

# Data Science

## Unit –IV

B.Tech. VI Semester AIM  
2024-25

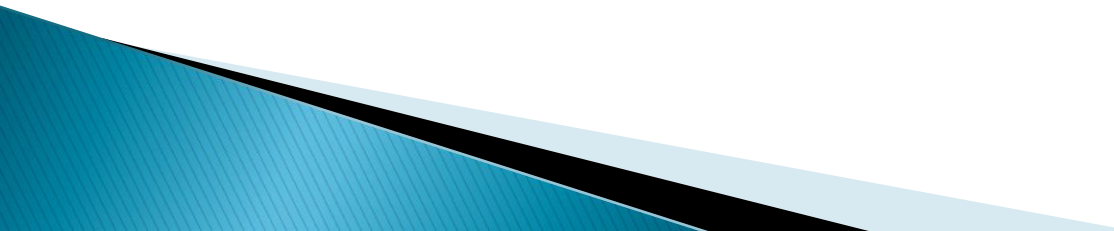
By

Dr. Loshma Guniseti,  
Professor and Head , AIML Dept.

## **Text Book:**

Data Science Concepts and Practice, Vijay Kotu, Bala Deshpande, 2nd Edition, Morgan Kaufmann Publishers.

## **Reference Books:**

- An Introduction to Data Science, Jeffrey S. Saltz, Jeffrey M. Stanton, Sage Publications.
  - The Art of Data Science, Roger D Peng, Elizabeth Matsui, Lean Publishing.
  - Data Science for Business, Foster Provost, Tom Fawcett, O'Reilly Media.
- 

# Course Outcomes

After successful completion of this course, the student will be able to:

CO	Course Outcomes	Knowledge Level
1	Discuss the fundamental concepts of Data Science.	K2
2	Illustrate Exploratory Data Analysis.	K2
3	Explain the Concepts of Recommendation Engines.	K2
4	<b>Explain various Anomaly Detection Techniques.</b>	<b>K2</b>
5	Discuss Feature Selection techniques.	K2

# Syllabus

**UNIT-I: Introduction:** AI, Machine Learning and Data Science, What is Data Science? Case for Data Science, Data Science Classification, Data Science Algorithms.

**Data Science Process:** Prior Knowledge, Data Preparation, Modeling-Training and Testing Datasets, Learning Algorithms, Evaluation of the Model, Ensemble Modeling, Application, Knowledge.

**UNIT-II: Data Exploration:** Objectives of Data Exploration, Datasets- Types of Data, Descriptive Statistics-Univariate Exploration, Multivariate Exploration, Data Visualization, Roadmap for Data Exploration.

**UNIT-III: Recommendation Engines:** Need, Applications, Concepts, Types, Collaborative Filtering- Neighborhood-Based Methods, Matrix Factorization; Content-Based Filtering- Building an Item Profile, User Profile Computation, Implementation Steps, Hybrid Recommenders.

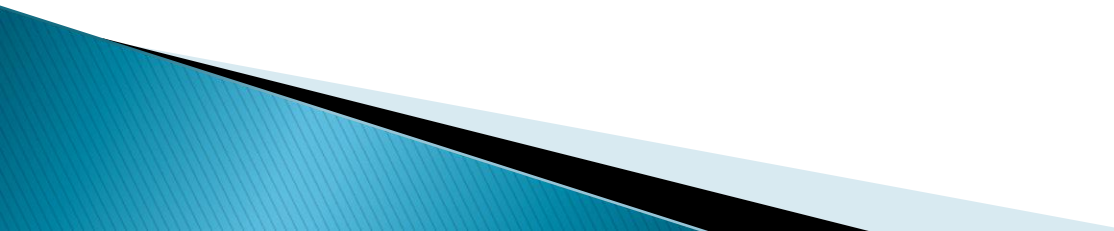
**UNIT-IV: Anomaly Detection:** Concepts - Causes of Outliers, Anomaly Detection Techniques; Distance-Based Outlier Detection- Working, Implementation Steps; Density-Based Outlier Detection- Working, Implementation Steps; Local Outlier Factor- Working, Implementation Steps.

**UNIT-V: Feature Selection:** Classifying Feature Selection Methods, Principal Component Analysis, Information Theory-Based Filtering, Chi-Square-Based Filtering, Wrapper-Type Feature Selection- Backward Elimination.

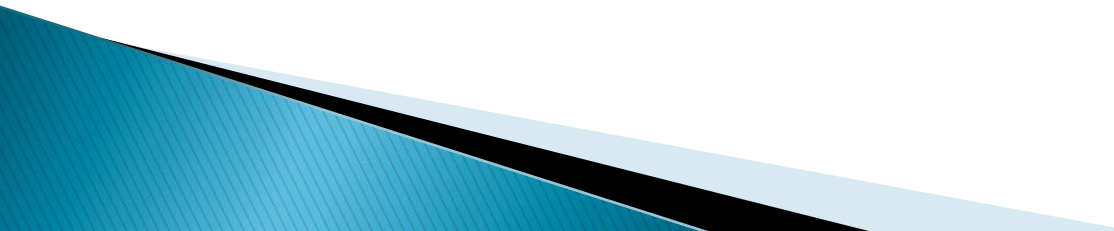
# Unit IV Contents

- Anomaly Detection
- Concepts
  - Causes of Outliers
- Anomaly Detection Techniques
- Distance-Based Outlier Detection
  - Working
  - Implementation Steps
- Density-Based Outlier Detection
  - Working
  - Implementation Steps
- Local Outlier Factor
  - Working
  - Implementation Steps

# Anomaly Detection

- Anomaly detection is the process of finding outliers in a given dataset.
  - Outliers are the data objects that stand out amongst other data objects and do not conform to the expected behavior in a dataset.
  - Anomaly detection algorithms have broad applications in business, scientific, and security domains where isolating and acting on the results of outlier detection is critical.
  - For identification of anomalies, algorithms such as classification, regression, and clustering can be used
- 

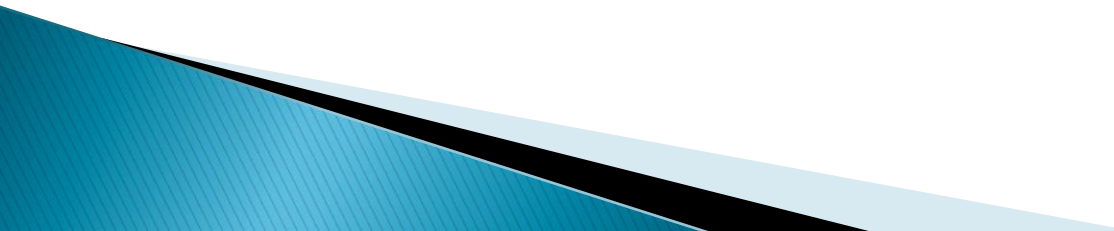
# Concept of Outlier

- An outlier is a data object that is markedly different from the other objects in a dataset. Hence, an outlier is always defined in the context of other objects in the dataset.
  - A high-income individual may be an outlier in a middle-class neighborhood dataset, but not in the membership of a luxury vehicle ownership dataset.
  - By nature of the occurrence, outliers are also rare and, hence, they stand out amongst other data points.
  - For example, the majority of computer network traffic is legitimate, and the one malicious network attack would be the outlier.
- 

# Causes of Outliers

- Outliers in the dataset can originate from either error in the data or from valid inherent variability in the data.
- Outliers are the data objects that stand out amongst other data objects and do not conform to the expected behavior in a dataset.
- Some of the most common reasons why an outlier occurs in the dataset:
  - **Data errors**

For example, in a dataset of human heights, a reading such as 1.70 cm is obviously an error and most likely was wrongly entered into the system.






# Causes of Outliers

- **Normal variance in the data**

An individual earning a billion dollars in a year or someone who is more than 7 ft tall falls under the category of outlier in an income dataset or a human height dataset respectively

- **Data from other distribution classes**

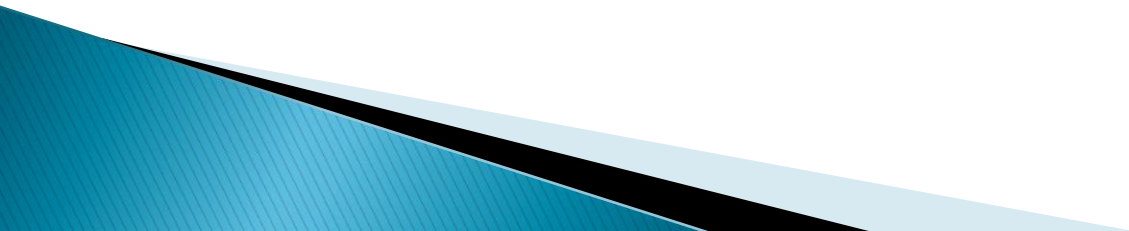
The number of daily page views for a customer-facing website from a user IP address usually range from one to several dozen. However, it is not unusual to find a few IP addresses reaching hundreds of thousands of page views in a day. This outlier could be an automated program from a computer (also called a bot) making the calls to scrape the content of the site or access one of the utilities of the site, either legitimately or maliciously



