IST 707 Applied Machine Learning

By Prof. Kelvin King

Assignment 5

**Decision Tree**

Submitted by:

Samarth Sandesh Mengji  
SUID: 718473878

NetID: [smengji@syr.edu](mailto:smengji@syr.edu)

Date of Submission: 10/8/2023

Syracuse University – School of Information Studies

**Authorship Attribution Report**

**Section 1: Data Preparation**

In this report, we aim to attribute the correct authorship of the disputed essays from Federalist papers. The dataset was first split using train\_test\_split method into training and testing data. The dataset has total of 11 disputed authorship which we try to identify the right author using our analysis.

The training and testing data was in ratio of 80/20.

The feature set for our classification model was based on the frequency distribution of common function words in the essays.

**Section 2: Decision Tree Model**

**Decision Tree Model**

We built a decision tree classification model using the default settings. The decision tree model was trained on the training dataset, and we evaluated its performance on the testing dataset.

**Section 3: Prediction**

After training the decision tree model, we used them to predict the authorship of the disputed essays. **The predictions for the disputed essays are as follows:**

Disputed Essay 1: Author: Madison

Disputed Essay 2: Author: Madison

Disputed Essay 3: Author: Jay

Disputed Essay 4: Author: Madison

Disputed Essay 5: Author: Madison

Disputed Essay 6: Author: Madison

Disputed Essay 7: Author: Madison

Disputed Essay 8: Author: Madison

Disputed Essay 9: Author: Madison

Disputed Essay 10: Author: Madison

Disputed Essay 11: Author: Madison

**Model Evaluation Metrics**

Accuracy: 0.9333333333333333

Precision: 0.8888888888888888

Recall: 0.8333333333333334

F1 Score: 0.8222222222222223

**Conclusion**

With an accuracy of approximately 93.33% and a precision of 88.88%, the model successfully identified James Madison as the likely author of the disputed essays.

The recall score of 91.6% indicates that the model correctly identified a significant portion of Madison's essays.

In conclusion, the decision tree model predicts that James Madison is the probable author of the disputed essays. These results are consistent with the findings of previous clustering methods.

**Assignment 2 Report: Bike Rental Predictions with Decision Trees**

**Section 1: Data Preparation**

In this section, we separated the original dataset into training and testing data for classification experiments. We used two datasets, hour.csv and day.csv, for predicting hourly and daily rental counts, respectively. The following splits were made:

**Hourly Rental Count Model:**

Training Data: 80% of hourly data

Testing Data: 20% of hourly data

**Daily Rental Count Model:**

Training Data: 80% of daily data

Testing Data: 20% of daily data

**Section 2: Build and Tune Decision Tree Models**

Built decision tree models using default settings and then tuned the parameters to evaluate if better models could be generated. Based on the max\_depth and min\_samples\_leaf hyperparameters.

Hourly Rental Count Model - Default Settings:

Mean Squared Error (MSE): 3392.90

R-squared (R^2): 0.8929

Daily Rental Count Model - Default Settings:

Mean Squared Error (MSE): 968219.27

R-squared (R^2): 0.7585

Hourly Rental Count Model - Tuned Settings:

Best Hyperparameters: {'max\_depth': None, 'min\_samples\_leaf': 4}

Mean Squared Error (MSE): 2778.51

R-squared (R^2): 0.9123

Daily Rental Count Model - Tuned Settings:

Best Hyperparameters: {'max\_depth': None, 'min\_samples\_leaf': 8}

Mean Squared Error (MSE): 763530.46

R-squared (R^2): 0.8096

Best hyperparameters were searched using grid search.

**Section 3: Prediction**

In this section, we observed the results and compared the models with default and tuned settings.

The following are the insights noticed:

The default hourly rental count model performed well with an R-squared value of 0.8929, indicating a good fit.

The daily rental count model with default settings had a lower R-squared value of 0.7585, which suggests that the model might not capture the underlying patterns effectively.

The hourly rental count model improved significantly after hyperparameter tuning with an R-squared value of 0.9123 and a lower MSE. The best hyperparameters were found to be max\_depth=None and min\_samples\_leaf=4 using grid search.

Similarly, the daily rental count model improved after tuning, with an R-squared value of 0.8096 and a lower MSE. The best hyperparameters were max\_depth=None and min\_samples\_leaf=8.

**Conclusion**

After tuning the decision tree models with hyperparameters, the models showed better results significantly for predicting bike rental counts. By choosing the right hyperparameters, model predictive accuracy increases.

I certify that this assignment represents my work. I have not used any unauthorized or unacknowledged assistance or sources in completing it, including free or commercial systems or services offered on the internet.

Reference:

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html>

<https://stackoverflow.com/questions/31421413/how-to-compute-precision-recall-accuracy-and-f1-score-for-the-multiclass-case>

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>

<https://towardsdatascience.com/gridsearchcv-for-beginners-db48a90114ee>

<https://www.mygreatlearning.com/blog/gridsearchcv/>