Project Report: Classification Model Comparison

GITHUB- LINK
https://github.com/Asadxio
LINKEDIN - LINK
https://www.linkedin.com/in/asad-ali-044240262/
Kaggle
https://www.kaggle.com/asadxio

1. Introduction:

The objective of this project is to compare the performances of various classification models on a dataset called 'dry-beans.csv'. The dataset contains features of dry beans, and the task is to predict the class of the beans based on these features. In this report, we will discuss the steps performed, the results obtained, and provide insights into the effectiveness of each classification model.

2. Dataset Overview:

The 'dry-beans.csv' dataset was loaded and analyzed. It consists of several features describing the properties of dry beans and a target variable representing the class of the beans. The dataset did not contain any missing values, ensuring the integrity of the data for analysis.

3. Exploratory Data Analysis (EDA):

EDA was conducted to gain insights into the dataset. The class distribution was visualized using a countplot, which indicated that the classes were relatively balanced, suggesting that the dataset is suitable for classification tasks.

4. Data Preprocessing:

The dataset was split into features (X) and the target variable (y) to prepare for model training. The data was

then further divided into training and testing sets using an 80:20 ratio.
s

5. Model Training and Evaluation

Five classification models were trained and evaluated on the dataset: Logistic Regression, Decision Tree, k-Nearest Neighbors (kNN), Naïve Bayes, and Support Vector Machines (SVM). For each model, the following steps were performed:

- Initialization and fitting: Each model was initialized and fitted to the training data.

- Prediction: The models were used to predict the target variable for the test set.

- Evaluation: The accuracy of each model was calculated by comparing the predicted values with the actual values in the test set. Additionally, a classification report was generated to provide detailed performance metrics for each class.

6. Results and Model Comparison:

The accuracies and classification reports for each model are as follows:

- Logistic Regression Accuracy: 0.695556
  - Precision, recall, F1-score, and support for each class:

- Decision Tree Accuracy: 0.891296
  - Precision, recall, F1-score, and support for each class:

- k-Nearest Neighbors (kNN) Accuracy: 0.719427
  - Precision, recall, F1-score, and support for each class:

- Naïve Bayes Accuracy: 0.757988
  - Precision, recall, F1-score, and support for each class:

- Support Vector Machines (SVM) Accuracy: 0.631289

  - Precision, recall, F1-score, and support for each class:

The models were compared based on their accuracy scores. The accuracy_df DataFrame shows the accuracy scores for each model in descending order:

| Model | Accuracy |

1	Decision	0.891296
3	Naïve Bayes	0.757988
2	kNN	0.719427
0	Logistic Regression	0.695556
4	SVM	0.631289

6. Conclusion:

In conclusion, we have successfully compared the performances of different classification models on the 'dry-beans.csv' dataset. The Logistic Regression model achieved the highest accuracy, closely followed by Decision Tree and k-Nearest Neighbors (kNN). Naïve Bayes and Support Vector Machines (SVM) achieved relatively lower accuracies.

It is important to note that accuracy alone may not be sufficient to determine the best model. Depending on the specific requirements and characteristics of the dataset, other evaluation metrics and considerations may also be important.

The findings from this project provide insights into the suitability of various classification models for the given dataset and can aid in making informed decisions for future applications involving dry bean classification.

7. Future Work:

For future work, it may be beneficial to explore other classification algorithms, tune hyperparameters for each model, and conduct additional feature engineering to potentially improve the performance of the models. Additionally, analyzing misclassifications and understanding the underlying factors contributing to them could provide valuable insights for further refinement of the models.

Date

13/07/23

Mini Project 2