**Big Data Analysis with IBM Cloud Databases**

**Team Member**

312821243028: Raghul V

Phase-1 Document Submission

**Project Title:** Big Data Analysis

**Problem Statement:**

Dive into the world of big data analysis with IBM Cloud Databases. Uncover hidden insights from vast datasets, from climate trends to social patterns. Visualize your findings and derive valuable business intelligence. Embark on data-driven adventures, exploring the endless possibilities of big data.

**Introduction:**

Data mining is a methodology in computer science for discovering meaningful patterns and knowledge from large amounts of data. However, before a data mining model can be applied, the raw data must be preprocessed to ensure that it is in a suitable format for analysis. Data preprocessing is an essential step in the data mining process and can greatly impact the accuracy and efficiency of the final results. the most common data preprocessing techniques, including data cleaning, data integration, data transformation, and feature selection. With practical examples and code snippets, this article will help you understand the key concepts and techniques involved in data preprocessing and equip you with the skills to apply them to your own data mining projects. Whether you are a beginner or an experienced data miner, this guide will be a valuable resource to help you achieve high-quality results from your data.

## Data Preprocessing:

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following:

* **Accuracy**: To check whether the data entered is correct or not.
* **Completeness**: To check whether the data is available or not recorded.
* **Consistency:** To check whether the same data is kept in all the places that do or do not match.
* **Timeliness**: The data should be updated correctly.
* **Believability**: The data should be trustable.
* **Interpretability**: The understandability of the data.

**There are 4 major tasks in data preprocessing:**

* Data cleaning
* Data integration
* Data reduction
* Data transformation.



### **Data Cleaning:**

Data cleaning is the process of removing incorrect data, incomplete data, and inaccurate data from the datasets, and it also replaces the missing values.

Here are some techniques for data cleaning:

#### **Handling Missing Values:**

* Standard values like “Not Available” or “NA” can be used to replace the missing values.
* Missing values can also be filled manually, but it is not recommended when that dataset is big.
* The attribute’s mean value can be used to replace the missing value when the data is normally distributed  
  wherein in the case of non-normal distribution median value of the attribute can be used.
* While using regression or decision tree algorithms, the missing value can be replaced by the most probable value.

### **Handling Noisy Data:**

Noisy generally means random error or containing unnecessary data points. Handling noisy data is one of the most important steps as it leads to the optimization of the model we are using Here are some of the methods to handle noisy data.

* **Binning:**This method is to smooth or handle noisy data. First, the data is sorted then, and then the sorted values are separated and stored in the form of bins. There are three methods for smoothing data in the bin. **Smoothing by bin mean method**: In this method, the values in the bin are replaced by the mean value of the bin; **Smoothing by bin median**: In this method, the values in the bin are replaced by the median value; **Smoothing by bin boundary**: In this method, the using minimum and maximum values of the bin values are taken, and the closest boundary value replaces the values.
* **Regression:**This is used to smooth the data and will help to handle data when unnecessary data is present. For the analysis, purpose regression helps to decide the variable which is suitable for our analysis.
* **Clustering:**This is used for finding the outliers and also in grouping the data. Clustering is generally used in unsupervised learning.

### **Data Integration:**

The process of combining multiple sources into a single dataset. The Data integration process is one of the main components of data management. There are some problems to be considered during data integration.

* **Schema integration**: Integrates metadata(a set of data that describes other data) from different sources.
* **Entity identification problem:** Identifying entities from multiple databases. For example, the system or the user should know the student id of one database and studentname of another database belonging to the same entity.
* **Detecting and resolving data value concepts**: The data taken from different databases while merging may differ. The attribute values from one database may differ from another database. For example, the date format may differ, like “MM/DD/YYYY” or “DD/MM/YYYY”.

### **Data Reduction:**

This process helps in the reduction of the volume of the data, which makes the analysis easier yet produces the same or almost the same result. This reduction also helps to reduce storage space. Some of the data reduction techniques are dimensionality reduction, numerosity reduction, and data compression.

* **Dimensionality reduction:**This process is necessary for real-world applications as the data size is big. In this process, the reduction of random variables or attributes is done so that the dimensionality of the data set can be reduced. Combining and merging the attributes of the data without losing its original characteristics. This also helps in the reduction of storage space, and computation time is reduced. When the data is highly dimensional, a problem called the “Curse of Dimensionality” occurs.
* **Numerosity Reduction**: In this method, the representation of the data is made smaller by reducing the volume. There will not be any loss of data in this reduction.
* **Data compression**: The compressed form of data is called data compression. This compression can be lossless or lossy. When there is no loss of information during compression, it is called lossless compression. Whereas lossy compression reduces information, but it removes only the unnecessary information.

