DATA MINING PROJECT

Muhammad Adeel (20i-0722) & Faziha Ikhlaq (20i-0473)

SUBMISSION DATE: 12/05/2024

BS-CS DATA MINING (SECTION-A)

Task 1

Explanation:

In task 1 of the data mining project, exploratory data analysis (EDA) was conducted using Python's Pandas, Matplotlib, and Seaborn libraries. The script loads training and testing datasets from CSV files, displaying the first few rows of each dataset to provide an initial insight into the data's structure. Subsequently, it visualizes the distribution of numeric features in the training dataset through histograms. This EDA process serves as a foundational step in understanding the data's characteristics, facilitating further analysis and modeling in subsequent tasks of the project.

Task 2

Explanation:

- **Components:**

- Utilizes TensorFlow and Keras libraries.

- Comprises TransformerEncoder, Discriminator, and ContrastiveLearning layers.

- **TransformerEncoder:**

- Extracts features using self-attention mechanisms.

- Captures sequential dependencies in the data.

- Utilizes positional encoding for encoding sequence information.

- **Discriminator:**

- Identifies anomalies in the data.

- Utilizes Multi-Head Attention and point-wise feed-forward networks.

- **Contrastive Learning:**

- Enhances model performance in anomaly detection.

- Incorporates Multi-Head Attention and feed-forward networks.

- **Training Process:**

- Utilizes mean squared error loss function.

- Optimized using the Adam optimizer.

- Iterates through epochs to train the model.

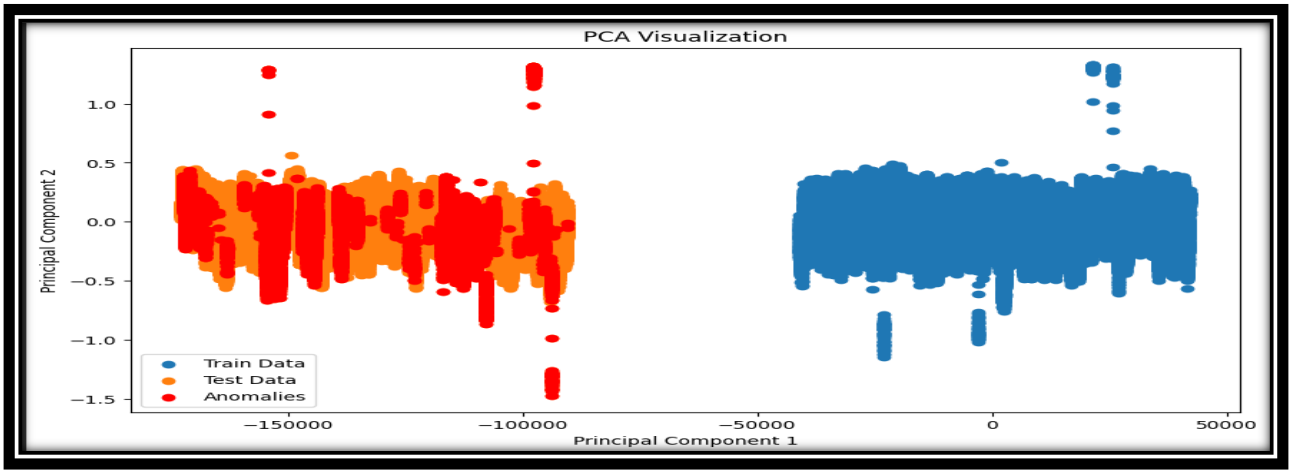
- Adjustments such as batch size and epochs can be made for optimization.

