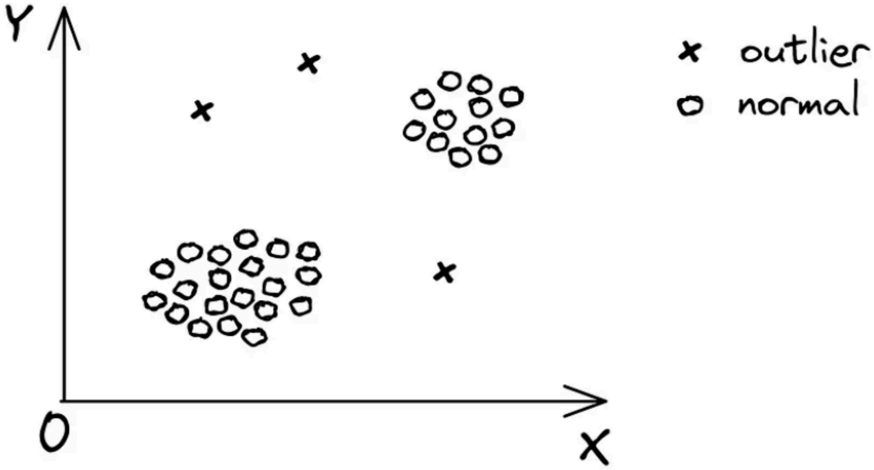
Name: Romesh Umesh Lulla **Class:** D17C **Roll No.:** 31

ADS Experiment 6

Aim: To implement Outlier detection using distance based/density based method

Theory:
Outlier detection is a crucial task in data analysis and machine learning, aimed at identifying observations that deviate significantly from the majority of the data points. Two common methods for outlier detection are distance-based and density-based approaches. Let's delve into the theory behind each of these methods:



1. Distance-based Outlier Detection:

The distance-based approach to outlier detection focuses on measuring the dissimilarity or distance between data points. The fundamental assumption is that outliers are located far away from the dense regions of the data distribution. Common distance metrics used for this purpose include Euclidean distance, Manhattan distance, Mahalanobis distance, and Minkowski distance.

- Nearest Neighbor Methods: One common technique is to compute the distance between a data point and its k-nearest neighbors. If a data point has significantly larger distances to its neighbors compared to the rest of the dataset, it is considered an outlier.

- DBSCAN (Density-Based Spatial Clustering of Applications with Noise): While primarily a density-based clustering algorithm, DBSCAN can also be used for outlier detection. It identifies outliers as data points that lie in low-density regions or regions with insufficient neighboring points.

- Isolation Forest: It's an ensemble method based on decision trees that isolates outliers by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The number of splits required to isolate the outlier is used as a measure of its abnormality.

- **Elliptical Envelope:** The Elliptical Envelope method assumes that inlying data follows a multivariate Gaussian distribution, estimating a robust covariance matrix to capture data relationships. Outliers are identified as points lying outside an ellipse constructed based on this covariance matrix and the data mean. The Mahalanobis distance measures outlier deviation from the ellipse. Widely used in fraud detection and anomaly detection, Elliptical Envelope offers an efficient approach for outlier detection in multivariate datasets.

2. Density-based Outlier Detection:

Density-based methods focus on identifying regions of low data density, assuming that outliers exist in these sparsely populated areas. These methods typically cluster the data into dense regions and label data points in low-density regions as outliers.

- Local Outlier Factor (LOF): LOF measures the density of a data point with respect to its neighbors. It assigns an outlier score based on the ratio of the local density of a data point to the local densities of its neighbors. Data points with significantly lower density compared to their neighbors are flagged as outliers.

- Mean Shift: This method identifies outliers by iteratively shifting data points towards the mode of the underlying data distribution. Outliers are detected as data points that fail to converge towards the mode within a certain number of iterations.

- OPTICS (Ordering Points To Identify the Clustering Structure): Similar to DBSCAN, OPTICS identifies outliers by clustering the data based on density. Outliers are detected as points lying in regions with low density or insufficient neighboring points.

- Local Density Estimation: Some methods estimate the density of data points within a defined radius or neighborhood. Data points with densities below a certain threshold are considered outliers.

In summary, distance-based methods focus on measuring the distance between data points, while density-based methods emphasize the local density of data points. Both approaches offer effective ways to detect outliers in various types of datasets, each with its own strengths and weaknesses depending on the characteristics of the data.

Conclusion:

Outlier detection methods play a crucial role in data analysis and machine learning by identifying observations that deviate significantly from the majority of the dataset. Various techniques, including distance-based, density-based, and model-based approaches, offer diverse strategies for detecting outliers based on different underlying assumptions about the data distribution and characteristics. While each method has its strengths and weaknesses, the choice of technique depends on the nature of the data and the specific requirements of the problem at hand. Overall, the selection and application of appropriate outlier detection methods are essential for improving data quality, enhancing decision-making processes, and ensuring the reliability of machine learning models in real-world applications.