2024

ASSIGNMENT 2

MACHINE LEARNING

ROHA ARSLAN(FA21-BSE-116)

COMSATS University Islamabad, Lahore Campus

Course Title:	Machine Learning				Course Code:	CSC354	Credit Hours:	3(3,0)
Resource Person:	Dr. Muhammad Sharjeel				Programme Name:	BSSE		
Semester:	6th	Batch:	FA21	Section:	A, B	Max Marks:	10	

Assignment 2: **Due Date/Time: Wednesday, 27th March, 23:59**

Submission: Upload the assignment solution to your GitHub account (private repository).

Important instructions: If the assignment requires to submit a python source code (ipython notebook), please write the following information at the start of your file.
 # Date
 # CSC354 – Assignmen2 – ML – Decision Trees
 # Your Full Name
 # You Complete Registration Number
 # A brief description of the task

Question1: [CLO-2] - [Bloom Taxonomy Level: <Applying>]

Download the Datasaurus Dozen dataset from the following link.

Link: https://www.openintro.org/data/csv/datasaurus.csv

Note: Please open the dataset file first for manual inspection before performing any experiments.

Use this dataset for a classification task using decision trees. Specifically, use J48 and Random Forest classifiers for predicting the type of 'dataset' within the Datasaurus Dozen. Start with a baseline model with default parameters. Then find optimal parameters for the model using both Random and Grid search methods. You are free to use any train/test split, however, only experiment with models' parameters, keeping rest of the settings constant throughout the experiments.

ANSWER: Code in python notebook, uploaded in github

Question2: [CLO-2] - [Bloom Taxonomy Level: <Applying>]

Download the Used Cars Prices dataset (cars-dataset) from your shared Google Drive folder.

Link: http://tinyurl.com/sp24ml

Note: Please open the dataset file first for manual inspection before performing any experiments.

Use this dataset for a regression task using decision trees. Specifically, use a Decision Tree Regressor for predicting the price of a car. Similar to Q1, start with a baseline model with default parameters and then try to find the optimal parameter settings using both Random and Grid search methods.

ANSWER: Code in python notebook, uploaded in github

Question3: [CLO-2] - [Bloom Taxonomy Level: <Applying>]

Write a short report explaining your experience after attempting both Q1 and Q2.

Report on Q1 and Q2:

Data Exploration:

The two datasets, "datasaurus.csv" and "cars-dataset.csv," were thoroughly examined to understand their feature distributions and underlying structure. Understanding them proved important for the next modelling tasks.

Baseline Model Establishment:

In Q1, baseline models were created with Random Forest and C4.5 decision tree classifiers; in Q2, the baseline model was a Decision Tree Regressor. To set a performance baseline, these early models were trained using default parameters.

Parameter Optimization:

Grid and Random search methods were used to optimize the parameters. While Q2 concentrated on lowering Mean Squared Error (MSE) to improve regression model performance, Q1 investigated different parameter combinations to increase classification accuracy.

Evaluation Metrics:

For Q1, the main evaluation metric was classification accuracy, while for Q2, regression model predictions were evaluated using Mean Squared Error (MSE). This selection of measures made it possible to directly compare baseline and optimized models and offered insights into the predictive behaviour of the models.

Optimization Duration:

When compared to other optimization techniques, Q1's Grid search for the Random Forest classifier takes noticeably longer (15–30 minutes). The extended length may be attributed to the exhaustive investigation across a wide range of parameters.

Performance Improvement:

The main objective was to outperform the baseline model for both experiments. This was accomplished by gradually adjusting model parameters to reduce mean square error (MSE) in regression tasks and increase accuracy in classification tasks.

Conclusion:

The results of the tests showed how well parameter optimization strategies work to improve model performance in various machine learning applications. The systematic study of parameter space consistently resulted in gains in predicting accuracy, even with different datasets and tasks. Additionally, determining the proper evaluation standards was crucial in evaluating model performance and directing optimization efforts. Given the circumstances, the investigations provided insight on how model behaviour is affected by parameter tuning and emphasized the significance of optimization in the creation of machine learning models.