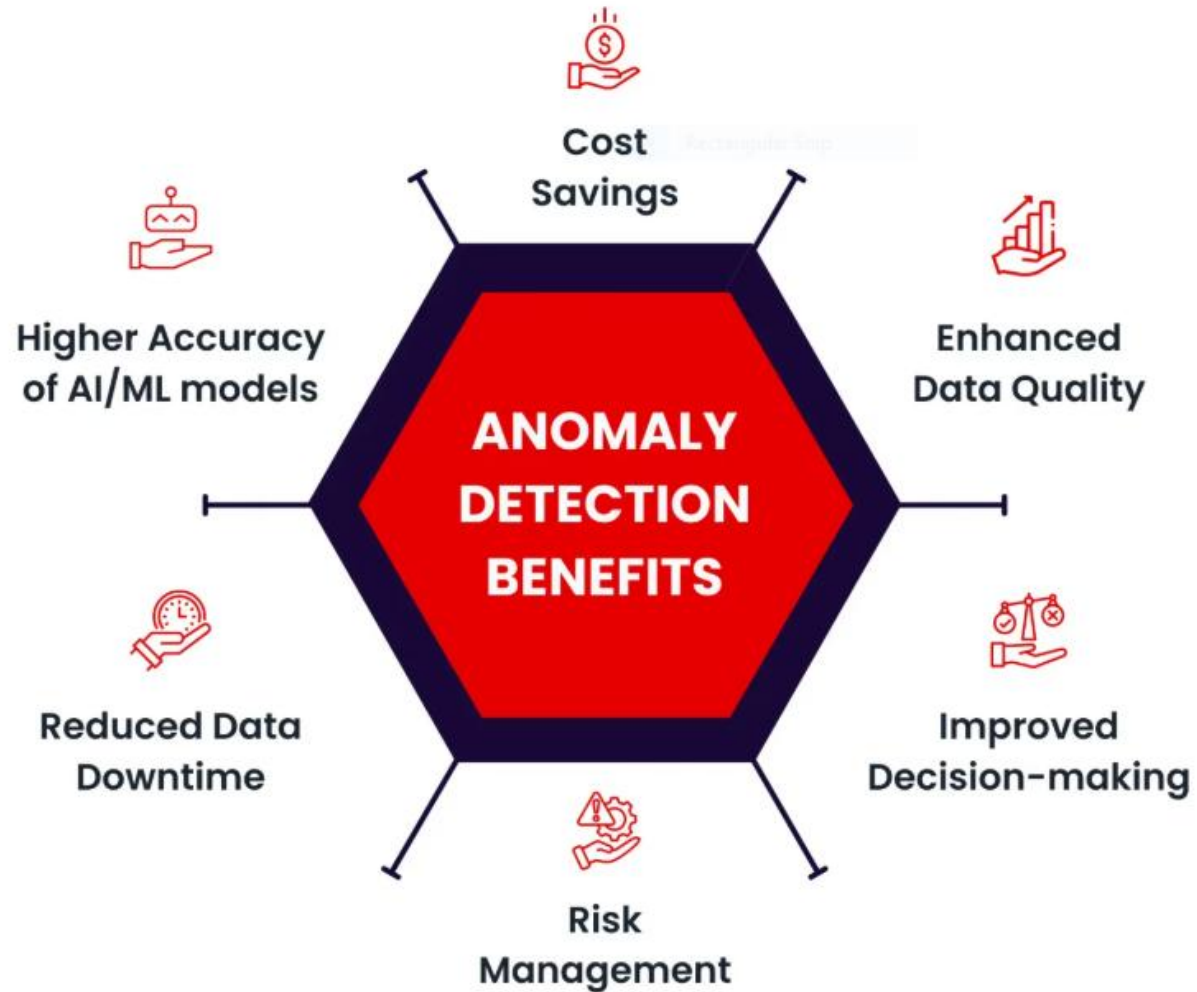
# Causes of Outliers

- **Distributional assumptions**

For example, if the data measured is usage of a library in a school, then during term exams there will be an outlier because of a surge in the usage of the library.

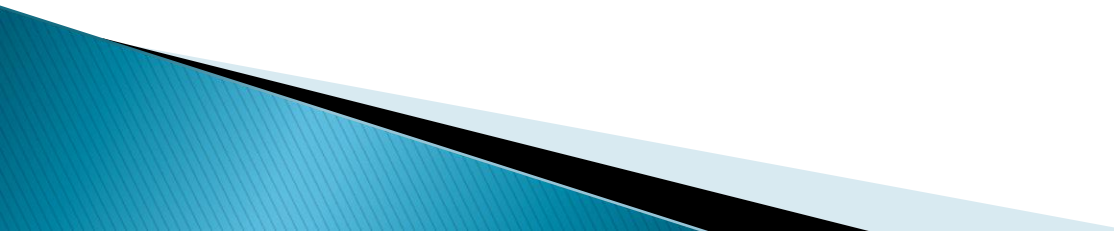


# Anomaly Detection



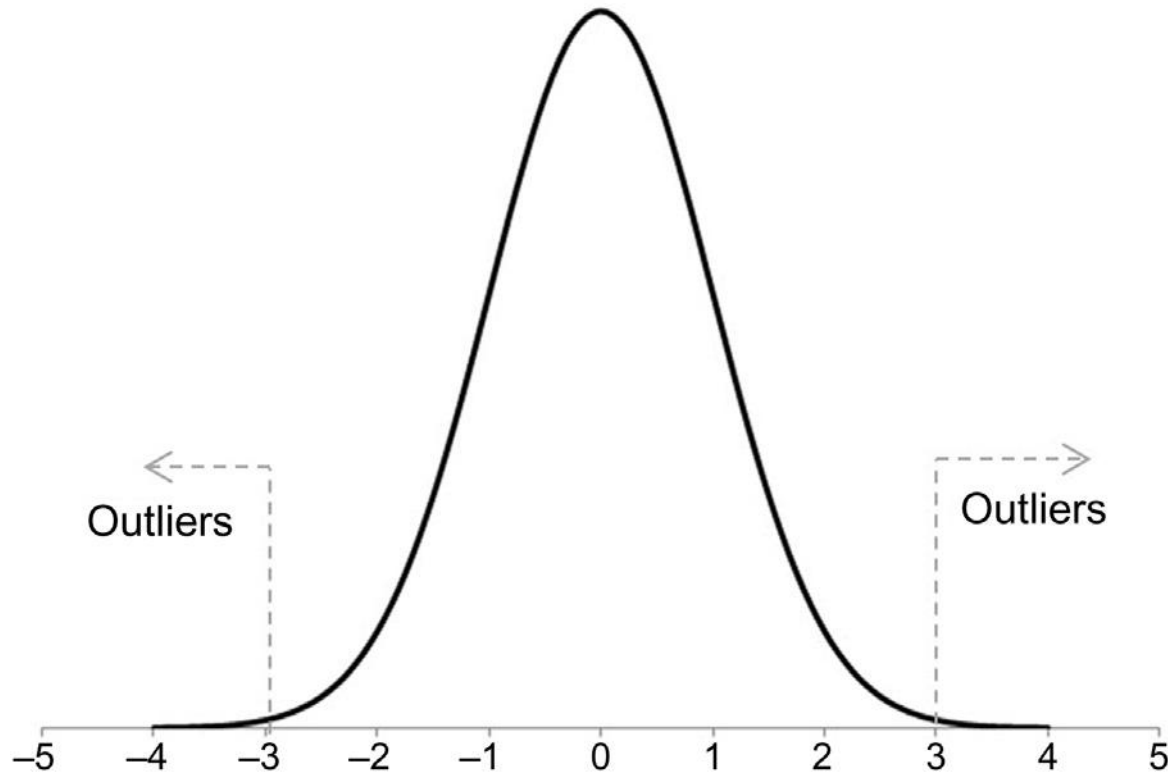
# Anomaly Detection

## Outlier Detection Using Statistical Methods

- Outliers in the data can be identified by creating a statistical distribution model of the data and identifying the data points that do not fit into the model or data points that occupy the ends of the distribution tails.
  - The underlying distribution of many practical datasets fall into the Gaussian (normal) distribution.
  - The parameters for building a normal distribution (i.e., mean and standard deviation) can be estimated from the dataset and a normal distribution curve can be created
- 

# Anomaly Detection

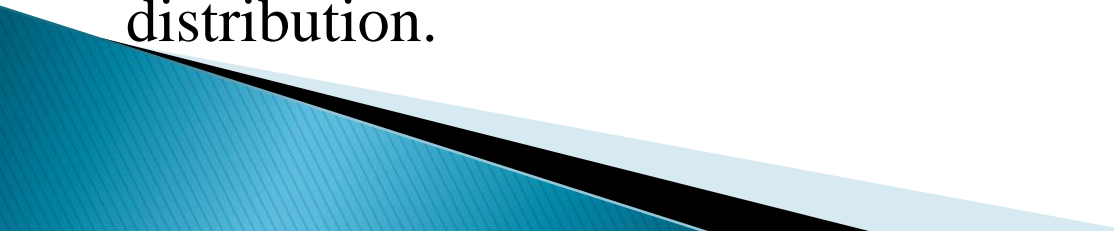
## Outlier Detection Using Statistical Methods



Standard normal distribution and outliers

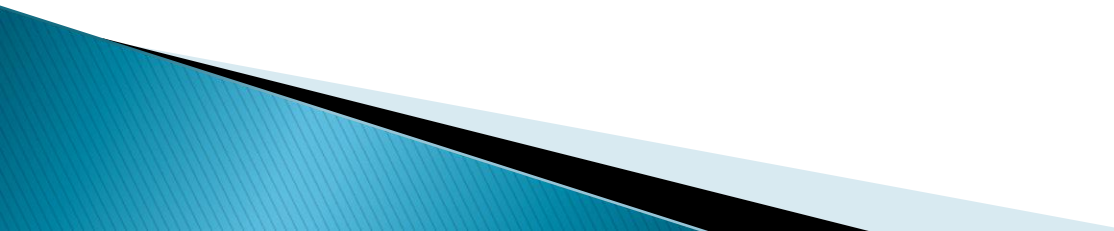
# Anomaly Detection

## Outlier Detection Using Statistical Methods

- Outliers can be detected based on where the data points fall in the standard normal distribution curve.
  - A threshold for classifying an outlier can be specified, say, three standard deviations from the mean.
  - Any data point that is more than three standard deviations is identified as an outlier.
  - Identifying outliers using this method considers only one attribute or dimension at a time.
  - More advanced statistical techniques take multiple dimensions into account and calculate the Mahalanobis distance instead of the standard deviations from the mean in a univariate distribution.
- 

# Anomaly Detection

## Outlier Detection Using Statistical Methods

- Mahalanobis distance is the multivariate generalization of finding how many standard deviations away a point is from the mean of the multivariate distribution.
  - Outlier detection using statistics provides a simple framework for building a distribution model and for detection based on the variance of the data point from the mean.
  - One limitation of using the distribution model to find outliers is that the distribution of the dataset is not previously known.
  - Even if the distribution is known, the actual data don't always fit the model
- 

# Anomaly Detection

## Outlier Detection Using Statistical Methods

### Mahalanobis distance

$$D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)$$

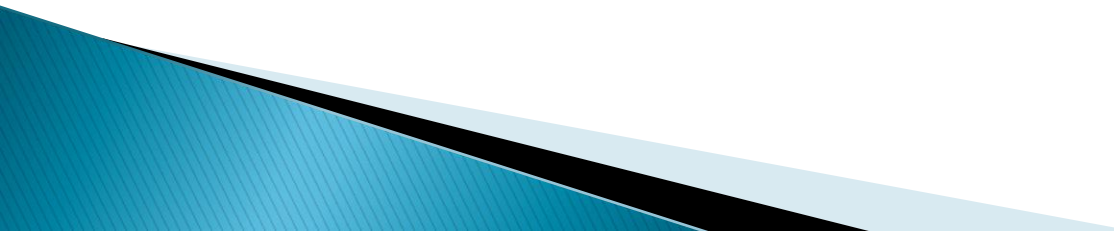
where, -  $D^2$  is the square of the Mahalanobis distance.