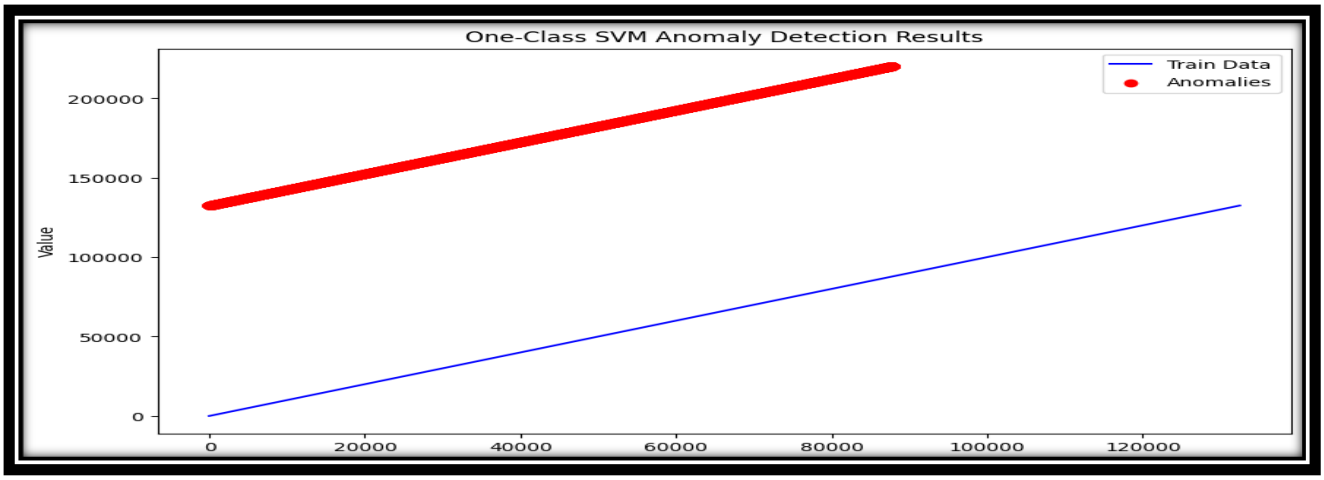
Task 3:

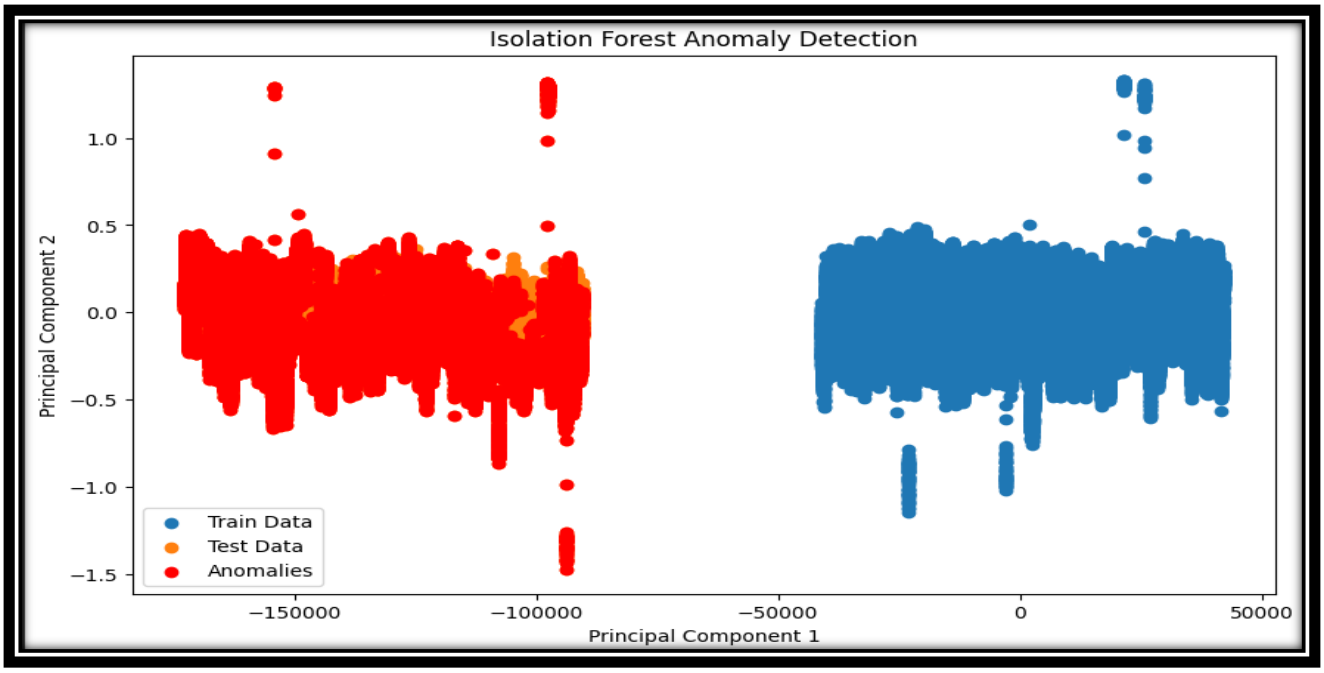
Explanation:

