WINE VARIETY PREDICTOR: ACCURATE WINE VARIETY PREDICTION FROM DESCRIPTIONS USING MACHINE LEARNING

Nishmitha Naik

13th June 2024

Summary

- **Objective:**
- To develop a predictive model capable of identifying wine varieties based on their descriptions, mimicking the expertise of a master sommelier. This model aims to assist wine enthusiasts, sellers, and sommeliers in identifying wines more accurately and efficiently, even without tasting them.
- **Approach:**
- Leverage wine data from Kaggle.
- Apply TF-IDF vectorization and train the model using Random Forest Classifier to predict wine varieties.
- **Goal:**
- The goal of this project is to create a reliable and efficient predictive model that can accurately identify wine varieties based on textual descriptions.

Outline

- Business Problem
- Data
- Results
- Conclusions

Business Problem

Predicting Wine Variety:

- The core business problem addressed by this project is the need to accurately predict wine varieties based solely on textual descriptions. This capability is essential for:
  - **Improving Wine Recommendations**: Enhancing personalized wine recommendations for customers.
  - **Enhancing Marketing Strategies**: Tailoring marketing efforts based on accurate wine variety identification.
  - **Scaling Expertise**: Bridging the gap where human sommeliers cannot scale to meet demands, ensuring consistent quality and expertise in wine identification.

Data & Methods

- **Dataset**

- **Source**: The dataset for this project is sourced from Kaggle and includes 130,000 rows of wine reviews and associated metadata. This rich dataset provides a solid foundation for training and testing predictive models.

- **Approach**

- **Text Preprocessing**: Clean and preprocess the text data to make it suitable for model training.
- **Feature Extraction**: Utilize TF-IDF vectorization to convert text data into numerical features.
- **Model Training**: Train a Random Forest Classifier on the preprocessed and vectorized text data.
- **Evaluation**: Assess model performance using accuracy as the primary metric and refine the model as necessary.

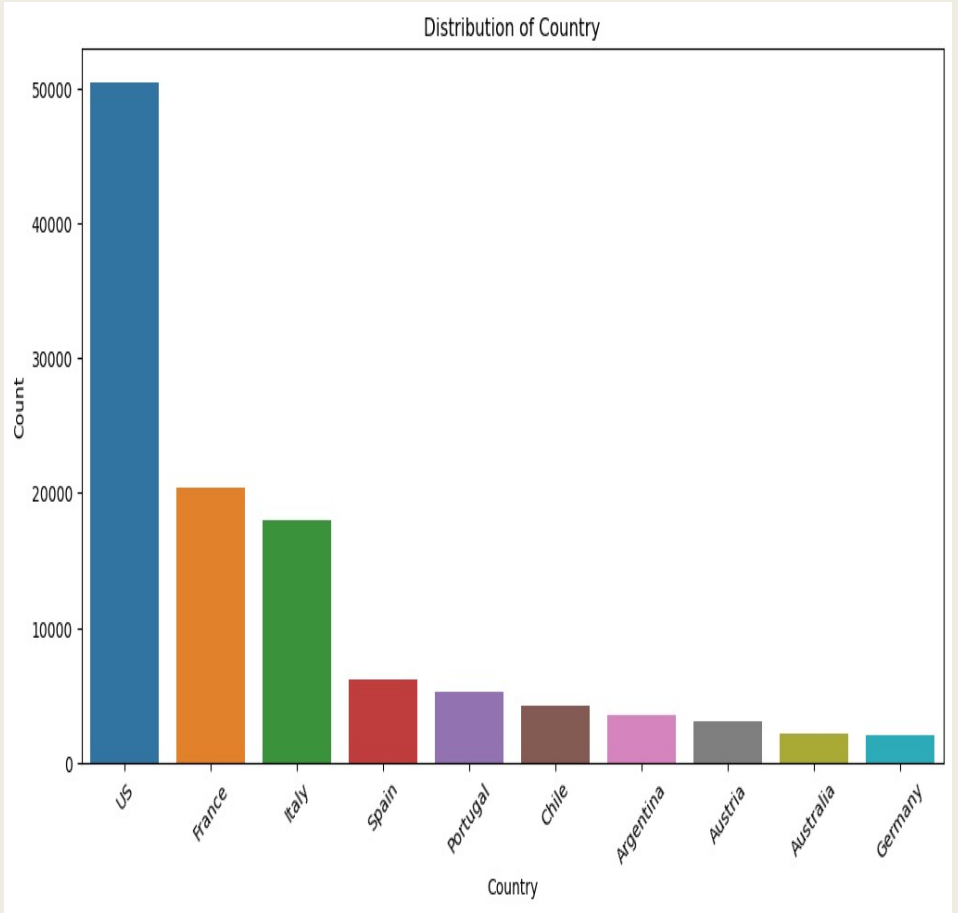
Exploratory Data Analysis

- **Data Visualization**:

- Utilized count plots to understand the distribution of categorical data.
- Focused on Country and Variety columns to identify the most common wine-producing countries and wine varieties.