- $x$  is the vector of the observation (row in a dataset),
- $m$  is the vector of mean values of independent variables (mean of each column),
- $C^{-1}$  is the inverse covariance matrix of independent variables.



# Anomaly Detection

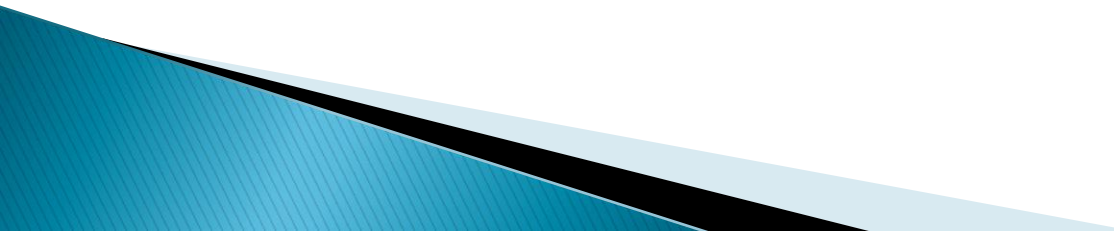
## Outlier Detection Using Data Science

- Outliers exhibit a certain set of characteristics that can be exploited to find them.
  - Following are classes of techniques that were developed to identify outliers by using their unique characteristics.
  - Each of these techniques has multiple parameters and, hence, a data point labeled as an outlier in one algorithm may not be an outlier to another.
    - Distance-based
    - Density-based
    - Distribution-based
    - Clustering
    - Classification techniques
- 

# Anomaly Detection

## Outlier Detection Using Data Science

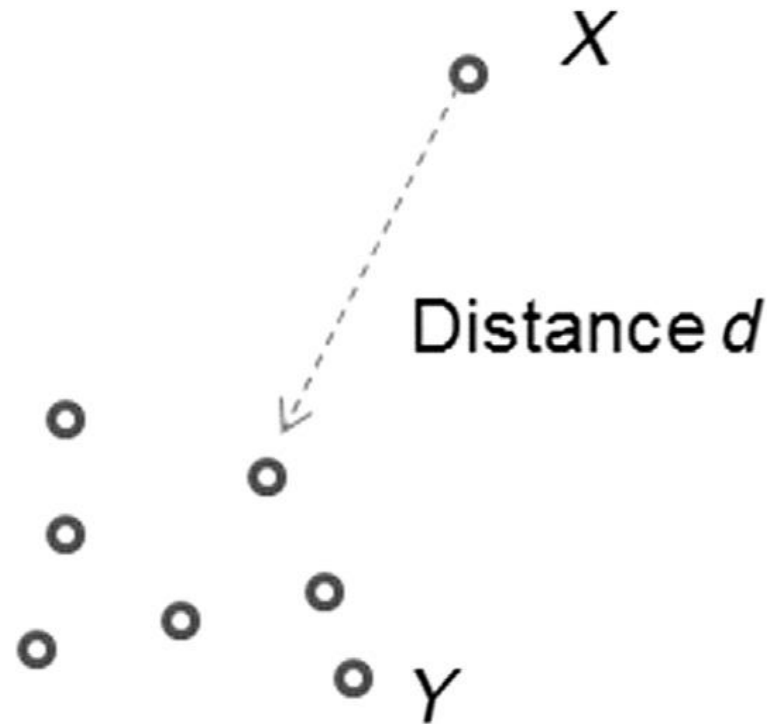
### Distance-based

- By nature, outliers are different from other data objects in the dataset.
  - In multidimensional Cartesian space they are distant from other data points.
  - If the average distance of the nearest  $N$  neighbors is measured, the outliers will have a higher value than other normal data points.
  - Distance-based algorithms utilize this property to identify outliers in the data
- 

# Anomaly Detection

## Outlier Detection Using Data Science

- Distance-based

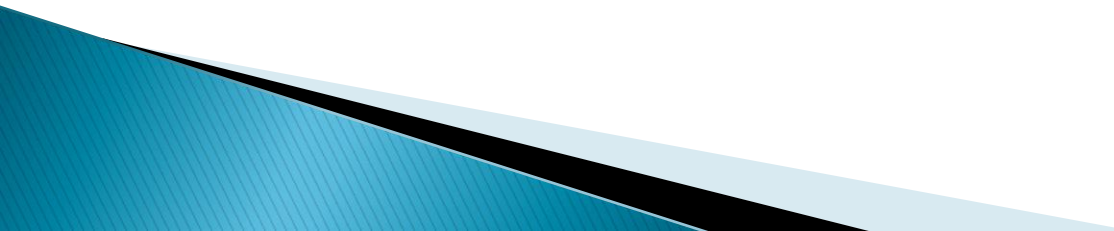


Distance-based outlier

# Anomaly Detection

## Outlier Detection Using Data Science

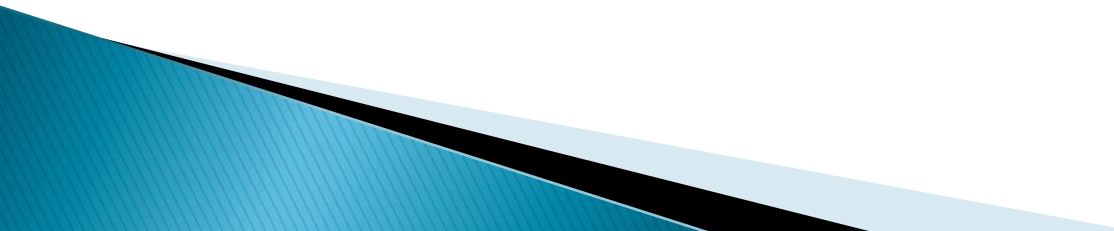
### Density-based

- The density of a data point in a neighborhood is inversely related to the distance to its neighbors.
  - Outliers occupy low-density areas while the regular data points congregate in high-density areas. This is derived from the fact that the relative occurrence of an outlier is low compared with the frequency of normal data points.
- 

# Anomaly Detection

## Outlier Detection Using Data Science

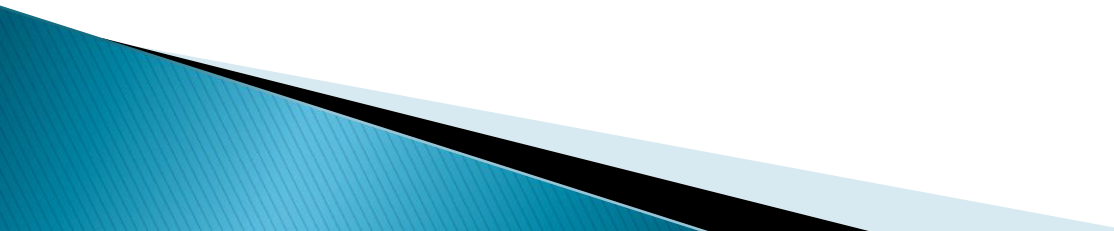
### Distribution-based

- Outliers are the data points that have a low probability of occurrence and they occupy the tail ends of the distribution curve.
  - So, if one tries to fit the dataset in a statistical distribution, these anomalous data points will stand out and, hence, can be identified.
  - A simple normal distribution can be used to model the dataset by calculating the mean and standard deviation.
- 

# Anomaly Detection

## Outlier Detection Using Data Science


### Clustering

- Outliers by definition are not similar to normal data points in a dataset. They are rare data points far away from regular data points and generally do not form a tight cluster.
  - Since most of the clustering algorithms have a minimum threshold of data points to form a cluster, the outliers are the lone data points that are not clustered.
  - Even if the outliers form a cluster, they are far away from other clusters.
- 

# Anomaly Detection

## Outlier Detection Using Data Science

### Classification techniques

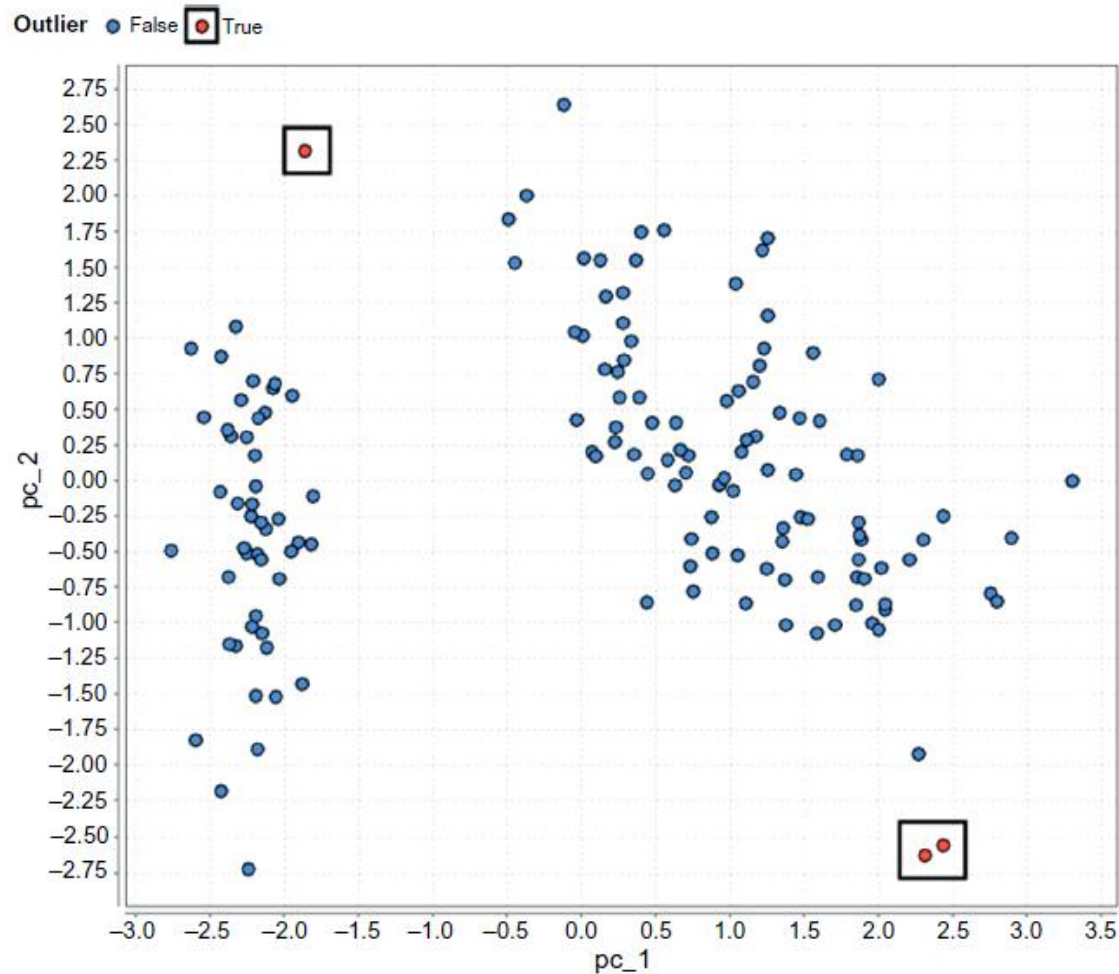
- Nearly all classification techniques can be used to identify outliers, if previously known classified data are available.
  - In classification techniques for detecting outliers, a known test dataset is needed where one of the class labels should be called “Outlier.”
  - The outlier-detection classification model that is built based on the test dataset can predict whether the unknown data is an outlier or not.
  - The challenge in using a classification model is the availability of previously labeled data.
  - Outlier data may be difficult to source because they are rare. This can be partially solved by stratified sampling where the outlier records are oversampled against normal records.
- 

