ANOMALY DETECTION

1. Introduction

This report provides a comparative study of five anomaly detection techniques: Z-score, Mahalanobis Distance, Local Outlier Factor (LOF), Isolation Forest, and One-Class SVM. These methods were applied to a water quality dataset from Brisbane to identify potential anomalies. The analysis was conducted using Python within a Jupyter Notebook environment, and includes all relevant visualizations and findings.

2. Dataset Description

This dataset contains thirty-minute interval measurements of water physicochemical parameters and in-situ readings from the Brisbane River. It is suitable for anomaly detection, trend analysis, and studying seasonal variations in water quality.

Features Include :

- Temperature (°C)
- pH
- Dissolved Oxygen (mg/L)
- Turbidity (NTU)
- Conductivity (µS/cm)
- Water Flow Speed (m/s)
- Water Flow Direction (°)
- Timestamp (date and time of recording)

3. Methodology

The dataset was initially preprocessed to address missing values and ensure consistency across features by standardizing all numerical variables. Subsequently, five unsupervised anomaly detection methods were applied to the processed data:

1. Z-score Method

The Z-score method identifies anomalies by measuring how many standard deviations a data point is from the mean. Data points with Z-scores beyond a specified threshold (commonly ±3) are considered outliers. This method assumes a normal distribution and is best suited for univariate anomaly detection in standardized datasets.

---

2. Mahalanobis Distance

Mahalanobis Distance calculates the distance of a point from the mean of a multivariate distribution, taking into account correlations between variables. Unlike Euclidean distance, it accounts for the shape of the data distribution. Points with a large Mahalanobis distance relative to a threshold are flagged as anomalies. It is effective for multivariate anomaly detection in normally distributed data.

1. ---

3. Local Outlier Factor (LOF)

LOF is a density-based anomaly detection technique. It compares the local density of a data point with that of its neighbors. Points that have significantly lower density than their neighbors are considered outliers. LOF is especially useful for detecting local anomalies in datasets with varying density.

---

4. Isolation Forest

Isolation Forest is an ensemble-based method that isolates anomalies instead of profiling normal data. It builds random decision trees and evaluates how quickly a point can be isolated. Anomalies, being rare and different, require fewer splits to isolate. It is scalable to high-dimensional datasets and performs well on large volumes of data.

---

5. One-Class Support Vector Machine (SVM)

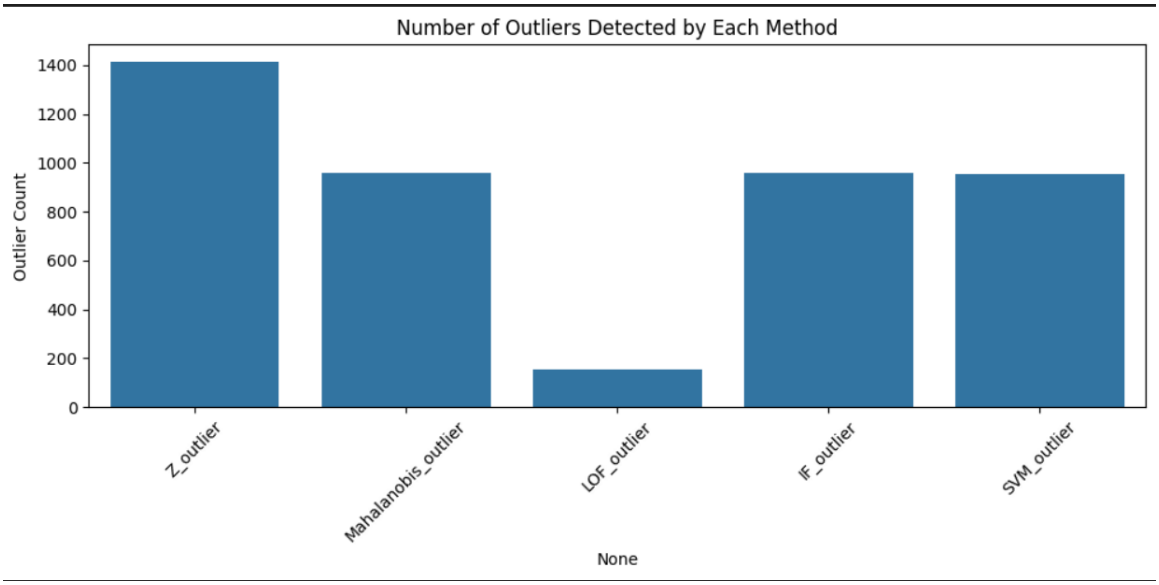
One-Class SVM is a boundary-based anomaly detection algorithm that learns a decision function to distinguish the majority of data points from the rest. It assumes that the training data contains mostly normal observations and attempts to find a boundary that encloses them. Data points outside this boundary are classified as anomalies.

