Munavar Hussain

CS5644

HOMEWORK #2

Question 1:

The dataset in the given appears to be a voting record dataset, with attributes indicating votes on various issues where (represented by "y" for yes, "n" for no, and "?" for missing values). The target variable is political affiliation, specifically whether a person is a "Democrat" or "Republican."

Key Points:

1. **Dataset:**
    a. Voting records with each row representing a person, and columns indicating their votes on different bills or issues.
    b. Missing values ("?") exist for some votes.
2. **Prediction Task:**
    a. The goal is to predict a person's political affiliation (Democrat or Republican) based on their voting record.
3. **Why?**
    a. This task is likely important for understanding patterns in voting behavior, and it could be useful in political science research or applications such as voter profiling and winning statistics of both the parties.

The notebook explores different strategies for handling missing data (dropping rows, imputing values, etc.) and evaluates models like Decision Trees and Naive Bayes using precision, recall, and F1-score to assess their performance.

Data Preprocessing:

1. Handling Missing Data

- **Dropping rows with missing values:** In the initial phase, rows with missing values ("?") were dropped entirely. This is simple but sometimes not recommended, as it reduces the dataset size and may exclude important information.

- **Treating missing values directly:** I replaced missing values with the most common value mainly (mode imputation).

- **Imputation of missing values:** This involved filling in missing values using some statistical technique, like the mean, mode, or a more complex method like KNN imputation, ensuring the data remained intact and complete. In my notebook I just imputed y's with 1, no's with 0 and ?'s with –1 respectively

2. Exploratory Data Analysis:

- Before applying the models, it's crucial to understand the data distribution, the presence of missing values, and the relationship between features. Visualizing the distribution of votes (e.g., number of 'yes', 'no', 'missing' per feature) and comparing class balances (Democrat vs. Republican) was part of this stage.

Models:

1. Decision Tree:

a. **Why?** Decision Trees are interpretable and can handle missing values reasonably well. They can also model non-linear relationships, making them well-suited for categorical data like voting records. In this context, Decision Trees help to determine which votes (features) are most critical in classifying political affiliation. As there are missing values in the dataset I used Decision trees.

2. Naive Bayes:

a. **Why?** Naive Bayes is a simple, yet effective, model for classification tasks, particularly when features are categorical and relatively independent. Although the "naive" assumption of feature independence might not hold for voting data, it provides a useful baseline.
b. **Parameters:** Naive Bayes does not have many parameters to tune, but since it performs well with categorical data, it was a good fit for this task.

Comparison of Models:

- **Why Decision Tree over Naive Bayes?**
  - Decision Trees offer more flexibility and interpretability by explicitly showing which features (votes) are driving the classification. They can model interactions between votes, whereas Naive Bayes treats all features as independent.
  - However, Naive Bayes can still be useful if computational efficiency is a priority or if the dataset is relatively simple.

Model Performance:

- Precision, recall, and F1-score were used to compare the models. All the models performed similarly, but their behavior with different data processing techniques (handling missing data) is where the main difference lies. Decision Trees were slightly better at handling missing values with imputation, while Naive Bayes performed better with fewer missing values.

By exploring how different data preprocessing strategies (dropping, treating, imputing missing values) impacted these models. This analysis shows that pre-processing is essential to any kind of real-world data.

Model	Mean/StdDev of Precision	Mean/StdDev of Recall	Mean/StdDev of F1 Score
NB – Discard Missing	0.92+-0.014	0.92+-0.014	0.92+-0.009
NB – Unique Value	0.92+-0.014	0.92+-0.014	0.92+-0.009
NB – Impute Value	0.95+-0.014	0.95+-0.014	0.94+-0.009
DT – Discard Missing	0.95+-0.0047	0.94+-0.0047	0.94+-0.0047
DT – Unique Value	0.95+-0.0047	0.95+-0.0047	0.95+-0.0047
DT -- Impute Value	0.94+-0.0047	0.94+-0.0047	0.94+-0.0047

Conclusion:

In HW2, we utilized a voting record dataset to predict political affiliation (Democrat or Republican) with an emphasis on data processing techniques and their effect on model performance. We looked at a number of methods for dealing with missing data, including removing rows, appending values, and using exploratory data analysis (EDA) to figure out the distributions and correlations between features. Naive Bayes and Decision Trees were the two models used. While Naive Bayes fared better in recall, correctly recognizing more genuine positive instances, Decision Trees showed superior accuracy, suggesting fewer false positives. Comparable F1 scores were produced by both models, indicating equal overall efficacy. The selection of the models demonstrated the trade-off between recall and precision: Decision Trees are better at reducing false positives, while Naive Bayes is better at catching all pertinent cases. This analysis highlights the significance of data processing and model selection in real-world data science, where understanding the dataset and carefully tuning model parameters can significantly affect outcomes.

Question 2:

Explain in appropriate detail which classifier is suitable for which characteristics of datasets.

Answer:

1. Decision Trees

- **Characteristics of Suitable Datasets:**
  - **Non-linear Relationships:** Decision trees can capture complex, non-linear relationships between features.
  - **Categorical Features:** They work well with both numerical and categorical data without requiring transformation or scaling.
  - **Missing Values:** Decision trees can handle missing values naturally without the need for imputation.
- **Considerations:**
  - Prone to overfitting, particularly in deep trees; hence, depth restriction and trimming might be beneficial.
  - Not resistant to noise and outliers.

2. Naive Bayes

- **Characteristics of Suitable Datasets:**
  - **Independence Assumption:** Best suited for datasets where features are conditionally independent given the class label.
  - **Text Classification:** Effective for high-dimensional datasets like text data (e.g., spam detection) because it can handle many features efficiently.
  - **Categorical Features:** Works well with categorical data and can also handle continuous data through Gaussian or multinomial distributions.
- **Considerations:**
  - When there is a significant degree of correlation between the characteristics, the independence assumption may not perform well.
  - While simplicity might be advantageous, with complicated datasets it may result in underfitting.