Tutorial 5

1.

Noisy data is data that contains errors or inconsistencies. It can be caused by a variety of factors, such as human error, faulty equipment, or environmental conditions. Noisy data can make it difficult to draw accurate conclusions from data analysis.

Some common sources of noise in data include:

1. Measurement Errors	These errors can occur when data is collected or measured. For example, a sensor may not be calibrated correctly, or a human may make a mistake when entering data into a computer.
2. Outliers	Outliers are data points that are significantly different from the rest of the data. They can be caused by a variety of factors, such as human error, faulty equipment, or natural variation.
3. Missing Values	Missing values occur when data is not collected for a particular observation. This can happen for a variety of reasons, such as a sensor malfunction or a human error.

Some common methods for identifying noisy data include:

1. Visual Inspection	This involves looking at the data visually to identify any obvious errors or inconsistencies.
2. Statistical Analysis	This involves using statistical methods to identify data points that are significantly different from the rest of the data.

Some common methods for handling noisy data include:

1. Data Cleaning	This involves removing or correcting errors in the data.
2. Data Imputation	This involves filling in missing values with estimates.
3. Data Transformation	This involves changing the format of the data to make it easier to analyze.

2.

Noisy data is data that contains errors or inconsistencies. It can be caused by a variety of factors, such as human error, faulty equipment, or environmental conditions. Noisy data can make it difficult to draw accurate conclusions from data analysis.

There are many different techniques for handling noisy data. Some of the most common techniques include:

Nasir Uddin Ahmed
Student Id: S2015449

1. Smoothing	Smoothing is a technique that is used to remove random fluctuations from data. This can be done by averaging the values of nearby data points or by fitting a curve to the data.
2. Clustering	Clustering is a technique that is used to group data points that are similar to each other. This can be done by finding data points that are close together in space or by finding data points that have similar values.
3.Outlier Detection	Outlier detection is a technique that is used to identify data points that are significantly different from the rest of the data. This can be done by using statistical methods or by visually inspecting the data.

The type of data being handled and the sort of analysis being done will determine the optimal method for handling noisy data. If the data is continuous, for instance, smoothing might be a smart choice. Clustering could be a useful choice if the data is categorical. Outlier detection could be a wise choice if the data contains outliers.

Here are some best practices for selecting the appropriate technique for handling noisy data:

a. Consider the type of data. Some techniques are better suited for certain types of data than others.

b. Consider the analysis that is being performed. Some techniques are better suited for certain types of analysis than others.

c. Consider the amount of noise in the data. Some techniques are more effective at handling noise than others.

d. Experiment with different techniques. The best technique may not be obvious, so it is important to experiment with different techniques to find the one that works best for the data.

3.

The process of merging data from various sources into a single, cohesive perspective is known as data integration. This is crucial for data analysis since it enables you to combine data from several sources to create a more accurate and comprehensive picture of your data.

There are many different challenges that can arise during the data integration process. Some of the most common challenges include:

1. Data Format Incompatibility	Data from different sources may be stored in different formats, which can make it difficult to combine the data.
2. Missing or Inconsistent Data	Some data may be missing from one or more sources, or the data may be inconsistent between sources. This can make it difficult to analyze the data.
3.Data Duplication	The same data may be stored in multiple sources, which can lead to confusion and errors.

There are a number of techniques and tools that can be used to address these challenges. Some of the most common techniques include:

1. Data Cleansing	Data cleansing is the process of identifying and correcting errors in data. This can help to ensure that the data is accurate and consistent.
2.Data Transformation	Data transformation is the process of converting data from one format to another. This can help to make the data compatible with different systems and applications.
3.Data Deduplication	Data deduplication is the process of identifying and removing duplicate data. This can help to reduce the amount of data that needs to be stored and processed.

4.

The comprehensive data mining and predictive analytics software SAS Enterprise Miner has many features and options for integrating data. For data integration, SAS Enterprise Miner has a number of critical features and capabilities, including:

1. Data Cleansing and Preparation	SAS Enterprise Miner includes a variety of tools for cleansing and preparing data, including data validation, data imputation, and data normalization.
2. Data Integration	SAS Enterprise Miner can integrate data from a variety of sources, including relational databases, flat files, and web services.
3. Data Mining	SAS Enterprise Miner includes a variety of data mining algorithms for predictive modeling, including classification, regression, and clustering.
4.Predictive Analytics	SAS Enterprise Miner can be used to build predictive models that can be used to make predictions about future events.
5. Reporting and Visualization	SAS Enterprise Miner includes a variety of reporting and visualization tools that can be used to communicate the results of data mining and predictive analytics projects.

SAS Enterprise Miner is a data mining and predictive analytics tool offered by SAS Institute. While it has some data integration capabilities, it primarily focuses on advanced analytics and modeling. Here are a few ways SAS Enterprise Miner differs from other data integration tools and platforms.

1.Advanced Analytics	SAS Enterprise Miner is designed specifically for advanced analytics tasks, such as data mining, predictive modeling, and statistical analysis. It provides a wide range of algorithms and techniques to uncover patterns, relationships, and insights within data.
2. Focus on Predictive Modeling	SAS Enterprise Miner emphasizes predictive modeling and the development of predictive models. It offers a visual interface for building and evaluating models, allowing users to explore various algorithms and techniques to create accurate predictions.

3. Integration with SAS Ecosystem	SAS Enterprise Miner seamlessly integrates with other SAS products, such as SAS Visual Analytics and SAS Data Integration Studio. This integration enables users to leverage the full power of the SAS ecosystem for data preparation, analysis, and reporting.

The robust and flexible data integration tool SAS Enterprise Miner can be used to combine data from many sources, clean and prepare data, create predictive models, and present the findings of data mining and predictive analytics projects. Advantages and Disadvantages of using SAS Enterprise Miner for data integration given below.

Advantages	Disadvantages
1. It is a comprehensive and powerful data integration tool.	1. It is a complex and expensive tool.
2. It can integrate data from a variety of sources.	2. It requires a significant investment in training and expertise.
3. It includes a variety of reporting and visualization tools.	3. It can be difficult to integrate with other data integration tools and platforms.

SAS Enterprise Miner can be used to create a unified data set that can be used for analysis by following these steps:

1. Identify the data sources that need to be integrated.

2. Cleanse and prepare the data from each data source.

3. Integrate the data from the different data sources into a single data set.

4. Perform data mining and predictive analytics on the unified data set.

5. Communicate the results of the data mining and predictive analytics projects.