# Distance-Based Outlier Detection

- Distance or proximity-based outlier detection is one of the most fundamental algorithms for anomaly detection and it relies on the fact that outliers are distant from other data points.
- The proximity measures can be simple Euclidean distance for real values and cosine or Jaccard similarity measures for binary and categorical values

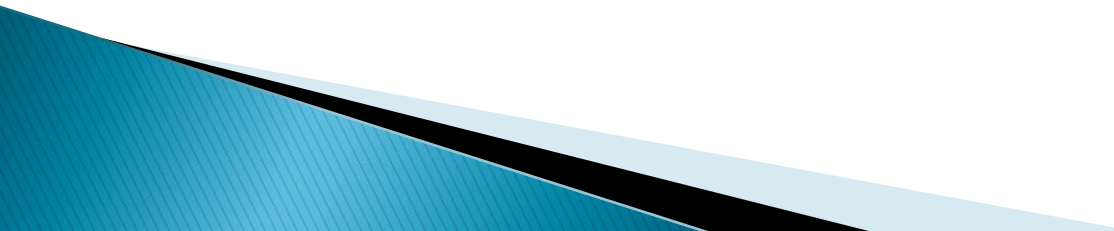


# Distance-Based Outlier Detection



Dataset with outliers.

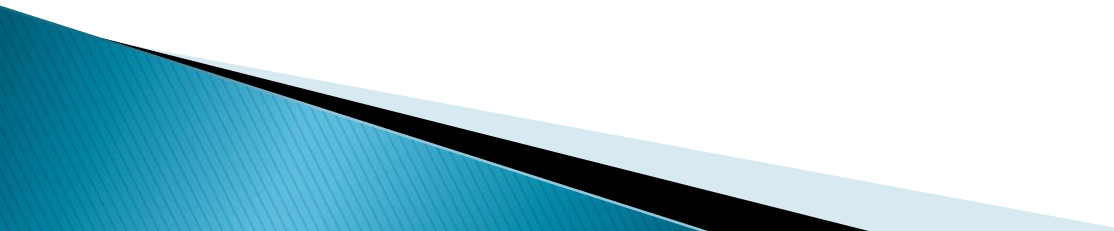
# Distance-Based Outlier Detection Working

- The fundamental concept of distance-based outlier detection is assigning a distance score for all the data points in the dataset.
  - The distance score should reflect how far a data point is separated from other data points.
  - A distance score can be assigned for each data object that is the distance to the  $k^{\text{th}}$ -nearest data object.
  - For example, a distance score can be assigned for every data object that is the distance to the third-nearest data object.
- 

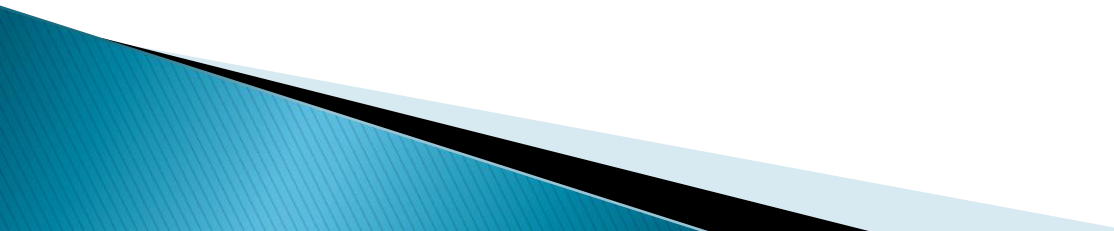
# Distance-Based Outlier Detection Working

- If the data object is an outlier, then it is far away from other data objects; hence, the distance score for the outlier will be higher than for a normal data object.
- If the data objects are sorted by distance score, then the objects with the highest scores are potentially outlier(s).
- In distance-based outlier detection, there is a significant effect based on the value of  $k$ , as in the  $k$ -NN classification technique.
- If the value of  $k=1$ , then two outliers next to each other but far away from other data points are not identified as outliers.

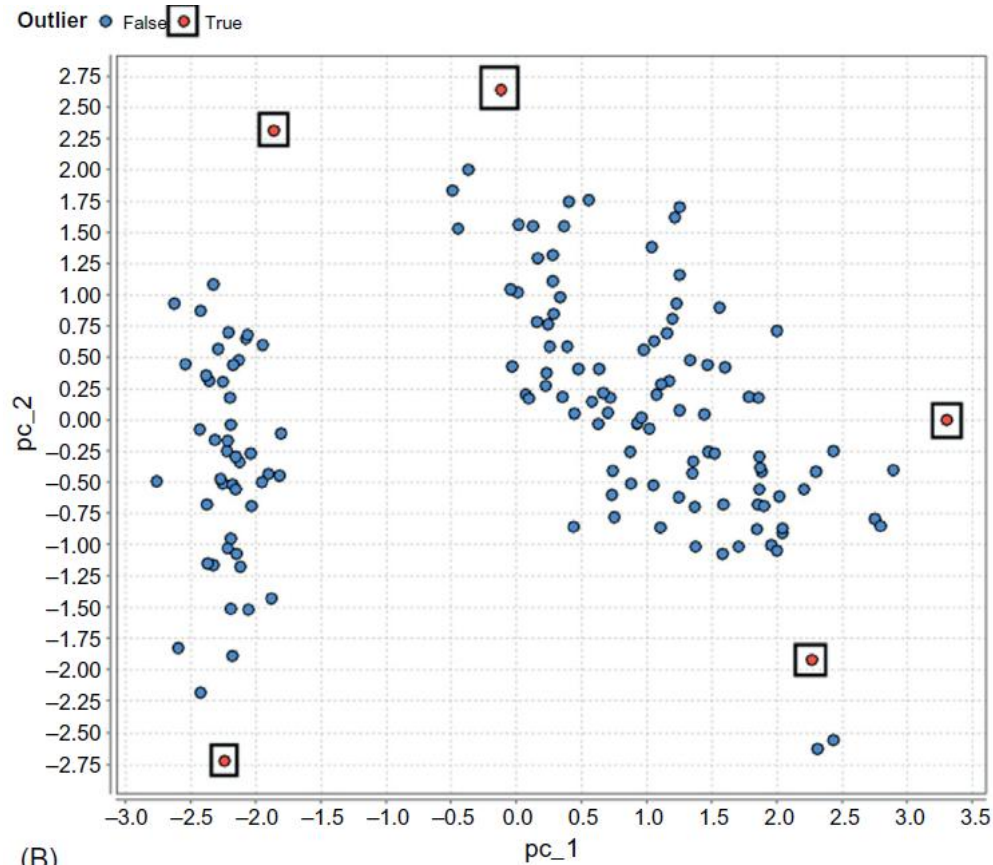
# Distance-Based Outlier Detection Working

- On the other hand, if the value of  $k$  is large, then a group of normal data points which form a cohesive cluster will be mislabeled as outliers, if the number of data points is less than  $k$  and the cluster is far away from other data points.
  - With a defined value of  $k$ , once the distance scores have been calculated, a distance threshold can be specified to identify outliers or pick the top  $n$  objects with maximum distances, depending on the application and the nature of the dataset.
- 

# Distance-Based Outlier Detection Implementation

- Commercial data science tools offer specific outlier-detection algorithms and solutions as part of the package either in the modeling or data cleansing sections.
  - In RapidMiner, unsupervised outlier-detection operator can be found in Data Transformation > Data Cleansing > Outlier Detection > Detect > Outlier Distance.
  - The example set used in this process is the Iris dataset with four numerical attributes and 150 examples.
- 

# Distance-Based Outlier Detection Implementation

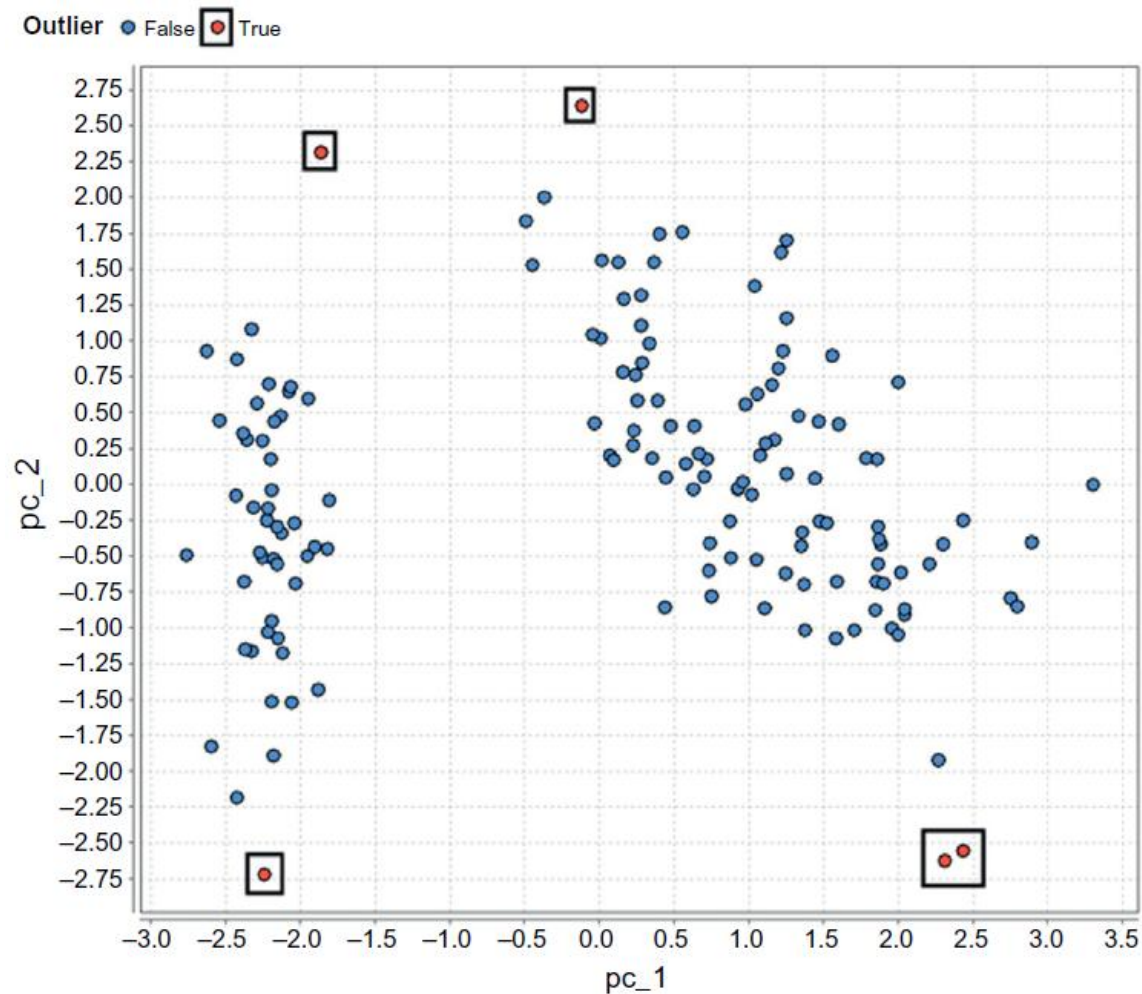


(B)