Insight 1 for Country :

1. US has the highest number of reviews.
2. France and Italy follow in the number of reviews.



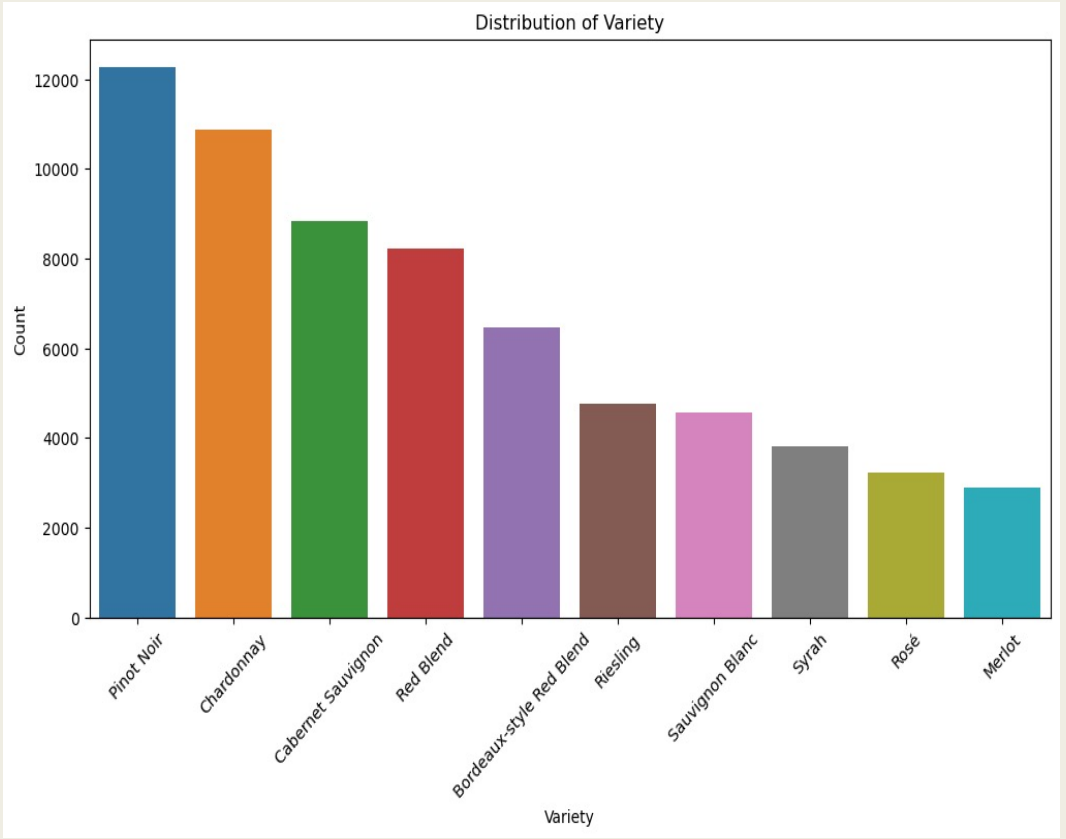
Exploratory Data Analysis

Insights 2 for Wine Variety:

1. Pinot Noir, Chardonnay, and Cabernet Sauvignon are the most reviewed varieties.
2. Red blends and white varieties like Riesling are also significant.

