Assignment 10: AutoML pipeline

Dataset Description and Preprocessing

The sports ball dataset was sourced from Hugging Face (Shanav12/sports_ball_dataset). It contains images of various sports balls labeled by their respective types. Since machine learning models typically work with numerical features, each image was preprocessed by resizing to 100×100 pixels, converting to grayscale, normalizing pixel values, and flattening into a 1D array. This transformation ensures that the classifiers can efficiently process the image data.

The Iris dataset, obtained from Hugging Face (scikit-learn/iris), is a well-known dataset in machine learning. It contains 150 samples with four numerical features (sepal length, sepal width, petal length, petal width) and three possible species labels: Iris-setosa, Iris-versicolor, and Iris-virginica. To simplify the classification task, the dataset was converted into a binary classification problem, where:

- 0 represents Iris-setosa
- 1 represents non-Iris-setosa (Iris-versicolor and Iris-virginica)

Additionally, the "Id" column was removed as it did not contribute to classification.

AutoML Pipeline and Configuration

The AutoML pipeline was implemented using LazyPredict, which automates the evaluation of multiple supervised learning classifiers. Both datasets were split into training and testing sets to ensure fair evaluation. The split ratio was 75% training and 25% testing for the sports ball dataset, and 80% training and 20% testing for the Iris dataset.

The evaluation metrics used include:

- Accuracy: Measures the proportion of correctly classified samples.
- Balanced Accuracy: Adjusts for class imbalances.
- F1-Score: Accounts for precision and recall, useful for imbalanced datasets.
- ROC-AUC Score: Measures the model's ability to distinguish between classes (only applicable for binary classification).

LazyPredict was configured to test multiple classification algorithms, such as RandomForestClassifier, Support Vector Machines (SVC), Gradient Boosting, and k-Nearest Neighbors (k-NN), among others. The models were evaluated on their default hyperparameters, and their performance was compared based on accuracy and computation time.

Results and Observations

Sports Ball Classification Results

The best-performing model for the sports ball dataset was RandomForestClassifier, achieving an accuracy of 69% and f1-score of 70%. Other strong models included:

- ExtraTreesClassifier (accuray 67%, f1-score 68%)
- NuSVC (64%, f1-score 64%)

The ROC-AUC score was unavailable because this was a multi-class classification problem. The relatively low accuracy suggests that distinguishing different types of sports balls based on grayscale pixel intensity alone is challenging. This may be due to similarities between different ball textures and shapes. Potential Improvements:

- Implementing Convolutional Neural Networks (CNNs) for better feature extraction.
- Using color images instead of grayscale, as color may provide more distinguishing features.
- Increasing dataset size to improve generalization.

Iris Dataset Classification Results

The Iris dataset yielded near-perfect classification performance, with most models achieving 100% accuracy and 100% f1-score. The best-performing classifiers included:

- AdaBoostClassifier
- RandomForestClassifier
- XGBClassifier

The ROC-AUC score of 1.00 indicates that these models perfectly separated the two classes. This is expected, as the Iris dataset is relatively simple, and the decision boundaries between classes are well-defined.Potential Improvements:

- Testing multi-class classification instead of binary classification to assess the models' performance on the original three-class problem.
- Using feature selection techniques to determine which attributes contribute most to classification.

Conclusion

This project successfully demonstrated the effectiveness of AutoML in handling two different classification problems. LazyPredict proved to be a valuable tool for rapid model benchmarking, allowing for quick comparisons of various classification algorithms. The sports ball classification task was more challenging, highlighting the limitations of traditional machine learning models when applied to image data. Deep learning approaches such as CNNs would likely improve performance. In contrast, the Iris dataset was easily classified, achieving perfect accuracy with several models. This highlights the power of classical machine learning techniques when applied to well-structured tabular data.