Top five outliers of Iris dataset k=1



# Distance-Based Outlier Detection Implementation



Top five outliers of Iris dataset k=5

# Distance-Based Outlier Detection Implementation

- **Step 1: Data Preparation**

- ▶ Even though all four attributes of the Iris dataset measure the same quantity (length) and are measured on the same scale (centimeters), a normalization step is included as a matter of best practice for techniques that involve distance calculation.
- ▶ The Normalize operator can be found in Data Transformation > Value modification > Numerical. The attributes are converted to a uniform scale of mean 0 and standard deviation 1 using Ztransformation.



# Distance-Based Outlier Detection Implementation

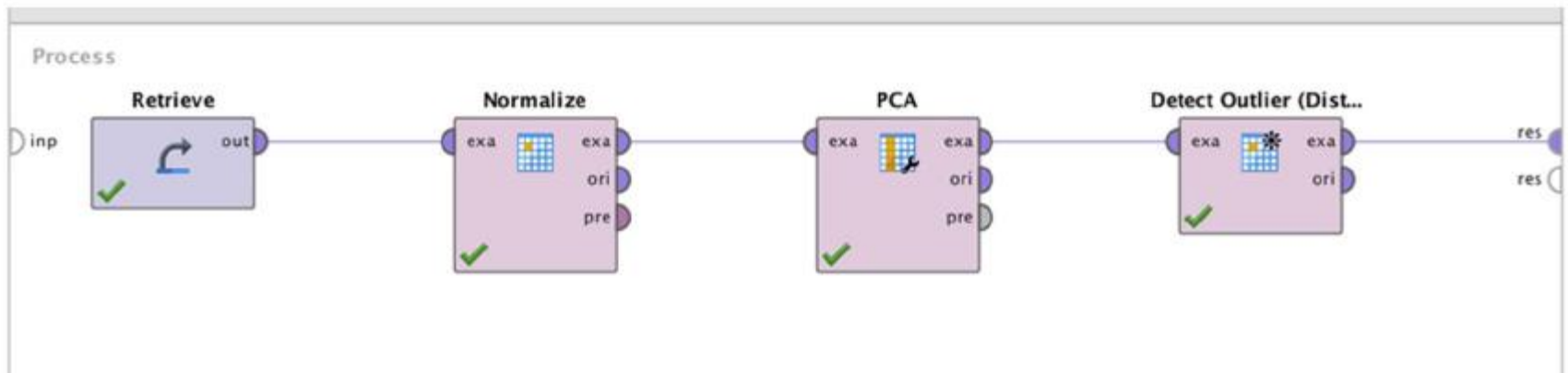
## ▶ Step 2: Detect Outlier Operator

- ▶ The Detect Outlier (Distances) operator has a data input port and outputs data with an appended attribute called outlier. The value of the output outlier attribute is either true or false. The Detect Outlier (Distances) operator has three parameters that can be configured by the user.
  - Number of neighbors
  - Number of outliers
  - Distance function

# Distance-Based Outlier Detection Implementation

## ► Step 2: Detect Outlier Operator

- In this example,  $k=1$ , number of outliers=10, and the distance function is set to Euclidian. The output of this operator is the example set with an appended outlier attribute.



Process to detect outliers based on distance

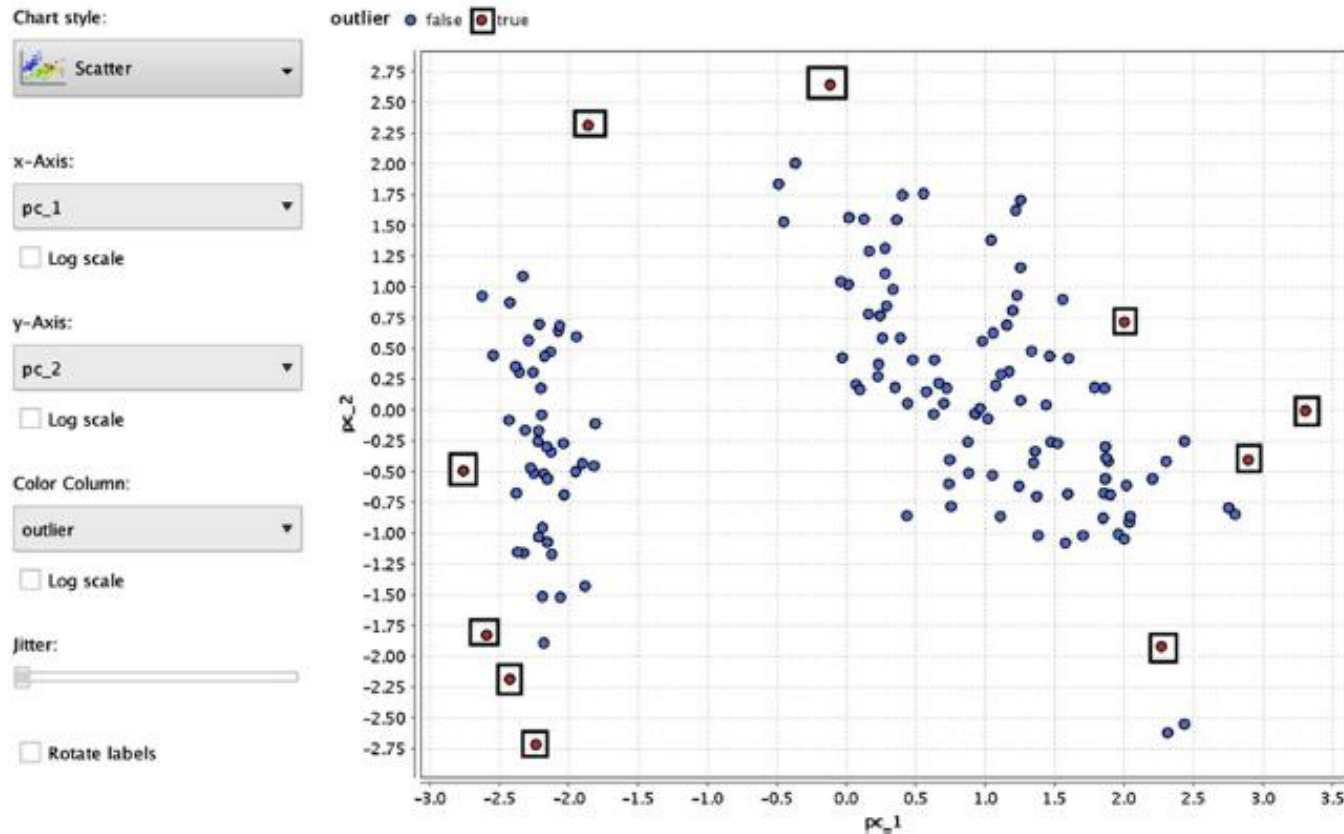
# Distance-Based Outlier Detection Implementation

- **Step 3: Execution and Interpretation**

- The result dataset can be sorted by outlier attribute, which has either a true or false value. Since 10 outliers have been specified in the parameter of the Detect outlier operator, that number of outliers can be found in the result set.

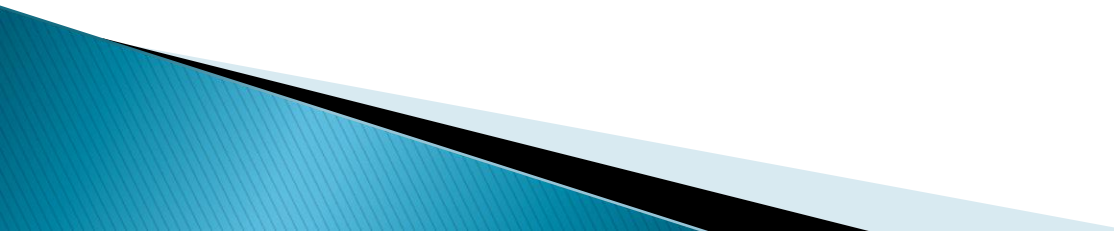
# Distance-Based Outlier Detection Implementation

## ► Step 3: Execution and Interpretation

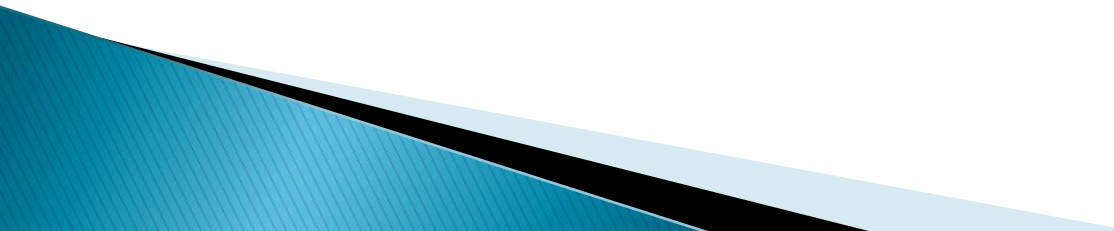


Outlier detection output

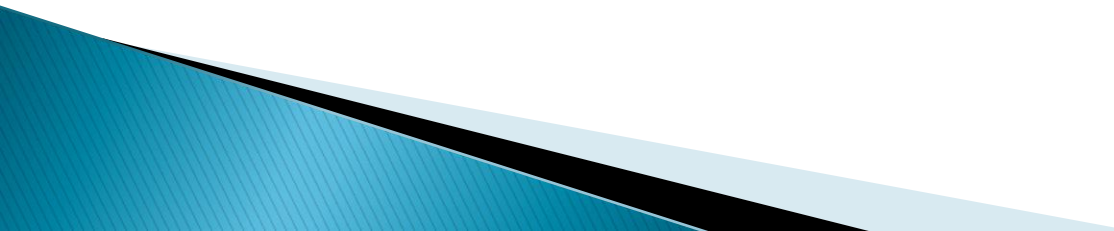
# Distance-Based Outlier Detection

- Distance-based outlier detection is a simple algorithm that is easy to implement and widely used when the problem involves many numeric variables.
  - The execution becomes expensive when the dataset involves a high number of attributes and records, because the algorithm has to calculate distances with other data points in high-dimensional space.
- 