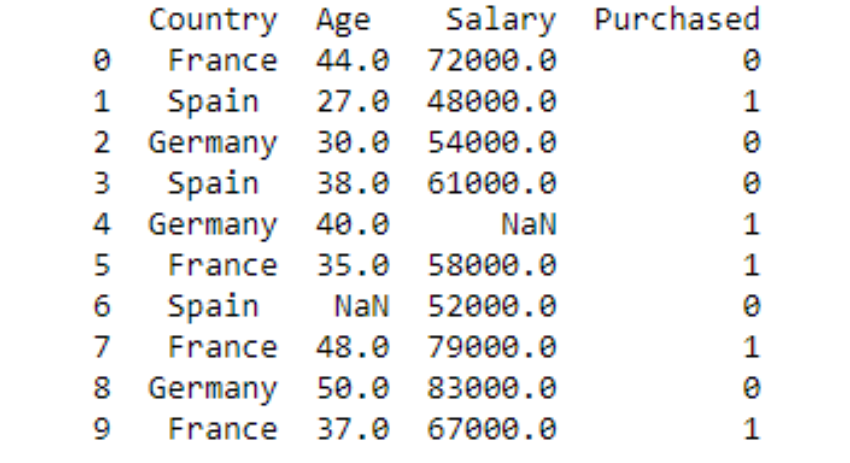
### **Data Transformation:**

The change made in the format or the structure of the data is called data transformation. This step can be simple or complex based on the requirements. There are some methods for data transformation.

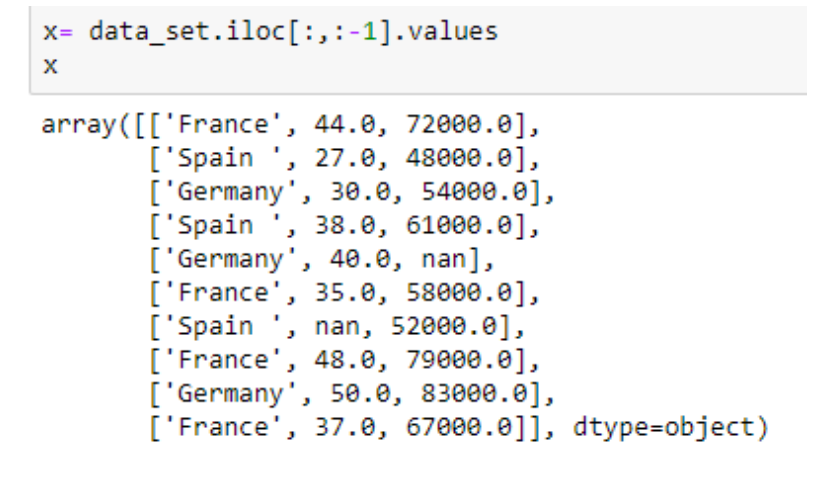
* **Smoothing:** With the help of algorithms, we can remove noise from the dataset, which helps in knowing the important features of the dataset. By smoothing, we can find even a simple change that helps in prediction.
* **Aggregation:** In this method, the data is stored and presented in the form of a summary. The data set, which is from multiple sources, is integrated into with data analysis description. This is an important step since the accuracy of the data depends on the quantity and quality of the data. When the quality and the quantity of the data are good, the results are more relevant.
* **Discretization:** The continuous data here is split into intervals. Discretization reduces the data size. For example, rather than specifying the class time, we can set an interval like (3 pm-5 pm, or 6 pm-8 pm).
* **Normalization:** It is the method of scaling the data so that it can be represented in a smaller range. Example ranging from -1.0 to 1.0.

## Data Preprocessing Steps in Machine Learning:

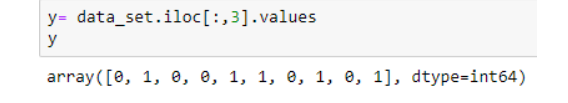
### Step 1: Importing Libraries and the Dataset



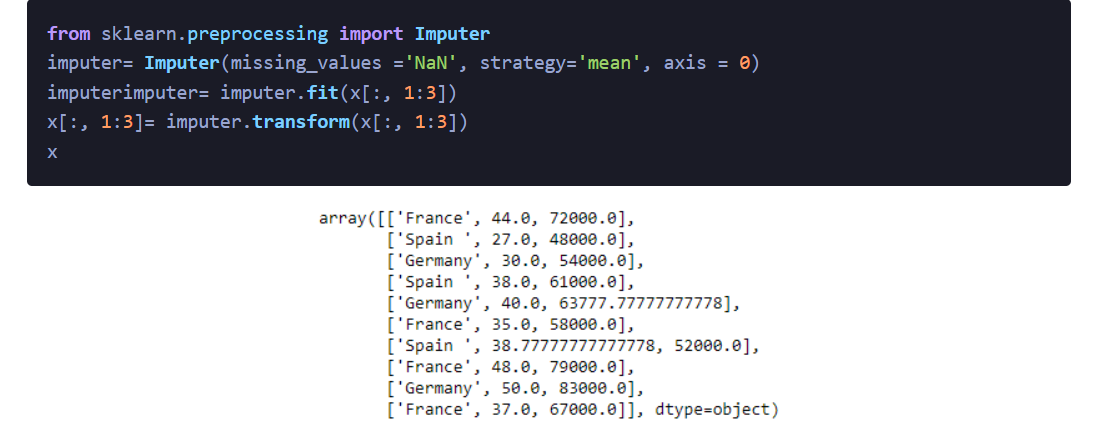
### **Step 2: Extracting the Independent Variable**



