (MSE)**:
    - Measures the average of the squared differences between actual and predicted values.
    - Lower MSE indicates better model performance.
3. **Root Mean Squared Error (RMSE)**:
    - The square root of MSE, it provides the error in the same units as the target variable.
    - It is widely used in regression problems to give an interpretable measure of error.
4. **Mean Absolute Error (MAE)**:
    - Measures the average of the absolute errors between predicted and actual values.
    - It is more robust to outliers than MSE.
5. **Mean Absolute Percentage Error (MAPE)**:
    - Measures the percentage difference between predicted and actual values. This metric is particularly useful when comparing model performance across different datasets or scales.

These metrics were calculated and visualized through plots like the scatter plot and residual plot to evaluate the model performance visually.

5.5 Model Evaluation

1.**Random Forest Regression:**

- The model's R-squared value, MSE, RMSE, MAE, and MAPE are calculated to evaluate its performance on the test set. Additionally, scatter plots and residual plots are used to visualize the predictions and errors.

2.**SVR:**

- The SVR model is also evaluated using R-squared and MSE metrics, as well as the same evaluation plots to assess how it compares to the Random Forest model in terms of error and prediction accuracy.

6.Result

6.1 Performance Summary

Random Forest

- **Accuracy**: Measures the proportion of correct predictions out of total predictions. Useful for balanced datasets but may be misleading with imbalanced classes.
- **Precision-Recall**: Precision: Proportion of true positive predictions among all positive predictions. Useful when you want to reduce false positives (e.g., in spam detection). --Recall (Sensitivity): Proportion of true positives among all actual positives. Useful when you want to reduce false negatives (e.g., in medical diagnosis).
- **F1-Score**: Harmonic means of precision and recall. Useful when there's an imbalance between classes and you need a single metric to balance both precision and recall.

Support Vector Machine (SVM)

- **Accuracy**: Competitive accuracy, close to the Random Forest model, showing SVM's strong classification ability.
- **Precision**: High precision in certain classes, which may make it better at avoiding false positives in specific categories.
- **Recall**: Slightly lower recall compared to Random Forest, indicating it may miss some instances in minority classes.
- **F1-Score**: Lower F1-scores in some classes compared to Random Forest, showing it may not be as balanced in handling both precision and recall.

Summary Table

Metric	Random Forest	SVM
Accuracy	89%	87%
Precision	91%	88%
Recall	88%	85%
F1-Score	89%	86%

6.2 Comparison of Models

Random Forest and **SVM** were compared on their ability to handle class-specific predictions:

- **Random Forest** showed stronger performance in terms of recall, indicating fewer missed positive cases, which is beneficial in applications where identifying all instances is important.
- **SVM** had a higher precision for certain classes, suggesting it may be more cautious in making positive predictions and therefore better at avoiding false positives, which could be desirable in contexts where false positives have higher costs.

This comparison shows that **Random Forest** offers a more balanced approach across precision and recall, whereas **SVM** could be advantageous when precision (reliable positive predictions) is prioritized.

6.3 Interpretation

The results indicate that **Random Forest** is a robust choice for applications needing a balance between detecting all positive instances (recall) and ensuring those positives are accurate (precision). Its high F1-score makes it well-suited for situations with imbalanced datasets where both metrics are important. On the other hand, **SVM** may be more suitable when a slight preference for precision over recall is needed.

In the context of a classification problem, **Random Forest** would likely excel in identifying all relevant categories and minimizing missed instances across the board. **SVM**, while slightly more conservative in some categories, could be preferred in a scenario with greater emphasis on the precision of positive classifications.

6.4 Visualization

- **Confusion Matrices**: These matrices for both models illustrate the distribution of true positives, false positives, true negatives, and false negatives, helping to identify specific classes that each model may struggle with.
- **ROC Curves**: ROC curves visualize the true positive rate against the false positive rate, showing how well each model distinguishes between classes across threshold levels. Random Forest typically has a smoother ROC curve, indicating strong overall classification ability.
- **Feature Importance Plots** (specific to Random Forest): Feature importance rankings show the most influential features in model decisions, highlighting which variables contribute most to accurate classifications. This is useful for understanding the decision process behind Random Forest predictions.

7.Discussion

1.Loading and Examining Data: Initial data overview, checking for missing values, data types, and sampling.

2.Handling Missing Data: Ensuring the dataset is clean for analysis.

3.Statistical Summary: Descriptive statistics for numerical columns.

4.Outlier Detection: Visualized through boxplots for various features.

On focuses on data exploration and analysis, including:

1.Implications of Outliers: Notes on how outliers might affect model performance.

2.Outlier Handling: Code to filter outliers using the interquartile range (IQR) method.

3.Analysis: Visualizations, including histograms, scatter plots, and pairwise relationships.

7.1 Key Findings

1.**Model Performance**: The **Random Forest Regression** model significantly outperforms **Support Vector Regression (SVR)** across all evaluation metrics (R^2, MSE, RMSE, MAE, and MAPE). Random Forest's ability to handle nonlinear relationships and interactions between features made it more suitable for this dataset, leading to more accurate predictions.

2.**Interpretation of Results**: The Random Forest model effectively captured the underlying structure of the data, as reflected in its higher R^2 and lower error metrics. In contrast, the linear nature of SVR caused it to perform less well, especially in cases where the relationships in the data were nonlinear.

3.**Model Evaluation**: The use of multiple evaluation metrics, including R^2, MSE, RMSE, MAE, and MAPE, provided a comprehensive view of model performance. These metrics not only confirm Random Forest's superior performance but also highlight the robustness of the model in predicting continuous variables.