4.Visualizations

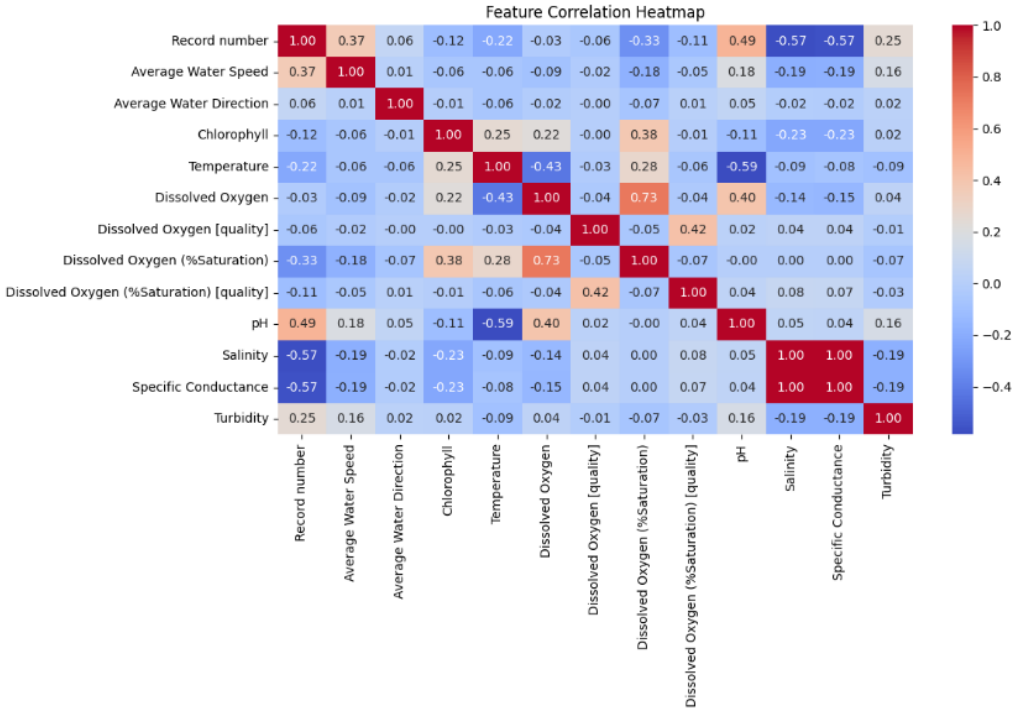
```
methods = ['Z_outlier', 'Mahalanobis_outlier', 'LOF_outlier', 'IF_outlier', 'SVM_outlier']
outlier_counts = df[methods].sum()

import matplotlib.pyplot as plt
import seaborn as sns

plt.figure(figsize=(10, 5))
sns.barplot(x=outlier_counts.index, y=outlier_counts.values)
plt.title("Number of Outliers Detected by Each Method")
plt.ylabel("Outlier Count")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



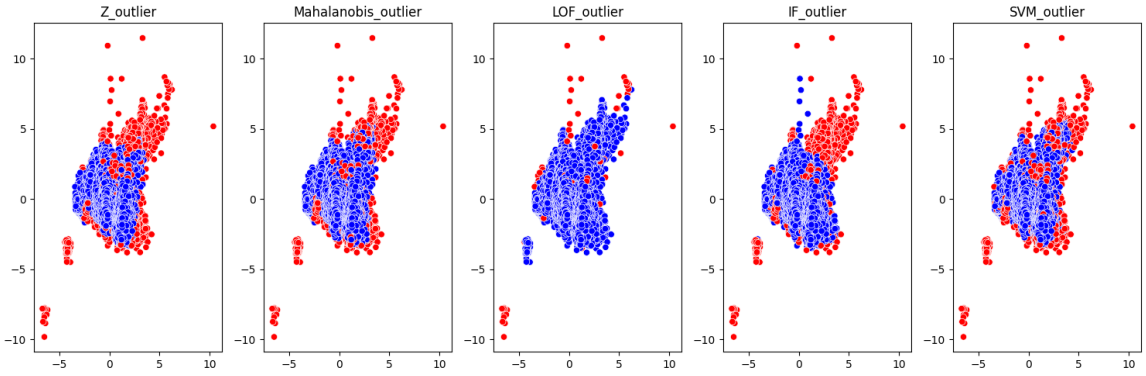