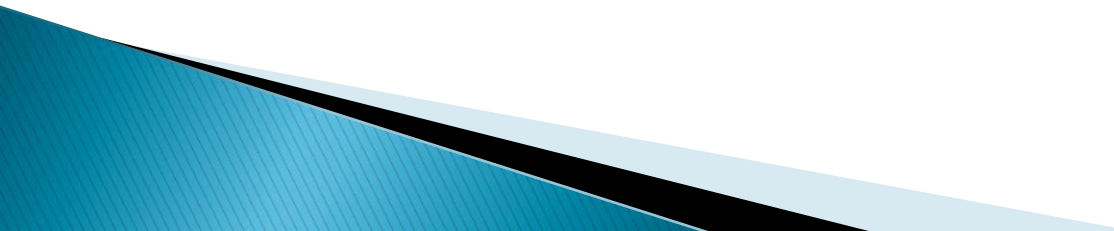
# Density-Based Outlier Detection

- Outliers, by definition, occur less frequently compared to normal data points. This means that in the data space outliers occupy low-density areas and normal data points occupy high-density areas.
  - Density is a count of data points in a normalized unit of space and is inversely proportional to the distances between data points. The objective of a density-based outlier algorithm is to identify those data points from low-density areas.
  - There are a few different implementations to assign an outlier score for the data points.
- 

# Density-Based Outlier Detection

- The inverse of the average distance of all  $k$  neighbors can be found.
  - The distance between data points and density are inversely proportional.
  - Neighborhood density can also be calculated by calculating the number of data points from a normalized unit distance.
  - The approach for density-based outliers is similar to the approach discussed for density-based clustering and for the  $k$ -NN classification algorithm.
- 

# Density-Based Outlier Detection Working


- Since distance is the inverse of density, the approach of a density-based outlier can be explained with two parameters, distance ( $d$ ) and proportion of data points ( $p$ ).
  - A point  $X$  is considered an outlier if at least  $p$  fraction of points lie more than  $d$  distance from the point. By the given definition, the point  $X$  occupies a low-density area. The parameter  $p$  is specified as a high value, above 95%.
  - One of the key issues in this implementation is specifying distance.
  - It is important to normalize the attributes so that the distance makes sense, particularly when attributes involve different measures and units.
  - If the distance is specified too low, then more outliers will be detected, which means normal points have the risk of being labeled as outliers and vice versa
- 



# Density-Based Outlier Detection : Implementation Steps

- The RapidMiner process for outlier detection based on density is similar to outlier detection by distance, which was reviewed in the previous section.
- The process developed for previous distance-based outliers can be used, but the Detect Outlier (Distances) operator would be replaced with the Detect Outlier (Densities) operator.

## Step 1: Data Preparation

- Data preparation will condition the data so the Detect Outlier (Densities) operator returns meaningful results. As with the outlier detection by distance technique, the Iris dataset will be used with normalization and the PCA operator so that the number of attributes is reduced to two for easy visualization
- 

# Density-Based Outlier Detection : Implementation Steps

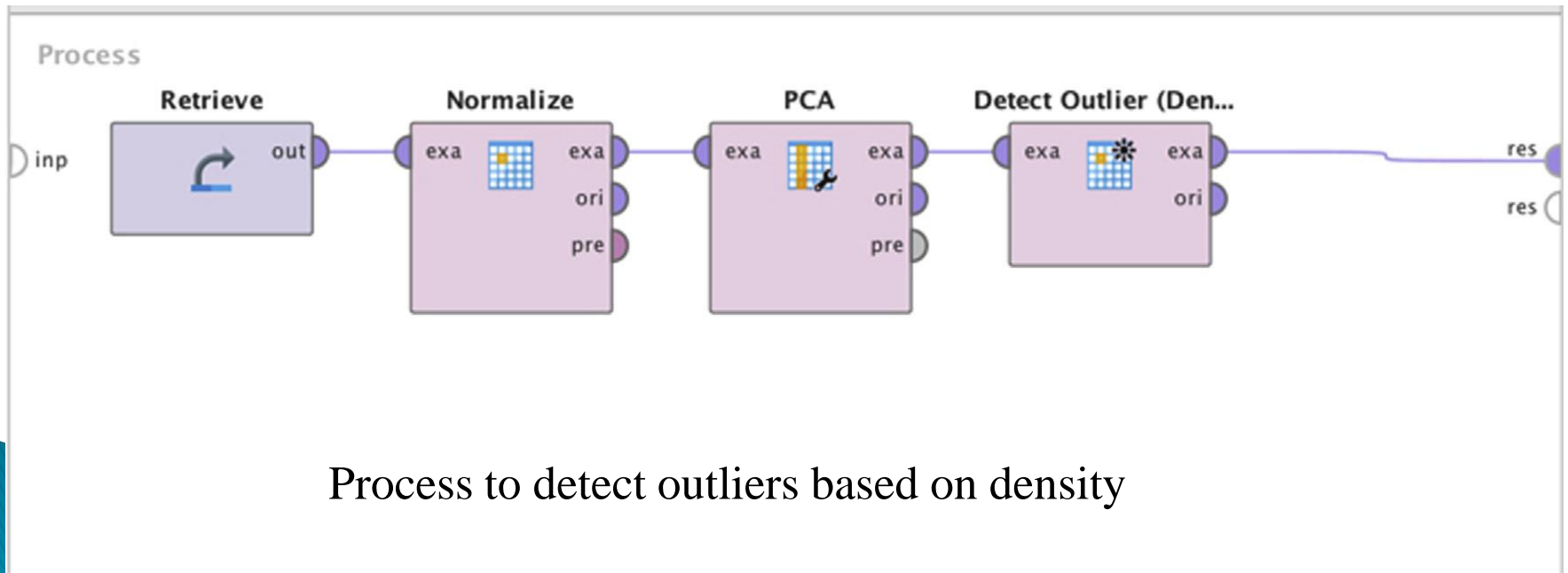
## Step 2: Detect Outlier Operator

- The Detect Outlier (Densities) operator can be found in Data Transformation> Data Cleansing> Outlier Detection, and has three parameters:
  - ✓ Distance (d): Threshold distance used to find outliers. For this example, the distance is specified as 1.
  - ✓ Proportion (p): Proportion of data points outside of radius d of a point, beyond which the point is considered an outlier. For this example, the value specified is 95%.
  - ✓ Distance measurement: A measurement parameter like Euclidean, cosine, or squared distance. The default value is Euclidean.

# Density-Based Outlier Detection : Implementation Steps


## Step 2: Detect Outlier Operator

- Any data point that has more than 95% of other data points beyond distance  $d$  is considered an outlier. The RapidMiner process with the Normalization, PCA, and Detect Outlier operators is shown in the figure. The process can be saved and executed



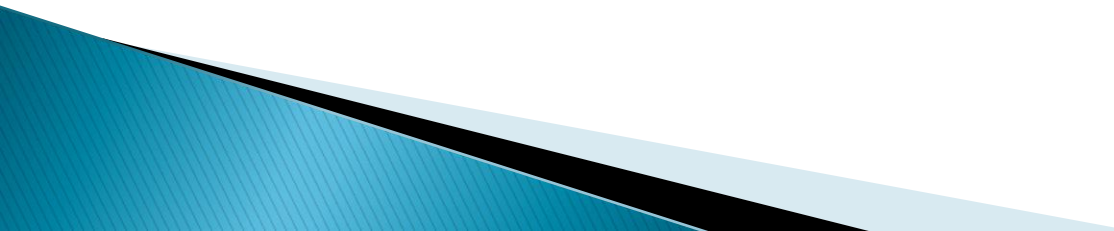
# Density-Based Outlier Detection : Implementation Steps

## Step 3: Execution and Interpretation

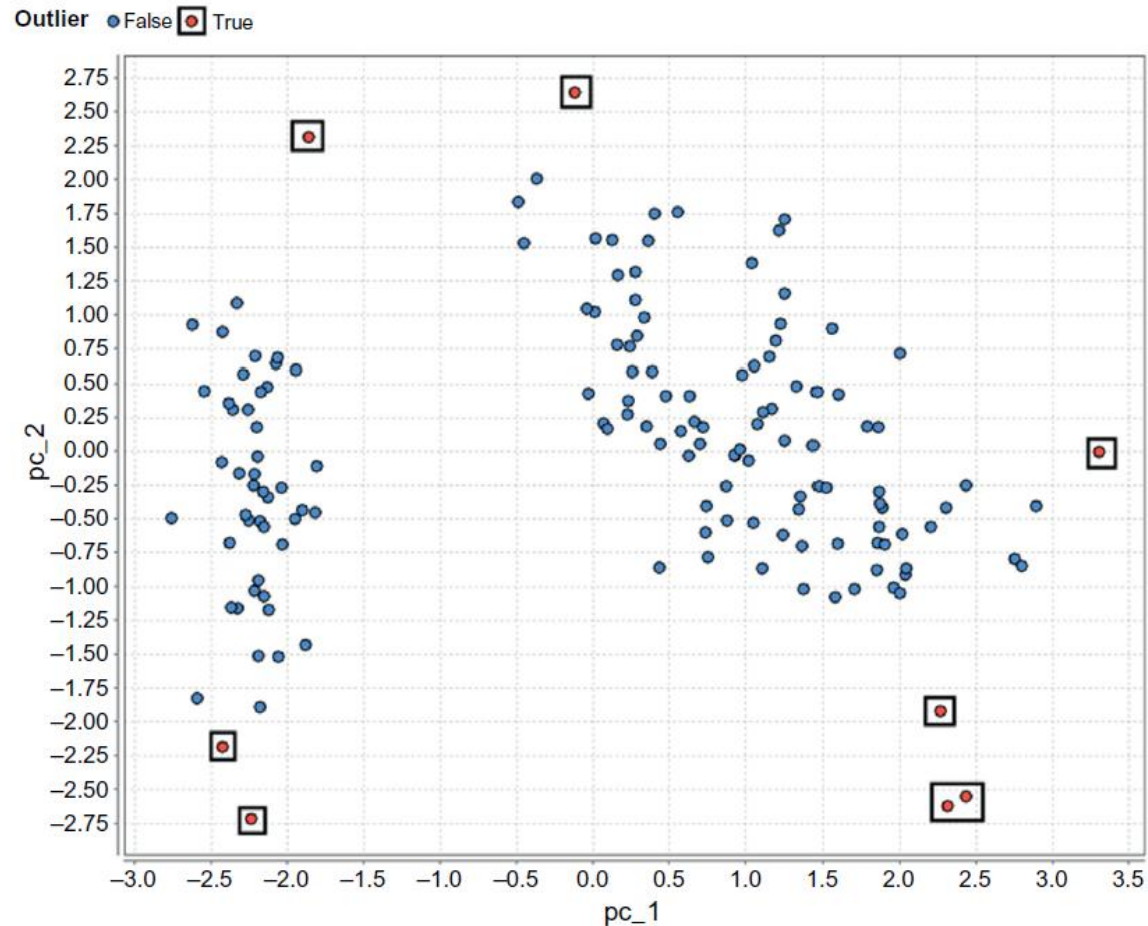
- The process adds an outlier attribute to the example set, which can be used for visualization using a scatterplot.
  - The outlier attribute is Boolean and indicates whether the data point is predicted to be an outlier or not. In the scatterplot, a few data points marked as outliers can be found.
  - The parameters  $d$  and  $p$  of the Detect Outlier operator can be tuned to find the desired level of outlier detection.
  - Density-based outlier detection is closely related to distance-based outlier approaches and, hence, the same pros and cons apply.
- 