- **Importance**:

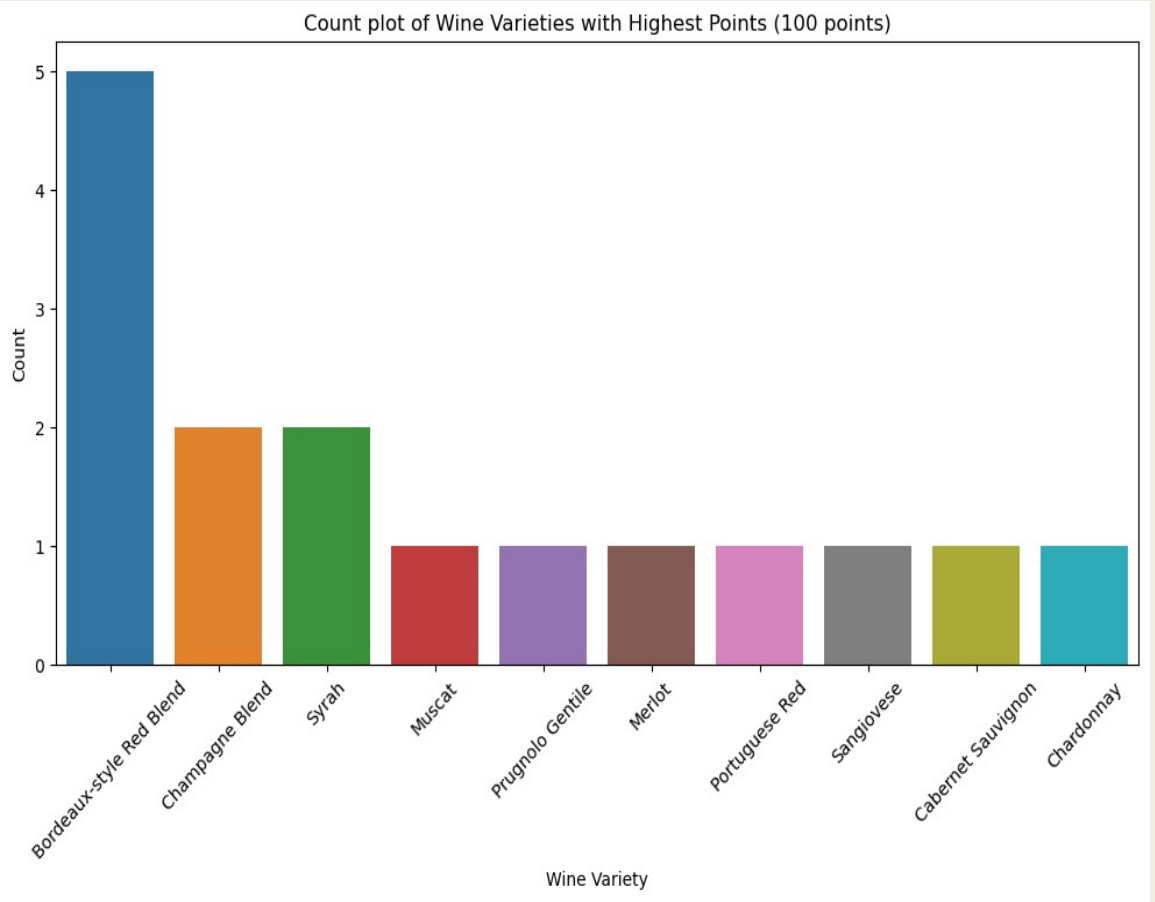
- Understanding the distribution helps in identifying dominant categories in the dataset.
- Provides insights into the most popular wine-producing regions and wine types, which is crucial for market analysis and recommendation systems.



Exploratory Data Analysis

Insights :

- **Varieties with Maximum Points**:
  - Bordeaux-style Red Blend is the most frequently awarded with the highest points.
  - Other notable varieties include Champagne Blend and Syrah.
- **Diverse Wine Varieties**:
  - High-scoring wines include Syrah, Moscat Pinotage, Merlot, and Chardonnay.
  - Exceptional quality is found across different wine styles and types.



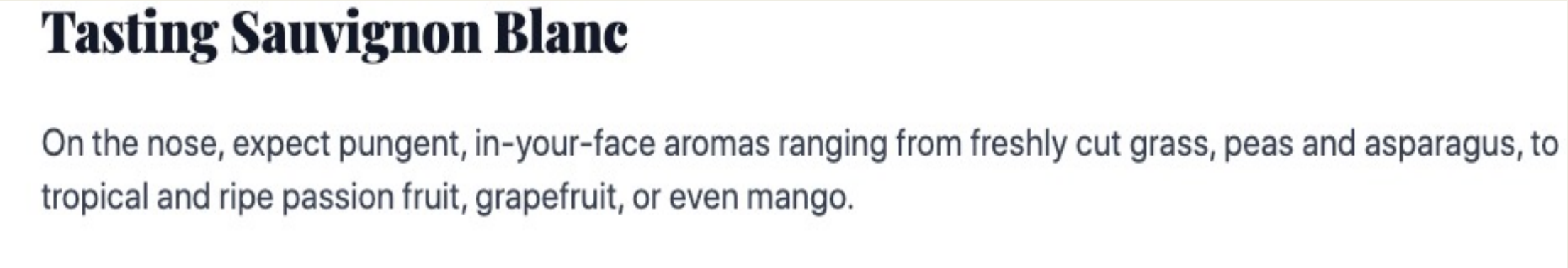
Model Evaluation Results :

- **Accuracy Score:**
  - Evaluated the model's performance by calculating the accuracy, which is the proportion of correct predictions out of all predictions made.
  - Accuracy: 0.4986875
  - Indicating its potential to simulate sommelier-like wine identification capabilities.