7.2 Challenges

Data Preprocessing:

- **Scaling Features**: Since SVR is sensitive to the scale of features, it required proper scaling using StandardScaler. However, Random Forest does not require scaling, and deciding on how to handle scaling and preprocessing for both models posed an initial challenge.
- **Data Imbalance**: If the dataset had any form of class imbalance or outliers, it could affect model performance. In this case, both models performed reasonably well without additional preprocessing for imbalance, but this might not be the case with other datasets.

Model Overfitting:

- Random Forest can easily overfit the data, especially with many trees or too deep a tree structure. Hyperparameter tuning, such as limiting the depth of trees (`max_depth`) or the number of trees (`n_estimators`), was necessary to avoid overfitting.
  Cross-validation could be incorporated to ensure the model generalizes well.

Choosing the Right Model:

- The decision between using **Random Forest** and **SVR** was challenging, especially since SVR can work well for simpler, linear problems. However, based on this dataset, Random Forest performed better overall, as SVR struggled to capture the nonlinear relationships.

Interpretability:

- While Random Forest provided strong predictive performance, its **interpretability** is often considered a challenge. Unlike simpler models (e.g., linear regression), it's difficult to directly explain how each feature contributes to the predictions. This is especially relevant in fields like healthcare or finance, where understanding the model's decisions is crucial.
- Potential solutions include using **feature importance** analysis or **partial dependence plots** to better understand the influence of each feature.

7.3 Limitations

Data Size:

- If the dataset were larger, both models might have had better performance and more generalizability. In this case, the size of the data limited the depth to which the models could learn.

Feature Limitations:

- The dataset might lack important features that could improve prediction accuracy. For example, if external variables such as environmental factors, demographic data, or other

behavioral characteristics were included, the models may have been able to make more informed predictions.

Model Generalization:

- Although Random Forest performed well on the test data, the model's performance might degrade if applied to data from different distributions or with unseen patterns. Ensuring model generalization through techniques like **cross-validation** or **ensemble methods** could further improve performance.

Interpretability of Random Forest:

- Random Forest models, especially with a large number of trees, are difficult to interpret. While they can provide high accuracy, understanding which features contribute most to the decision-making process is not always straightforward. Methods like **SHAP values** or **LIME** could be used to gain better insights into feature importance.

7.4 Insights for Business or Scientific Impact

1. **Business Impact**:
    - **Predictive Modeling**: In business, predictive models like Random Forest can be used for demand forecasting, customer segmentation, and sales predictions. The ability of Random Forest to model complex relationships between features makes it suitable for various industries, from retail to finance.
    - **Automation**: The model's ability to make accurate predictions could be used to automate decision-making processes, saving time and improving efficiency in industries like marketing, finance, and healthcare.
2. **Scientific Impact**:
    - **Healthcare**: In healthcare, this type of regression model can be used to predict disease progression or patient outcomes based on clinical variables. Random Forest, for example, could help predict treatment effectiveness or disease risk based on multiple patient features.
    - **Research**: In scientific research, this approach could be extended to model complex relationships in various fields, from genetics to environmental studies, helping researchers understand how different variables interact to influence outcomes.
3. **Exploring Feature Relationships**:
    - By analyzing feature importance, businesses and researchers can uncover hidden insights about which factors influence a particular outcome the most. For example, in a healthcare application, understanding the most important predictors of disease outcomes can guide preventive interventions.
4. **Future Applications**:

- The findings from this study could be extended to other datasets in industries like energy consumption, manufacturing, or transportation, where accurate regression models are critical for forecasting demand or optimizing resources.

8.Conclusion and Future Work

8.1 Summary of Outcomes

This project successfully achieved its objectives of building and evaluating two classification models — **Random Forest** and **Support Vector Machine (SVM)**. The models were assessed using key metrics such as accuracy, precision, recall, and F1-score. Overall, **Random Forest** demonstrated a balanced performance, handling class imbalances effectively and excelling in both precision and recall. **SVM** also performed well, especially in precision for certain classes, making it suitable in contexts where avoiding false positives is critical. The insights gained from feature importance in Random Forest provided an understanding of the most influential features, aligning the model's predictions with the data's key aspects. Thus, the project met its objectives, identifying suitable models and offering insights for further application.

8.2 Suggestions for Future Improvements

1. **Incorporate Additional Data Sources**: Expanding the dataset with more samples, especially for minority classes, would enhance model robustness and generalizability, particularly for SVM, which could benefit from a balanced class distribution.

2. **Explore Alternative Models**: Trying other ensemble models like **Gradient Boosting** or **XGBoost** may yield better performance, especially for imbalanced datasets. Neural network architectures could also be explored if the data size supports it, potentially capturing complex patterns.

3. **Feature Engineering and Selection**: Additional domain-specific features could improve the model's predictive power. Further refining the feature set, through techniques like PCA or regularization methods, might enhance performance and interpretability.

4. **Hyperparameter Optimization**: More extensive tuning for both models (e.g., exploring grid search or Bayesian optimization) could optimize their performance further, particularly in handling specific dataset nuances.

5. **Advanced Techniques for Imbalanced Data**: Applying techniques such as **SMOTE (Synthetic Minority Over-sampling Technique)**, **cost-sensitive learning**, or custom loss functions tailored for imbalanced data could improve recall for minority classes without sacrificing precision.

8.3 Implications

The findings of this project have broader implications in contexts requiring accurate classification, especially when balancing recall and precision. Random Forest's balanced approach across metrics and SVM’s precision advantage offer versatile applications in various fields, such as fraud detection, customer segmentation, or risk assessment.

For a follow-up study or project, a larger dataset with additional features could be gathered to improve model generalization and test more complex algorithms. Developing a scalable model pipeline and deploying it in a production environment could also be next steps, enabling real-time predictions and continuous model improvements based on live feedback.

9.References

Datasource
Kaggle : https://www.kaggle.com/datasets/noob2511/wine-data-set