# Density-Based Outlier Detection : Implementation Steps

## Step 3: Execution and Interpretation

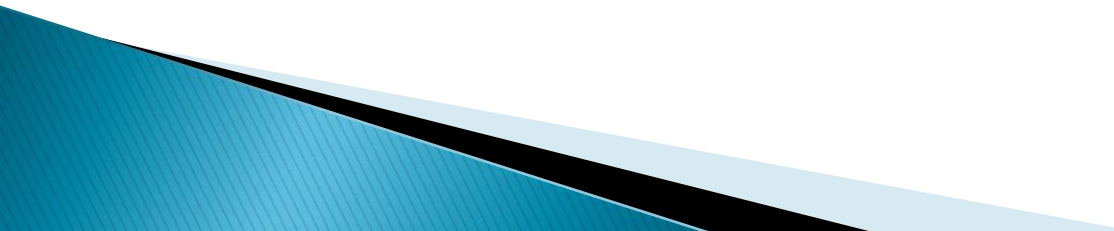
- As with distance based outlier detection, the main drawback is that this approach does not work with varying densities.
  - The next approach, local outlier factor (LOF) is designed for such datasets.
  - Specifying the parameter distance ( $d$ ) and proportion ( $p$ ) is going to be challenging, particularly when the characteristics of the data are not previously known.
- 

# Density-Based Outlier Detection : Implementation Steps



Output of density-based outlier detection.

# Local Outlier Factor

- The LOF technique is a variation of density-based outlier detection, and addresses one of its key limitations, detecting the outliers in varying density.
  - Varying density is a problem in simple density-based methods, including DBSCAN clustering
  - LOF takes into account the density of the data point and the density of the neighborhood of the data point as well.
  - A key feature of the LOF technique is that the outlier score takes into account the relative density of the data point.
- 

# Local Outlier Factor : Working

- Once the outlier scores for data points are calculated, the data points can be sorted to find the outliers in the dataset.
- The core of the LOF lies in the calculation of the relative density.
- The relative density of a data point  $X$  with  $k$  neighbors is given by the following equation:

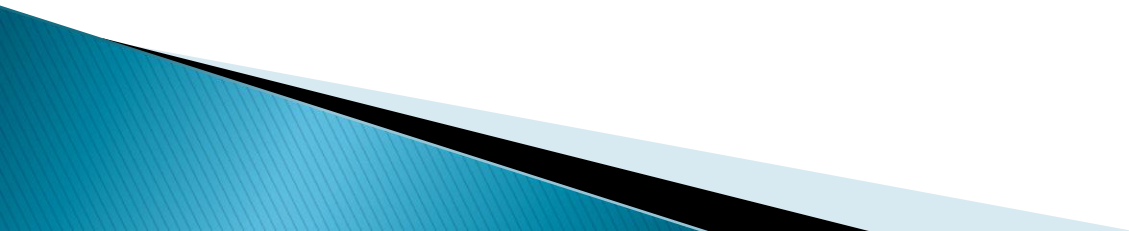
$$\text{Relative density of } X = \frac{\text{Density of } X}{\text{Average density of all data points in the neighborhood}}$$

- where the density of  $X$  is the inverse of the average distance for the nearest  $k$  data points

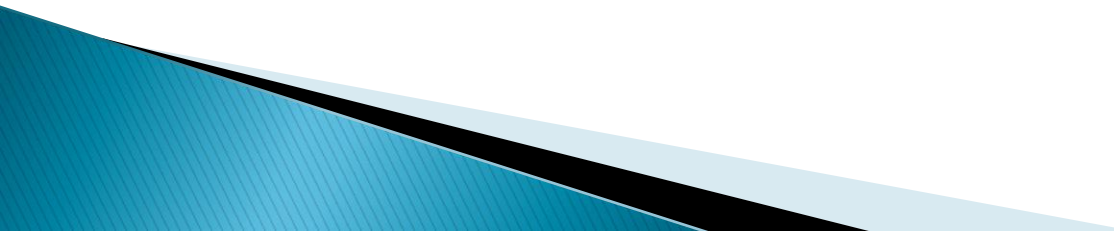


# Local Outlier Factor : Working

- The same parameter  $k$  also forms the locality of the neighborhood
- By comparing the density of the data point and density of all the data points in the neighborhood, whether the density of the data point is lower than the density of the neighborhood can be determined. This scenario indicates the presence of an outlier.



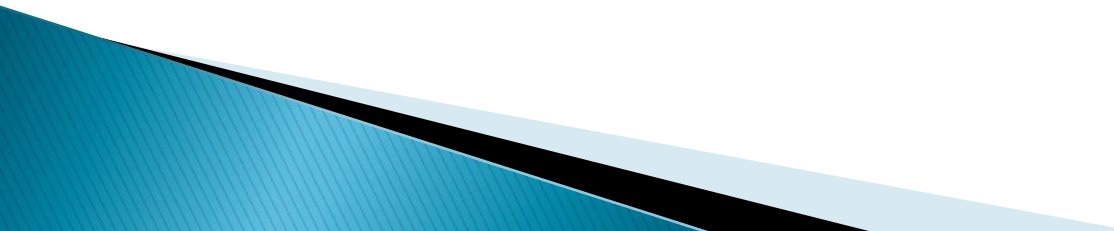
# Local Outlier Factor : Implementation

- A LOF-based data science process is similar to the other outlier processes explained in RapidMiner. The Detect Outlier (LOF) operator is available in Data Transformation > Data Cleansing > Outlier Detection.
  - The output of the LOF operator contains the example set along with a numeric outlier score.
  - The LOF algorithm does not explicitly label a data point as an outlier; instead the score is exposed to the user. This score can be used to visualize a comparison to a threshold, above which the data point is considered an outlier.
- 

# Local Outlier Factor : Implementation

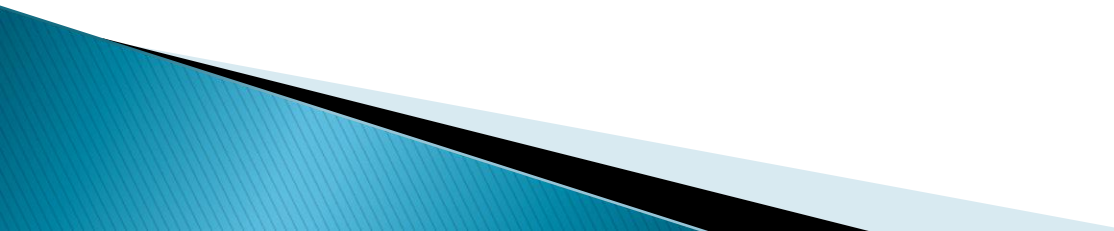
- Having the raw score means that the data science practitioner can “tune” the detection criteria, without having to rerun the scoring process, by changing the threshold for comparison.

## Step 1: Data Preparation

- Similar to the distance- and density-based outlier-detection processes, the dataset has to be normalized using Normalize operator. The PCA operator is used to reduce the four-dimensional Iris dataset to two dimensions, so that the output can be visualized easily.
- 

