1. Names of all group members
   Jawad Rajabi
   mohraj-0@student.ltu.se

2. Clear specification of the addressed grading criteria

   For grade 3: Develop 1 unsupervised and 1 supervised classification model for 5 datasets of your choice from 121 UCI datasets. Report accuracy results

   Iris, wine, abalone, adult and digit are for testing.
   Models that used are KMeans and DecisionTreeClassifier.
   The result is measured using silhouette score for KMeans and accuracy for DecisionTreeClassifier.

3. Description of the datasets used in the miniproject

Iris:

Dataset for classifying flowers with three different species: Setosa, Versicolor and Virginica.

Size: Total 150 samples

Classes: Setosa, Versicolor and Virginica.

Majority percentage: 33.33% (All three classes are evenly distributed with 50 samples)

Features: sepal length (cm), sepal width (cm), petal length (cm), petal width (cm)

Wine:

Dataset that analyzes chemical properties of wine to classify them into different categories.

Size: Total 178 samples

Classes: [0, 1, 2] (Corresponding to different wine categories)

Features: alcohol, malic acid, ash, alkalinity_of_ash, magnesium ,flavonoids etc.

Majority percentage: Varies depending on class distribution

Abalone:

Dataset that try to predict the age of abalone through its physical characteristics.

Functions: Length, Diameter, Height, Whole_weight, Shucked_weight, Viscera_weight, Shell_weight, Class_number_of_rings

Target column: Rings (represents the abalone's age in number of rings).

Problem in the code: The Rings is not correctly identified (incorrect column name is used)

Adult

Dataset to predict income level (above/below 50K) based on demographic and work-related attributes.

Features: age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, etc.

Target column: income (represents whether income is >50K or <=50K).

Problem in the code: The code is looking for the wrong target column which causes an error.

Digit

Handwritten digit data set for classification, where each image is represented by pixel values.

Features: Contains 64 features (one for each pixel in an 8x8 image)

Target column: class (represents the number 0-9 as shown in the image).

Problem in the code: The target column is not correctly identified in the dataset, which causes an error.

4. Description of the models used in the miniproject

KMeans (unsupervised learning):

A clustering algorithm that make groups data points based on similarity. The number of clusters is set to the uniqe number in the target variable.

Performance is measured by the silhouette which show how well clusters are separated.

DecisionTreeClassifier (supervised learning):

A classification algorithm that creates a tree structure based on decisions to predict the target variable.

Performance is measured by accuracy, which indicates the proportion of correct predictions on the test data.

5. Description of the experimental methodology (datasets' splits, cross-validation, performance metricsetc)

Dataset Splits:

The datasets are split into 70% for training and 30% for testing using train_test_split.

This helps to train models on a portion of the data and test its performance on unseen data.

Cross-Validation:

Det finns ingen avancerad korsvalidering (som k-fold) som används i koden.

70/30-delningen fungerar som ett enkelt sätt att testa modellen på ny data.

Performance Matrics:

Silhouette Score:

Used to check how well the data is grouped into clusters by the supervised model (KMeans). A score closer to 1 means that the clusters are good.

Accuracy:

Used to measure how well the supervised model (DecisionTreeClassifier) predicts the correct labels. Higher accuracy means better predictions.

6. Description of the experimental results

rice dataset:

Unsupervised (Silhouette score): 0.4799

Supervised (Accuracy): 1.0000

Wine dataset:

Unsupervised (Silhouette score): 0.2849

Supervised (Accuracy): 0.9630

For Abalone, Adult, and Digits, errors occur because the target columns are not found.

7. Conclusions

The Iris dataset performs very well for both unsupervised and supervised models, indicating that the data is well-structured and easy to cluster and classify.

The Wine dataset has a slightly lower silhouette score, indicating that the clusters are not as well separated. However, the high accuracy shows that the supervised model is very effective.

The errors that occur for the Abalone, Adult and Digits datasets indicate that the code's handling of target columns needs to be improved

Länk till GitHub
https://github.com/mohraj-0/D7041E