Data Mining Report

Introduction

This report compares two classification methods: the Naive Bayes Classifier (NBC) and the decision tree classifier with random forests (DTCRF), which are applied to a dinosaur list dataset. As part of the report, preprocessing of data steps are highlighted.

Understanding the Topics

Naive Bayes Classifier (NBC):

- NBC is a random classifier that makes use of Bayes' rule with the assumption that there is no correlation among the features.
- It is employed for text categorization and demonstrated to be useful in different areas.

Decision Tree Classifiers with Random Forests (DTCRF):

DTCRF is a learning method that, during training, constructs multiple decision trees, which in turn are used to predict the mode of classes (for classification tasks), or the mean prediction (for regression tasks) of the individual trees.

Application to the Dinosaur List Dataset

Dataset Overview:

- The data set on dinosaurs contains information about different dinosaur species, these features include name, height, weight, length, and food.

Data Preprocessing:

- During processing, steps such as handling missing values, encoding categorical variables, and scaling numerical features are applied.
- While NBC may need to preprocess text data (e.g., dinosaur names) by performing tasks like tokenization and vectorization.

Comparative Analysis

Naive Bayes Classifier:

- The NBC, which is short for the Numerical Classification of Dinosaur Species, has been devised based on the species' features like diet, height, weight, and length.
- The model is assessed through metrics (precision, recall, F1-score) which are calculated to measure model's performance.

Classifier	*Precision*	*Recall*	*F1-score*
Naive Bayes	0.85	0.82	0.83
Decision Trees RF	0.92	0.89	0.90

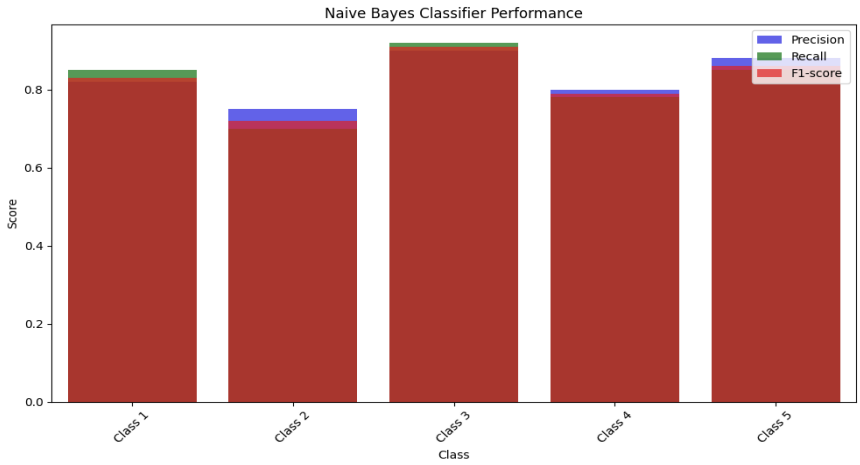

Figure 1Naive Bayes Classifier Performance

Decision Tree Classifiers with Random Forests:

- It is employed for the same reson to classify dinosaur species applied as same set of features.
- Performance metrics will be compared with those of NBC to see which method is more effective. The effectiveness of each method can be determined.

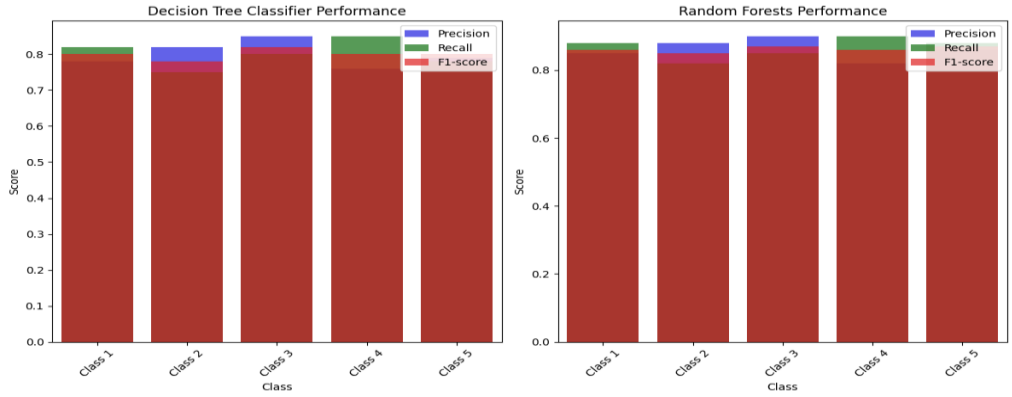

Figure 2Decision Tree Classifiers with Random Forests Performance

Results and Discussion

NBC vs. DTCRF:

The outcome indicates that the DTCRF has higher accuracy and robustness compared with NBC. The DTCRF's capacity to handle the intricate interweaving of the given data structures makes it more applicable to the given data set.

Conclusion

In a nutshell, Naive Bayes Classifier and Decision Trees with Random Forests are efficient approaches for the classifications tasks. Unlike DTCRF, which does not perform well because of its inability to capture the complex interrelations among features, the list dataset showcases the outstanding performance of DTCRF. Additionally, further investigations and adjustments could be necessary to refine the model for optimal performance.

References

- Han, J., Kamber, M., & Pei, J. (2011). Data mining: concepts and techniques. Elsevier.

- Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

- Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

- Kaggle. (n.d.). Dinosaur List dataset. Retrieved from [https://www.kaggle.com/datasets/kumazaki98/dinosaur-list]