Outcome 1 :

- **Test 1 : Predicting Wine Variety from a new description:**
- **Outcome**
- **Result**: The model predicts wine variety accurately based on the provided description.
- **Conclusion**: This approach demonstrates the effectiveness of text preprocessing, TF-IDF vectorization, and the Random Forest classifier in accurately predicting wine varieties from textual descriptions.

Example taken from the Internet



Wine Predictor Model Outcome :

```
# Function to predict wine variety from a new description
def predict_wine_variety(Description):
    # Preprocess the input description
    preprocessed_description = preprocess_text(Description)
    # Transform the preprocessed description using the trained TF-IDF vectorizer
    description_tfidf = tfidf_vectorizer.transform([preprocessed_description])
    # Predict the variety using the trained classifier
    predicted_variety = clf.predict(description_tfidf)
    return predicted_variety[0]

# Example of predicting wine variety from a new description
new_description = "On the nose, expect pungent, in-your-face aromas ranging from freshly cut grass, peas and asparagus, to tropical and ripe
predicted_variety = predict_wine_variety(new_description)
print("Predicted Variety:", predicted_variety)
```

```
Predicted Variety: Sauvignon Blanc
```

Outcome 2 :

- **Test 2 : Predicting Wine Variety from the Dataset:**
- **Outcome**

- **Result**: The model predicts wine variety accurately based on the provided description from the dataset .
- **Conclusion**: This approach demonstrates the effectiveness of text preprocessing, TF-IDF vectorization, and the Random Forest classifier in accurately predicting wine varieties from textual descriptions.

Example taken from the Dataset

	Country	Description	Designation	Points	Price	Province	Region_1	Region_2	Taster_name	Taster_twitter_handle	Title	Variety	Winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	25.0	Sicily & Sardinia	Etna	Unknown	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia

Wine Predictor Model Outcome :

```
# Function to predict wine variety based on an existing description from the dataset
def predict_variety_from_index(index):
    if index < 0 or index >= len(df_wine_data):
        return "Index out of bounds"
    description = df_wine_data.loc[index, 'Description']
    predicted_variety = predict_wine_variety(description)
    return predicted_variety

# Example of predicting wine variety for an existing description from the dataset
index = 0  # Replace with the desired index
predicted_variety = predict_variety_from_index(index)
print(f"Predicted Variety for index {index}:", predicted_variety)
```

```
Predicted Variety for index 0: White Blend
```

Conclusion:

Model Effectiveness

- **Proficiency**: The model effectively extracts relevant information from wine descriptions.
- **Accuracy**: Capable of accurately predicting wine varieties.

Generalization Capability

- **New Descriptions**: Predicts accurately for new wine descriptions.
- **Existing Descriptions**: Consistent performance on existing data.
- **Learned Patterns**: Generalizes well to unseen data.

Practical Applicability

- **Wine Classification**: Useful for sommeliers, wine enthusiasts, and industry professionals.
- **Efficiency**: Facilitates quick identification of wine varieties based on textual descriptions.
- **Machine Learning**: Demonstrates the practical use of ML in real-world wine classification tasks.

Future Considerations:

- **Scaling Up Data Processing:**
- Use distributed computing (e.g., Apache Spark) and cloud services (AWS, Google Cloud) for handling large datasets efficiently.
- **Advanced Model Architectures:**
- Explore deep learning models (RNNs, CNNs, Transformers) and transfer learning for improved predictions.
- **Hyperparameter Tuning:**
- Optimize model performance with grid search, random search, or Bayesian optimization.
- **Feature Engineering:**
- Enhance text features using techniques like Word2Vec or FastText, and incorporate non-text features (e.g., country, price).
- **Real-Time Predictions:**
- Develop a system for real-time predictions via web or mobile applications.
- **Integration with Sommelier Expertise:**
- Validate predictions and enhance model training with professional sommelier input.