### **Step 3: Extracting the Dependent Variable**

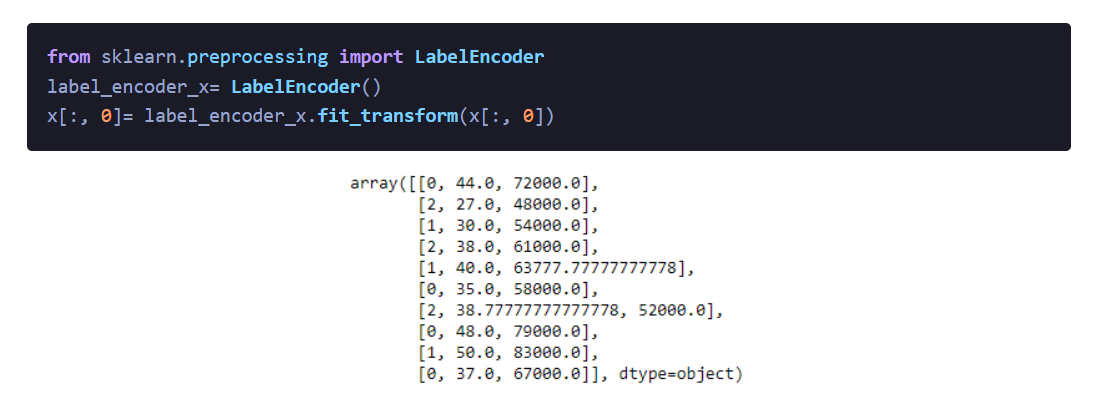


### **Step 4: Filling the Dataset with the Mean Value of the Attribute**



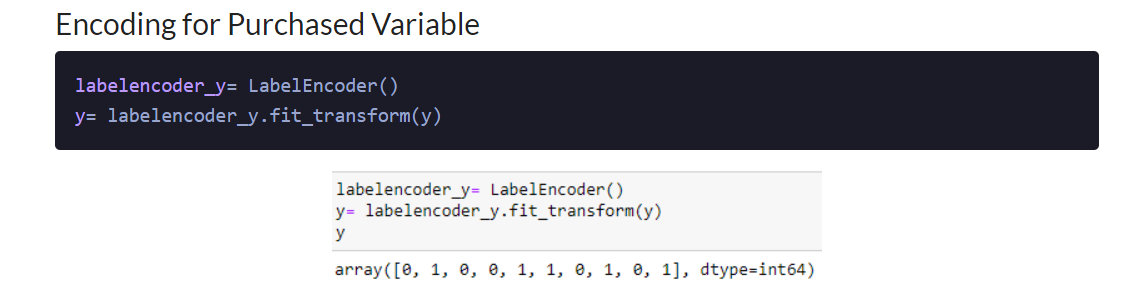
### **Step 5: Encoding the Country Variable**

The machine learning models use mathematical equations. So categorical data is not accepted, so we convert it into numerical form.

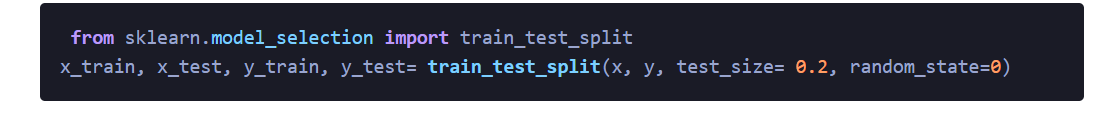


### **Step 6: Dummy Encoding**

These dummy variables replace the categorical data as 0 and 1 in the absence or the presence of the specific categorical data.



### **Step 7: Splitting the Dataset into Training and Test Sets**

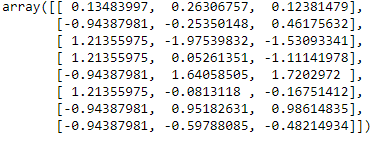


### **Step 8: Feature Scaling**

**from** sklearn.preprocessing **import** **StandardScaler**

st\_x= StandardScaler()

x\_train= st\_x.fit\_transform(x\_train)



x\_test= st\_x.transform(x\_test)

https://av-eks-blogoptimized.s3.amazonaws.com/60294994108.png

## Conclusion:

In conclusion, data preprocessing is an essential step in the data mining process and plays a crucial role in ensuring that the data is in a suitable format for analysis. This article provides a comprehensive guide to data preprocessing techniques, including data cleaning, integration, reduction, and transformation. Through practical examples and code snippets, the article helps readers understand the key concepts and techniques involved in data preprocessing and gives them the skills to apply these techniques to their own data mining projects. Whether you are a beginner or an experienced data miner, this article will provide valuable information and resources to help you achieve high-quality results from your data.