INFO 6105 Project

Wine Review

Lingyun Huang
Yi-Chieh Huang
Yi-Ning Ruan

OUTLINE

1. Problem Definition
2. Data Collection and Preprocessing
3. Machine Learning
4. Model Evaluation and Interpretation
5. Future Work and Conclusion

Problem Definition

Objective

- To create a predictive model that can identify wines through blind tasting
- The aim is to employ machine learning for predicting wine variety based on textual descriptions/reviews.

Target Variables

- Variety: The specific type or grape variety used in producing the wine
- Winery: Representing the name of the company or establishment responsible for producing the wine.

Data Collection and Preprocessing

- **Data source** : Kaggle dataset
- **Preprocessing**: Deal with missing values
- **Feature engineering**: TF-IDF (Term Frequency-Inverse Document Frequency), to transform the textual data into numerical format.

Machine Learning

1. **Naive Bayes:**

   The Naive Bayes Classifier is a versatile machine learning algorithm commonly used for text classification tasks like predicting wine variety and winery based on text descriptions. Its strength lies in its probabilistic nature, estimating the conditional probability of a class given features while assuming feature independence.

Machine Learning

2. **Random Forest:**

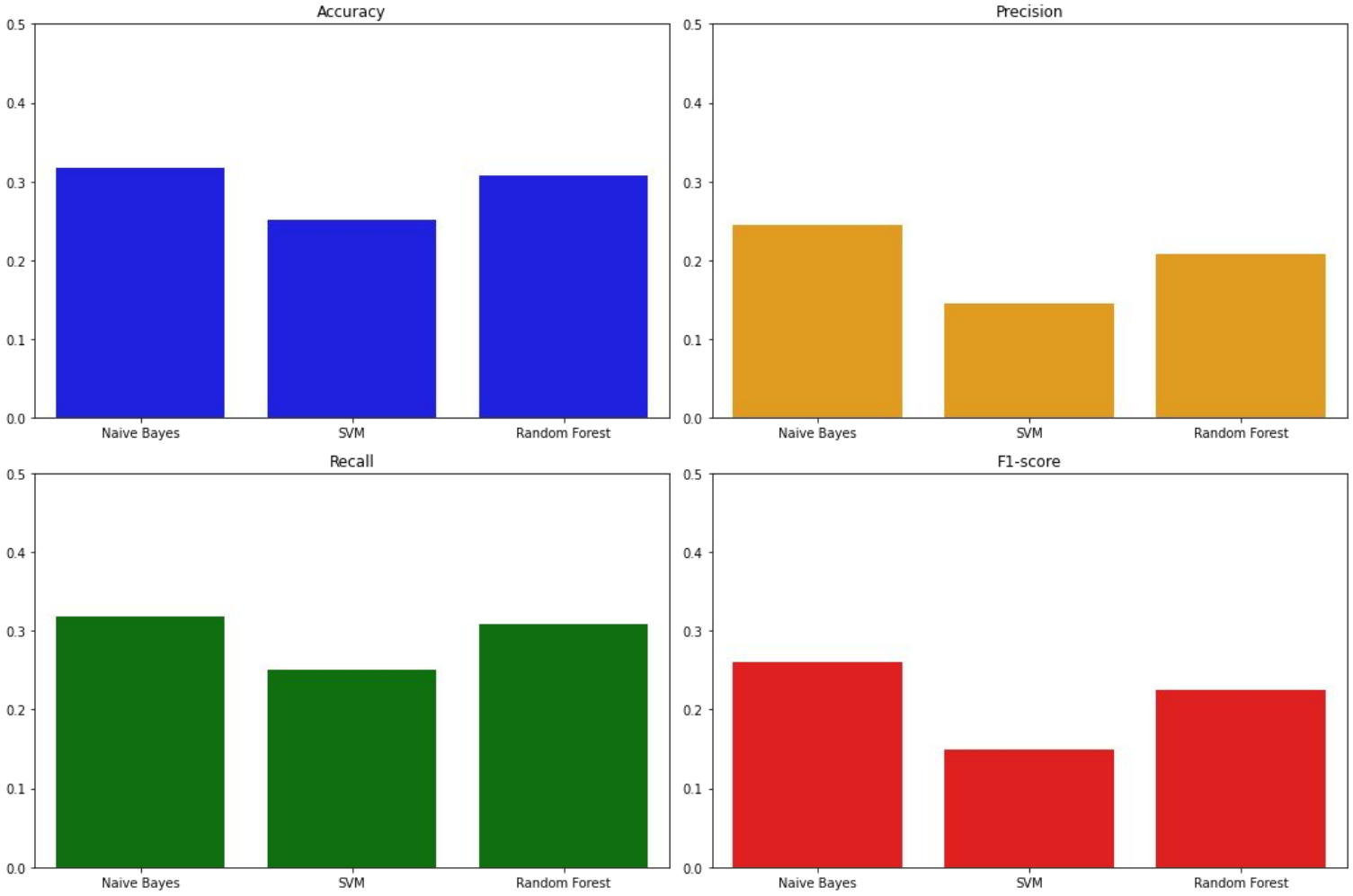
Wine reviews can contain nuanced and complex information, and Random Forest can capture intricate patterns and relationships that might be missed by a linear model like Logistic Regression. By aggregating predictions from multiple decision trees, Random Forest can provide a robust prediction of wine varieties based on the textual data.

Machine Learning

3. **SVM:**

Support Vector Machine is a powerful algorithm suitable for both binary and multiclass classification tasks. When applied to the task of predicting wine varieties and wineries from text descriptions, SVM aims to find a hyperplane that best separates different wine varieties in the high-dimensional space of the text features. Since wine reviews can have multiple words and features, the high-dimensional aspect of the data aligns well with SVM's strengths. SVMs can capture complex relationships between the words in reviews and the wine varieties, even when these relationships are not linear.

Model Evaluation and Interpretation



Naive Bayes achieved the highest accuracy, precision, recall, and F1-score for predicting wine variety.

Conclusion and Future Work

- Our computer's memory reached its limits due to the plety of data. To address this obstacle, we strategically turned to sampling, boasting 130,000 records into a mere 1% representation. While we managed to achieve an accuracy score of 0.7 using the complete dataset, the accuracy plummeted to 0.3 following the sampling process. This starkly highlights the critical role that data quantity plays in refining prediction precision.

Conclusion and Future Work

- In the future, with more robust computing equipment, we can leveraging larger datasets to produce a better model and achieve more superior results.

Thank You!