Bahria University

Lahore Campus

Data Mining

Assignment # 2

Submitted To:

Mr. Muhmmad Mudassir

Submitted By:

Murtaza Anwaar (03-134211-032)

Contents

1. Dataset and Preprocessing

For this assignment, the **Mall_Customers** dataset was chosen, which was sourced from **Kaggle**. This dataset contains 200 samples with various features such as Age, Annual Income, Spending Score, etc. The data was preprocessed as follows:

- **Dropped Unnecessary Features:** Unneeded columns were removed from the dataset to streamline the analysis.
- **Label Encoding:** Categorical variables were encoded into numerical values using label encoding.
- **Feature Scaling:** The dataset was scaled to standardize the feature values and improve model performance.
- **Missing Values:** There were no missing values, so no further imputation was required.

1. Unsupervised Learning Techniques

Two unsupervised learning techniques were applied to the dataset:

1.1. K-Means Clustering

This technique was used to identify clusters in the data.

- *Silhouette Score: 0.36*, indicating the clusters were not very well separated.

1.2. Agglomerative Clustering

Another clustering technique is based on hierarchical clustering.

- ***Silhouette Score: 0.44***, which showed a slightly better clustering performance than K-Means but still indicates room for improvement.

2. Feature Selection

The Recursive Feature Elimination (RFE) technique was used to identify the most important features for classification. RFE recursively removes less important features and retains the ones that are most predictive. After applying RFE, the selected features were:

- Annual Income

- Spending Score

These features were retained for further experiments.

3. Supervised Learning Classifiers

Five supervised learning classifiers were applied to the dataset to predict customer behavior:

3.1. Random Forest Classifier

- Accuracy: 0.95

3.2. Support Vector Machine (SVM)

- Accuracy: 1.00 (Perfect accuracy)

3.3. K-Nearest Neighbors (KNN)

- Accuracy: 1.00 (Perfect accuracy)

3.4. Gradient Boosting

- Accuracy: 0.93

3.5. XGBoost

- Accuracy: 0.93

These classifiers were evaluated using various performance metrics such as accuracy, precision, recall, F1-score, and the confusion matrix.

The results showed that SVM and KNN achieved perfect accuracy (1.00), while Random Forest, Gradient Boosting, and XGBoost performed slightly lower, but still achieved high accuracy (~0.93-0.95).

4. Second Dataset: Iris Dataset and Manual Feature Extraction

For the second part of the assignment, the Iris Dataset was chosen. This dataset contains 150 samples with 4 features describing Iris flowers, including sepal length, sepal width, petal length, and petal width. The goal was to extract features that could improve classifier performance.

The following manual feature extraction techniques were applied:

- Mean and Standard Deviation: The mean and standard deviation of each feature were calculated for the entire dataset and then repeated for each individual sample to create additional features.

This method improved the accuracy of the classifier, *bringing accuracy to 100%.*

5. Automated Feature Extraction

Two automated feature extraction techniques were applied using Weka:

5.1. Principal Component Analysis (PCA)

This technique reduced the dimensionality of the dataset by transforming the features into principal components, retaining the most important features while discarding less relevant ones.

5.2. Information Gain

This method was used to calculate the relevance of each feature in predicting the target variable. Features with higher information gain were retained, and features with lower information gain were discarded.

Both PCA and Information Gain were applied to reduce the feature space and improve classifier performance.

6. GitHub Repository

All the content for this assignment, including datasets, code files, results, and images, has been uploaded to the following GitHub repository:

MA-247/Data-Mining---Assignment-2