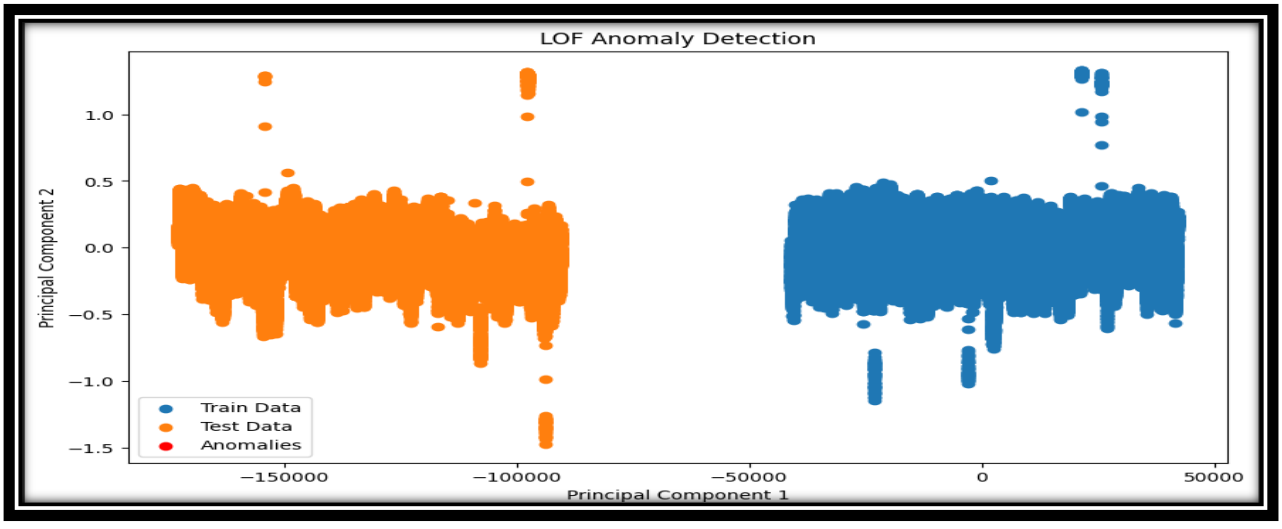
In Task 3 of the data mining project, the emphasis lies on anomaly detection using a range of methodologies. The process initiates with Principal Component Analysis (PCA), which serves as a dimensionality reduction technique to extract pertinent features from the dataset. PCA transforms the dataset into a lower-dimensional space, facilitating easier visualization and subsequent anomaly identification. Following PCA, methods such as One-Class SVM, Isolation Forest, Local Outlier Factor (LOF), and DBSCAN are employed for anomaly detection. One-Class SVM detects anomalies by considering a single class of data, while Isolation Forest identifies anomalies by pinpointing data points with shorter path lengths within trees. LOF computes the local density deviation of each data point, enabling the detection of outliers. DBSCAN, a density-based clustering algorithm, partitions the data into clusters and labels points outside dense regions as anomalies. Additionally, two bonus techniques, Dynamic Time Warping (DTW) and Seasonal Hybrid Extreme Studentized Deviate (S-H-ESD), offer specialized anomaly detection capabilities for time series data. DTW measures the similarity between time series, while S-H-ESD detects anomalies through seasonal decomposition and statistical analysis. These techniques collectively provide a comprehensive approach to anomaly detection across various data types and structures.

