Machine Learning Mini-Project-1
README File

Group No: 7

Group Members:

1. Tanay Raghavendra (18EC10063)
2. Ayan Chakraborty (18EC10075)
3. Debjoy Saha (18EC30010)

Project Info: Customer Purchase Prediction using Decision Tree-based Learning Model

Project Code: PPDT

File Descriptions: We provide the basic functionality of each file for ease of navigation.

```
./
├── ML_MiniProject1_Group7_Report.pdf
├── code
│   ├── Cross_Validation_main.py
│   ├── dataset.py
│   ├── dataset_size_variation.py
│   ├── decision_tree.py
│   ├── main.py
│   ├── plots
│   │   ├── Best_Decision_Tree_Visualization.png
│   │   ├── Decision_Tree_Sklearn.png
│   │   ├── Decision_Tree_Visualization.png
│   │   ├── DevDataset_size_variation.png
│   │   ├── ID3_final.png
│   │   ├── TrainDataset_size_variation.png
│   │   ├── Xval.png
│   │   ├── Xval_4.png
│   │   ├── Xval_7.png
│   │   ├── gini_final.png
│   │   └── sck_final.png
│   └── preds
│    ├── own_depth_4.csv
│    ├── own_depth_7.csv
│    └── scikit_depth_7.csv
│
```

code/

decision_tree.py
Decision Tree class implementation and functions for decision tree training with simultaneous continuous and categorical variables. Provides choice for pruning parameter `max_depth` and information gain (Gini-index or Entropy)

dataset.py
Class to return the dataset objects for training. Reads the train, validation, test CSV files, and preprocesses data. Converts all data to a suitable format based on attribute value and fills in NaN values. Function `get_dataset` returns pandas data-frame for X, training variable, y, target variable.

main.py
Trains the decision tree at different depths and information gains. Provides comparison with sklearn implementation. Provides model statistics.

Cross_Validation_main.py
Runs K-fold cross-validation training with varying pruning parameters depth values.

dataset_size_variation.py
Trains the decision tree training with different percentages of training data and finds best.

plots/

- Best_Decision_Tree_Visualization.png: Visualization of the optimal decision tree among all trained models.
- Decision_Tree_Visualization.png: Visualization of the pruned decision tree among all trained models.
- Decision_Tree_Sklearn.png: Visualization of the pruned decision tree trained using the sklearn implementation.
- DevDataset_size_variation.png: Validation dataset variation for different dataset sizes for training with different depths.
- gini_final.png: Comparison of training and validation set performances for Gini-Index at different depths.
- ID3_final.png: Comparison of training and validation set performances for entropy gain at different depths.
- sck_final.png: Training and validation accuracy comparison with sklearn decision tree at different depths.
- TrainDataset_size_variation: Training set performance with different dataset size for best and pruned trees.
- Xval - Basic cross-validation training
- Xval_4 - Cross validation training with depth = 4
- Xval_7 - Cross validation training with depth = 7

preds/

Containing the prediction results on the test data generated using our own models (of depth 4 and 7) and the scikit model.

- own_depth_4.csv: Predicted results using own model of maximum depth 4
- own_depth_7.csv: Predicted results using own model of maximum depth 7
- scikit_depth_7.csv: Predicted results using scikit model of maximum depth 7