Report

1. Introduction

The objective of this task is to evaluate the performance of an Auto-ML system using TPOT, which automates the selection and optimization of machine learning pipelines. This report summarizes the steps taken, the models and hyper parameters selected, and provides performance metrics for the best-performing model. The data-set used is the Iris data-set, a popular classification data set, making it an ideal candidate to test the effectiveness of automated machine learning.

2. Dataset and Problem Description

The Iris dataset consists of 150 samples of iris flowers, with the following attributes:
- Sepal Length
- Sepal Width
- Petal Length
- Petal Width

The task is to classify the samples into one of three species:
1. Setosa
2. Versicolor
3. Virginica

The dataset was divided into:
- 80% training data for model training and hyperparameter tuning.
- 20% testing data for evaluating the final model's performance on unseen data.

3. Steps

3.1. Dataset Preprocessing

- Loading the Dataset: The Iris dataset was loaded using the `load_iris()` function from scikit-learn.
- Splitting the Dataset: The data was split into training and testing sets using the `train_test_split()` function from scikit-learn, with 80% of the data used for training and 20% for testing.

3.2. AutoML Process

TPOT was used to automate the following steps:

1. Model Selection: TPOT tested multiple machine learning models, such as Random Forest, Gradient Boosting, Decision Trees, and others.

2. Hyperparameter Tuning: For each model, TPOT optimized hyperparameters using genetic programming.
3. Pipeline Optimization: TPOT generated and evaluated different machine learning pipelines to select the best one.

The configuration for TPOT was as follows:
- Generations: 5 generations, meaning TPOT would iterate 5 times over the possible pipelines.
- Population Size: 50, meaning TPOT would evaluate 50 pipelines in each generation.
- Random State: 42 to ensure reproducibility.

4. Best Model Selection

After testing several pipelines, TPOT selected the following machine learning pipeline as the best model:

```
pipeline = make_pipeline(
    StandardScaler(),
    GradientBoostingClassifier(learning_rate=0.1, max_depth=3, n_estimators=100)
)
```

Pipeline:

1. Preprocessing Step:
   - StandardScaler: This step standardized the feature values by scaling them to have zero mean and unit variance. This ensures that all features are on the same scale, improving the performance of many machine learning algorithms.

2. Model Selection:
   - Gradient Boosting Classifier: TPOT selected Gradient Boosting as the optimal classifier for this problem.
   - Hyperparameters:
     - learning_rate: 0.1, controlling how much each tree contributes to the final prediction.
     - max_depth: 3, limiting the depth of each individual decision tree to prevent overfitting.
     - n_estimators: 100, the number of boosting stages.

5. Model Performance Evaluation

5.1 Accuracy

The selected pipeline was evaluated on the test set, achieving the following accuracy:

```
accuracy = accuracy_score(y_test, y_pred)
print(f'Accuracy of the best model: {accuracy:.4f}')
```

- Accuracy: 96.67%

This high accuracy indicates that the selected Gradient Boosting model performed exceptionally well in classifying the iris species.

5.2 Classification Report

The following classification metrics were computed using the `classification_report` function:

Class	Precision	Recall	F1-Score
Setosa	1.00	1.00	1.00
Versicolor	1.00	0.93	0.96
Virginica	0.92	1.00	0.96

- Precision: All Setosa and Versicolor flowers were correctly classified with 100% precision.
- Recall: All Virginica flowers were correctly classified, with a recall score of 1.00.
- F1-Score: The F1-Score balances precision and recall, showing excellent performance across all classes.

5.3 Confusion Matrix

The confusion matrix visually represents the classification performance:

Predicted	Setosa	Versicolor	Virginica
Setosa	10	0	0
Versicolor	0	14	1
Virginica	0	0	10

6. Conclusion

The AutoML process using TPOT successfully identified the Gradient Boosting Classifier as the best model for the Iris dataset. The model achieved an accuracy of 96.67% on the test data, with excellent precision, recall, and F1-scores for all classes.

Conclusions:
- Model Selection: TPOT’s automatic search and optimization process efficiently selected the best-performing model and hyperparameters without manual intervention.
- Performance: The selected model performed exceptionally well in classifying the three iris species, with only one misclassification in the test data.
- Scalability: The process can be applied to more complex datasets, where the automation of model selection and tuning would yield significant time savings.

The AutoML system not only reduces the effort required for model selection and hyperparameter tuning but also ensures that optimal models are selected based on the dataset's characteristics.