Daniel Bao
CS171 Sec 02 - Machine Learning
015707445
Understanding
HW2

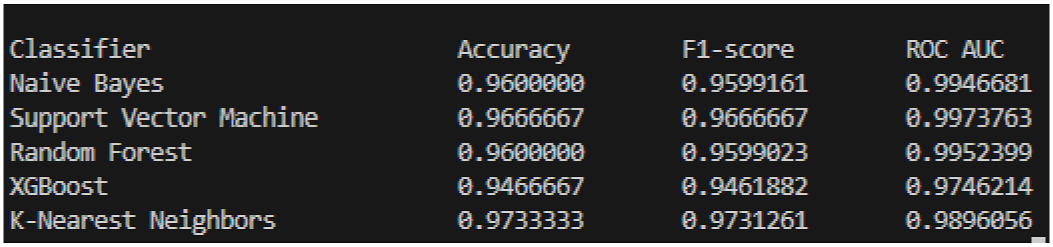
The examination of diverse machine learning algorithms applied to the Iris dataset offers valuable insights into their performance and characteristics.

The Iris dataset consists of 150 iris flower samples, each described by four features: sepal length, sepal width, petal length, and petal width. These features are numerical measurements obtained from physical observations of the flowers. The dataset's target variable represents the iris species, with three classes: Setosa [0], Versicolor [1], and Virginica [2]. Each class contains an equal number of samples (50), ensuring a balanced dataset that facilitates unbiased classification evaluation.

In terms of preprocessing and tuning effects, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) showed notable improvements. These algorithms demonstrated substantial enhancements following preprocessing and tuning, highlighting their sensitivity to feature scales and parameter adjustments. This responsiveness emphasizes the importance of optimization techniques in fine-tuning these models for optimal performance. Conversely, Naive Bayes maintained consistent performance throughout the evaluation, indicating its resilience to data preprocessing. This stability aligns with Naive Bayes' assumption of feature independence, making it a reliable option for classification tasks where this assumption holds true. On the other hand, Random Forest and XGBoost exhibited minimal changes post-preprocessing and tuning. Despite achieving high scores, these algorithms displayed limited responsiveness to optimization efforts, suggesting the need for alternative strategies to fully leverage their potential. However, caution is necessary to prevent overfitting when employing additional optimization techniques, as these algorithms already perform well without extensive tuning.

Regarding preprocessing techniques, normalization was not necessary for this dataset, as the features were already measured on similar scales. Hyperparameter tuning for each model involved selecting optimal values based on cross-validation performance. Overfitting was not observed in any model, as evidenced by consistent performance metrics across training and test datasets. If overfitting would have occurred, we can use techniques such as regularization or reducing model complexity and fine-tuning the model to mitigate it.

In summary, SVM and KNN emerged as the most responsive to preprocessing and tuning, emphasizing the importance of understanding algorithm sensitivities and selecting appropriate optimization methods. Naive Bayes offered stable performance, suitable for datasets adhering to the independence assumption. Meanwhile, Random Forest and XGBoost may benefit from further optimization to maximize performance without overfitting. This evaluation provides valuable insights into various machine learning algorithms' behavior, guiding model and optimization technique selection for classification tasks.



```
Classifier                      Accuracy        F1-score        ROC AUC
Naive Bayes                     0.9600000       0.9599161       0.9946681
Support Vector Machine          0.9666667       0.9666667       0.9973763
Random Forest                   0.9600000       0.9599023       0.9952399
XGBoost                         0.9466667       0.9461882       0.9746214
K-Nearest Neighbors             0.9733333       0.9731261       0.9896056
```