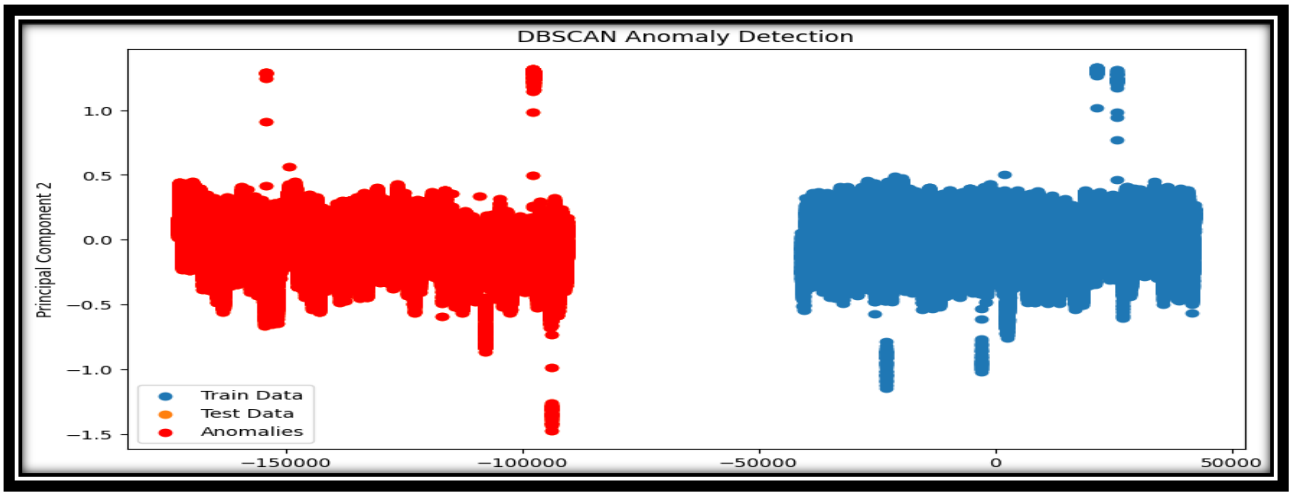
Results:




One-Class SVM Anomaly Detection Results
Train Data
Anomalies
Value
200000
150000
100000
50000
0
0
20000
40000
60000
80000
100000
120000


Isolation Forest Anomaly Detection
Principal Component 2
1.0
0.5
0.0
−0.5
−1.0
−1.5
Train Data
Test Data
Anomalies
−150000
−100000
−50000
0
50000
Principal Component 1


LOF Anomaly Detection
Principal Component 2
1.0
0.5
0.0
−0.5
−1.0
−1.5
Train Data
Test Data
Anomalies
−150000
−100000
−50000
0
50000
Principal Component 1



Task 4:

Explanation:

Task 4 of the data mining project entails performing an empirical examination of anomaly detection techniques, particularly focusing on comparing the efficacy of PCA-based anomaly detection with that of Isolation Forest. The process commences with preliminary data preparation, encompassing tasks such as managing missing values and standardizing features. For PCA-based anomaly detection, PCA is applied to extract principal components capturing 95% of the variance. Reconstruction errors are calculated using mean squared error (MSE) between original and reconstructed data. Anomalies are determined based on a selected threshold, and model performance is evaluated using classification report metrics.

On the other hand, Isolation Forest is implemented separately. Missing values are handled similarly, and the Isolation Forest model is trained on the preprocessed data. Predictions are generated, and model performance is evaluated using classification report metrics and ROC AUC score. The empirical analysis includes comparing F1 scores, time taken for training and prediction, and the AUC score for both methods. These metrics are visualized to provide a comprehensive comparison between PCA-based anomaly detection and Isolation Forest, aiding in understanding their relative effectiveness and efficiency in anomaly detection tasks.

Results:

PCA Model:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	63460
1	0.28	1.00	0.43	24381
accuracy			0.28	87841
macro avg	0.14	0.50	0.22	87841

weighted avg	0.08	0.28	0.12	87841

Isolation forest Model:

	precision	recall	f1-score	support
0	0.75	0.94	0.83	63460
1	0.51	0.18	0.26	24381
accuracy			0.72	87841
macro avg	0.63	0.56	0.55	87841
weighted avg	0.68	0.72	0.67	87841