# Day 40

Date: 26 February 2026

Topic Studied: Outliers – Detection, Types & Handling using Isolation Forest

Mode: Practical + Theory

Tools Used: Python, Machine Learning Libraries (Scikit-learn, Pandas, NumPy)

## Objective of the Session

The objective of today's training session was to understand what outliers are, explore their different types, learn why they are problematic in data analysis, and implement methods to detect and handle them, particularly using the Isolation Forest algorithm.

## What are Outliers?

Outliers are data points that significantly differ from the majority of the data. These abnormal values can appear due to:

- Measurement errors
- Data entry mistakes
- Experimental anomalies
- Rare but valid events

**Outliers can negatively impact machine learning models, leading to:**

- Biased predictions
- Reduced model accuracy
- Increased error rates
- Hence, detecting and handling outliers is a crucial step in data preprocessing.

## Types of Outliers

**1. Global Outliers (Point Outliers)**

- These are data points that stand far away from the rest of the dataset.
- Example: In a dataset of student marks ranging from 40 to 95, a value like 5 or 150 would be considered a global outlier.

**2. Contextual Outliers**

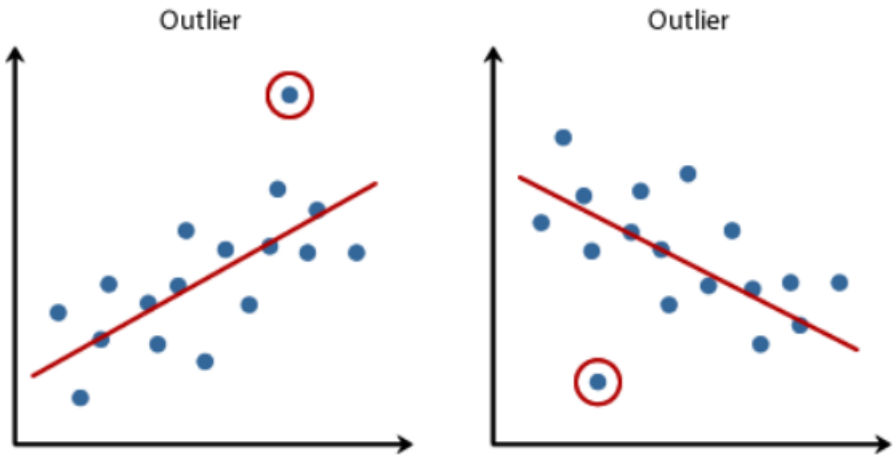
- These values are considered outliers only in a specific context.
- Example: A temperature of 30°C is normal in summer but abnormal in winter.

**3. Collective Outliers**

- A group of related data points that together behave abnormally.
- Example: A sudden spike in network traffic over a short time period.

**Problems Caused by Outliers**

- Distort statistical measures such as mean and standard deviation
- Affect model training and prediction accuracy
- Cause overfitting or underfitting
- Reduce reliability of analysis




**Methods to Detect and Handle Outliers**

- Some common techniques include:
- Z-score method
- Interquartile Range (IQR) method
- DBSCAN
- Isolation Forest

## Isolation Forest

What is Isolation Forest?

Isolation Forest is an unsupervised machine learning algorithm used for outlier detection. It works on the principle that:

- Outliers are easier to isolate than normal data points.

- Instead of profiling normal data, it isolates anomalies using randomly generated decision trees.

## How Isolation Forest Works

1. Randomly selects a feature
2. Randomly selects a split value
3. Builds multiple isolation trees
4. Data points that get isolated quickly are identified as outliers

**Advantages of Isolation Forest**

- Works well for large datasets
- Computationally efficient
- Handles high-dimensional data
- Does not assume any data distribution

## Practical Implementation

During the session, Python code was implemented to:

1. Load dataset

```
# Load dataset
import pandas as pd
df = pd.read_csv(path + "/healthcare-dataset-stroke-data.csv")
df.head(20)
```

2. Visualize data distribution

3. Apply Isolation Forest model

```
ist = IsolationForest(random_state=42)

outliers = ist.fit_predict(x_train)   # -1 = outlier, 1 = normal

mask = outliers == 1                  # keep only normal points
```

4. Detect anomalies

5. Remove or flag outliers

6. Analyze clean data

```
original dataset shape: (4088, 11)
Original dataset shape after smote: (7778, 11)
Accuracy: 87.76908023483367
```