PeterP22 GitHub Report on Mushroom Dataset

Comparative Machine Learning Analysis on Primary and Simulated Secondary Mushroom Datasets

Abstract: This report outlines an analytical study conducted on primary and secondary mushroom data sets utilising several machine learning (ML) classifiers. The primary objective was to evaluate the performance of these classifiers in terms of accuracy and F2 scores, while comparing the results between primary and secondary data.

1. Dataset Summary: The primary data set comprises 173 instances, each representing a different mushroom species. Each instance is described by a set of 20 attributes, which are a mix of both qualitative and quantitative characteristics, including cap-diameter, cap-shape, cap-surface, and cap-colour, among others. The secondary data set is a simulated dataset generated based on the primary data, consisting of 61,069 hypothetical mushroom instances, with 353 instances for each of the 173 species from the primary data. The mushrooms in the data set are labelled as either edible or poisonous (with the latter also encompassing those of unknown edibility). Both datasets were used in binary classification tasks to determine whether a mushroom was edible or poisonous.

2. Machine Learning Methods: We employed four different machine learning classifiers: Naive Bayes, Logistic Regression, Linear Discriminant Analysis, and Random Forest. These classifiers were trained and evaluated on both primary and secondary datasets.

3. Experiment Protocol: The data underwent pre-processing, including handling missing values, separating features, and labels, encoding categorical features, and splitting the data into training and testing sets. Each classifier was then cross-validated using 5-fold stratified cross-validation and evaluated based on accuracy and F2 score. The entire process was executed separately for the primary and secondary datasets.

4. Results: The classifiers exhibited variable performance on the two datasets. For the primary dataset, accuracy scores ranged from 0.449 (Naive Bayes) to 0.544 (Random Forest), and F2 scores ranged from 0.474 (Naive Bayes) to 0.712 (Random Forest). For the secondary dataset, the classifiers demonstrated higher performance, with accuracy scores ranging from 0.722 (Naive Bayes) to 1.000 (Random Forest), and F2 scores ranging from 0.768 (Naive Bayes) to 1.000 (Random Forest). The results are visually presented using boxplots.

5. Conclusion: The study highlights the variability in the performance of different machine learning classifiers on two distinct datasets. Random Forest demonstrated the highest accuracy and F2 scores on both datasets, with perfect performance on the secondary dataset. This exploration provides valuable insights for selecting machine learning models for binary classification tasks in similar contexts.