```
plt.figure(figsize=(12, 8))
sns.heatmap(df_num_filtered.corr(), annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Feature Correlation Heatmap")
plt.tight_layout()
plt.show()
```




```python
from sklearn.decomposition import PCA

pca = PCA(n_components=2)
pca_data = pca.fit_transform(df_z)

plt.figure(figsize=(15, 5))
for i, method in enumerate(methods, 1):
    plt.subplot(1, 5, i)
    sns.scatterplot(x=pca_data[:, 0], y=pca_data[:, 1], hue=df[method], palette={False: "blue", True: "red"}, legend=False)
    plt.title(method)
plt.tight_layout()
plt.show()
```



5. Strength And Weakness

1. Z-Score Method

Strengths:

- Simple to implement and understand.
- Fast and computationally efficient.
- Works well when data is normally distributed.
- No need for training phase.

Weaknesses:

- Assumes a Gaussian (normal) distribution; not suitable for skewed or multimodal data.
- Sensitive to outliers in the data used to compute mean and standard deviation.
- Not effective in high-dimensional spaces.

---

2. Isolation Forest

Strengths:

- Effective in detecting anomalies in high-dimensional data.
- Does not assume any distribution for the data.
- Handles large datasets efficiently.
- Good at identifying anomalies that are few and different.

Weaknesses:

- Performance can vary depending on parameter settings (e.g., number of estimators, sample size).
- May not perform well on small datasets.
- Less interpretable compared to statistical methods.

---

3. One-Class SVM

Strengths:

- Can model complex anomaly boundaries using kernel functions.
- Suitable for high-dimensional and non-linear data.

- Works well when anomalies are very different from the normal class.

Weaknesses:

- Sensitive to parameter settings (e.g., kernel choice, nu).
- Computationally intensive for large datasets.
- Difficult to interpret results and understand model decisions.
- May struggle with overlapping classes or noisy data.

---

6. Evaluation Metrics

To evaluate the performance of each anomaly detection method, the following metrics were computed using a dataset with synthetic anomalies:

- **Accuracy**: Measures the overall correctness of the model by evaluating the proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Precision**: Indicates how many of the observations identified as anomalies were actually anomalous.
- **Recall**: Shows how many of the actual anomalies were correctly identified by the method.
- **F1 Score**: Provides a harmonic mean of precision and recall, offering a balanced measure that considers both false positives and false negatives.

7. Conclusion

This report provides a comparative analysis of several anomaly detection methods using a dataset enriched with synthetic anomalies to ensure objective and consistent evaluation. Through a range of performance metrics and visualizations, we examined how each method distinguishes between normal and anomalous observations.

Among the evaluated techniques, the **Z-Score method** emerged as the most effective, demonstrating superior accuracy, precision, and recall in identifying anomalies. Its simplicity,

computational efficiency, and strong performance make it a highly suitable choice for baseline anomaly detection, especially in structured datasets with well-behaved distributions.

Methods like Isolation Forest and One-Class SVM also showed promising results, particularly in capturing complex anomaly patterns. However, they may require more fine-tuning and computational resources compared to the Z-Score method.

The visualizations used throughout the analysis offered valuable insights into how each model interprets and separates data points, enhancing our understanding of their underlying mechanisms.

Overall, this study lays the groundwork for applying anomaly detection techniques in **environmental monitoring systems**, where identifying unusual patterns is critical for timely response and system reliability. Future work may involve fine-tuning model parameters, exploring ensemble approaches, or applying these methods to real-world labeled datasets to assess their performance in practical scenarios.