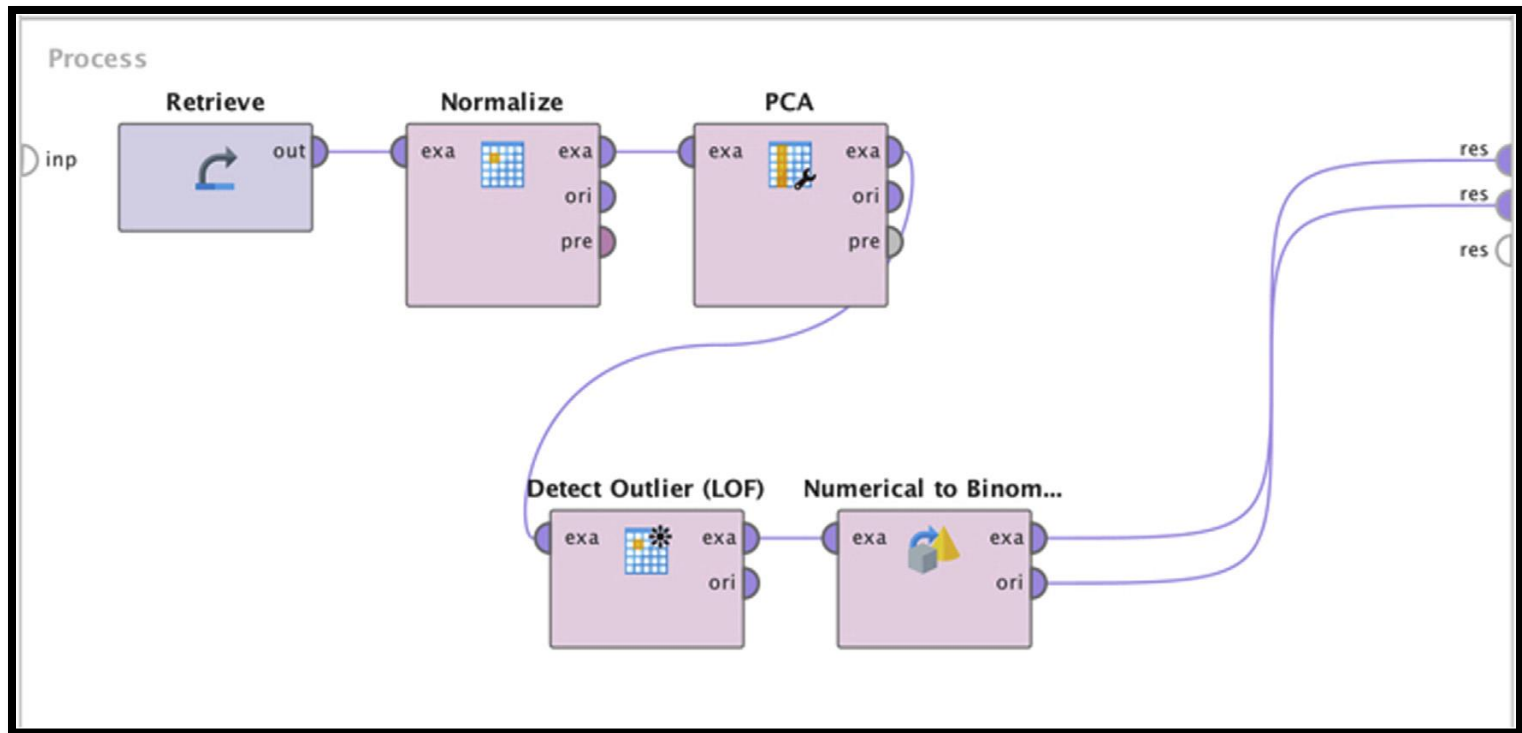
# Local Outlier Factor : Implementation

## Step 2: Detect Outlier Operator

- The LOF operator has minimal points (MinPts) lower bound and upper bound as parameters.
  - The MinPts lower bound is the value of  $k$ , the neighborhood number.
  - The LOF algorithm also takes into account a MinPts upper bound to provide more stable results
- 

# Local Outlier Factor : Implementation


## Step 2: Detect Outlier Operator



RapidMiner process for LOF outlier detection

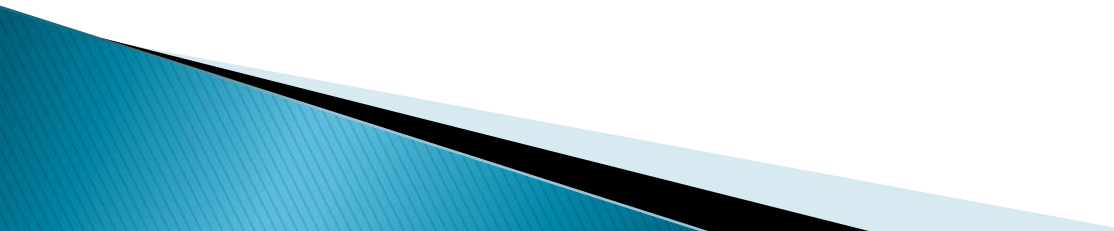
# Local Outlier Factor : Implementation

## Step 3: Results Interpretation

- After using the Detect Outlier operator, the outlier score is appended to the result dataset.
  - The result set with outlier score is represented as the color of the data point.
  - In the results window, the outlier score can be used to color the data points.
  - The scatterplot indicates that points closer to the blue spectrum (left side of the outlier scale in the chart legend) are predicted to be regular data points and points closer to the red spectrum are predicted to be outliers.
- 

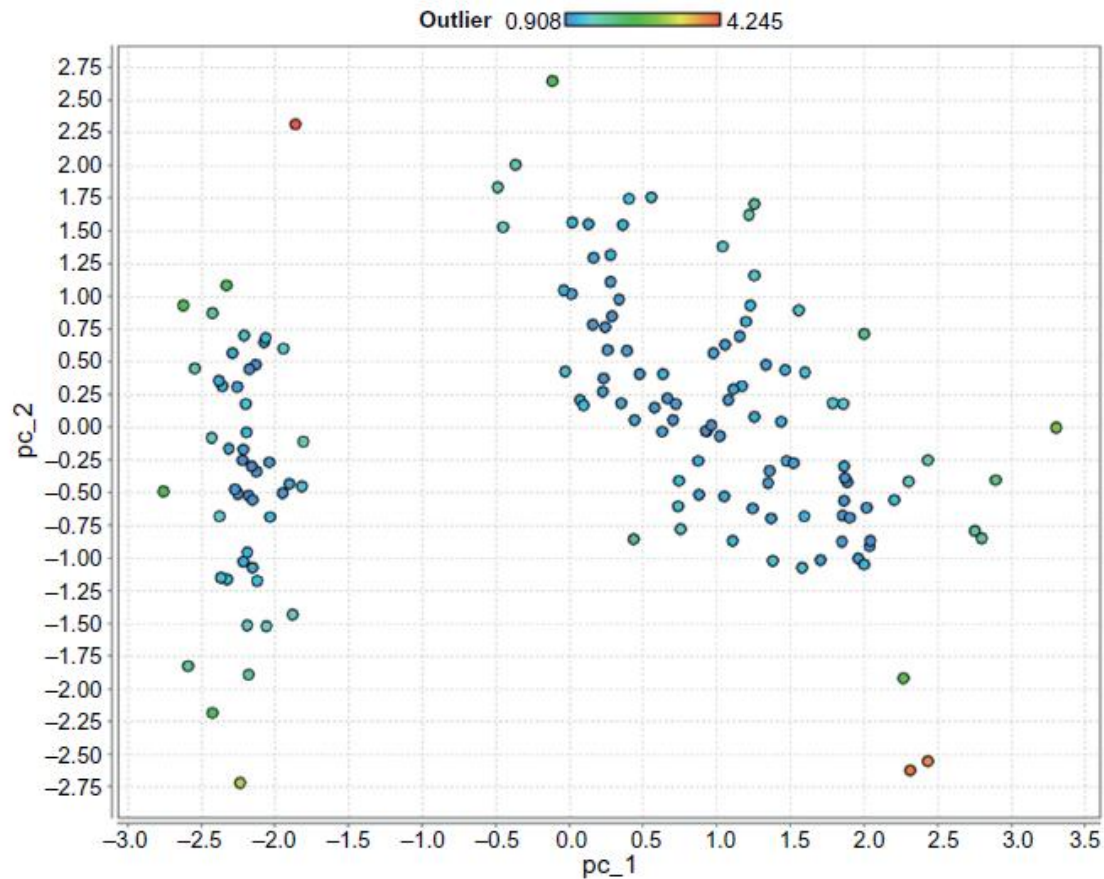
# Local Outlier Factor : Implementation

## Step 3: Results Interpretation

- If an additional Boolean flag indicating whether a data point is an outlier or not is needed, a Numeric to Binominal operator can be added to the result dataset.
  - The Numeric to Binominal operator converts the numeric outlier score to a binominal true or false based on the threshold specification in the parameter of the operator and to the score output from the LOF operator.
- 

# Local Outlier Factor : Implementation

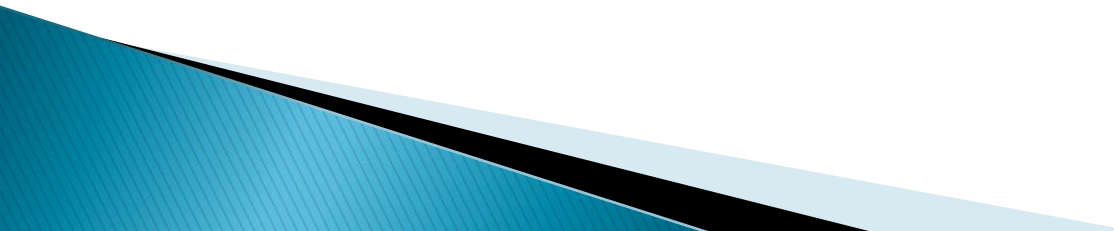
## Step 3: Results Interpretation



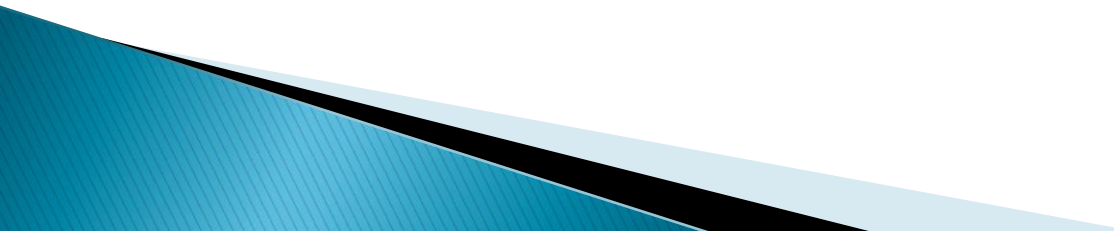
Output of LOF outlier detection



# Conclusion

- In theory, any classification algorithm can be used for outlier detection, if a previously classified dataset is available.
  - A generalized classification model tries to predict outliers the same way it predicts the class label of the data point.
  - However, there is one key issue in using classification models.
  - Since the probability of occurrence of an outlier is really low, say, less than 0.1%, the model can just “predict” the class as “regular” for all the data points and still be 99.9% accurate! This method clearly does not work for outlier detection, since the recall measure is 0%.
  - In practical applications, like detecting network intrusion or fraud prevention in high-volume transaction networks, the cost of not detecting an outlier is very high.
- 

# Conclusion

- The model can even have an acceptable level of false alarms, that is, labeling a regular data point as an outlier.
  - Therefore, special care and preparation is required to improve the detection of the outliers.
  - Stratified sampling methods can be used to increase the frequency of occurrence of outlier records in the training set and reduce the relative occurrence of regular data points.
  - In a similar approach, the occurrence of outliers and regular records can be sampled with replacements so that there are an equal number of records in both classes.
  - Stratified sampling boosts the number of outlier records in the test dataset with respect to regular records in an attempt to increase both the accuracy and recall of outlier detection.
- 

# Conclusion

- In practical applications, outlier-detection models have to be updated frequently as the characteristics of an outlier changes over time, and hence, the relationship between outliers and normal records changes as well.
  - In constant real time data streams, outlier detection creates additional challenges because of the dynamic distribution of the data and dynamic relationships within the data.
  - Outlier detection remains one of the most profound applications of data science as it impacts the majority of the population through financial transaction monitoring, fraud prevention, and early identification of anomalous activity in the context of security.
- 