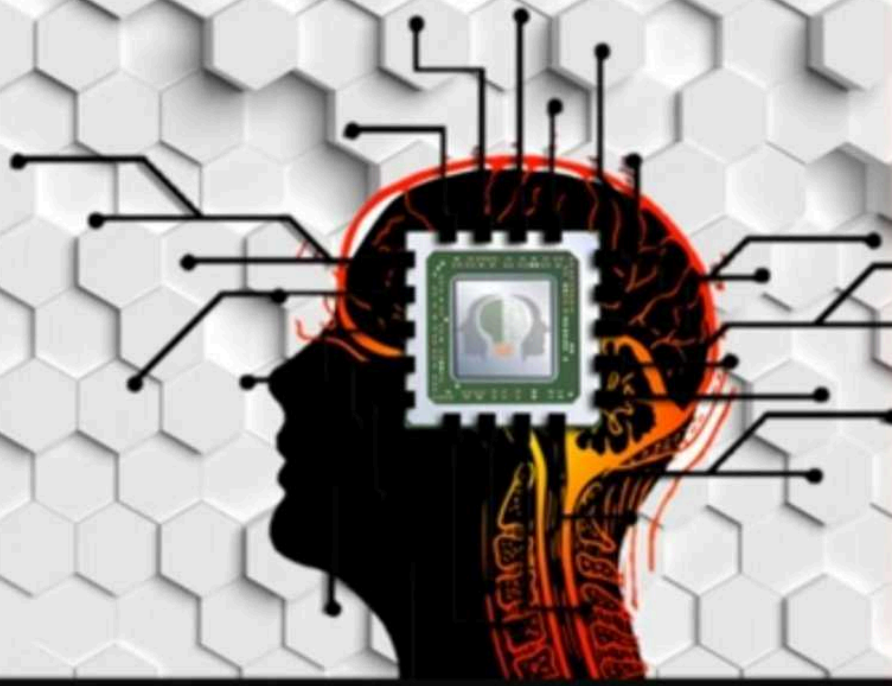
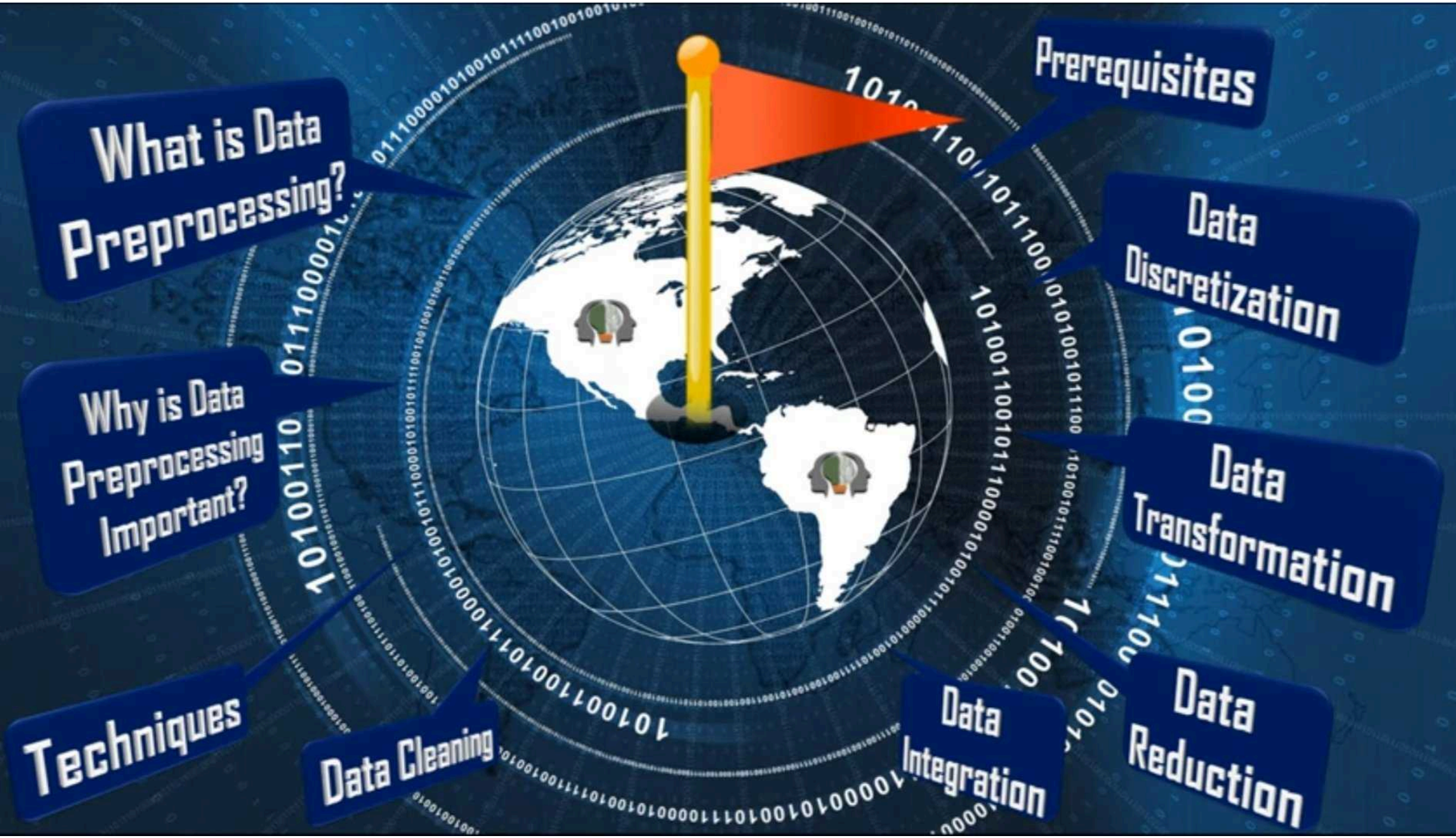
Data Preprocessing & Feature Engineering Tutorial

Introduction of Data Preprocessing - Techniques



Machine Learning & Data Science: Hands-on Python Course in Hindi


What is Data Preprocessing?
Why is Data Preprocessing Important?
Techniques
Data Cleaning
Data Integration
Data Reduction
Data Transformation
Data Discretization
Prerequisites

What is Data Preprocessing?

- Data
  - Text
  - Image
  - Video
  - Audio

- Data Preprocessing is a process to convert raw data into meaningful data using different techniques.

Why is Data Preprocessing Important?

- **Data in the real world is dirty**
  - Incomplete
  - Noisy
  - Inconsistent
  - Duplicate
- **Data Quality Elements**
  - Accuracy
  - Completeness
  - Consistency
  - Believability
  - Interpretability
- **Machine Learning algorithm follow the rule *(learn like Kids)***
  - GIGO – Garbage In Garbage Out

Steps/Technique in Data Preprocessing

- **Major steps in Data Preprocessing**
  - **Data Cleaning**
  - **Data Integration**
  - **Data Reduction**
  - **Data Transformation**
  - **Data Discretization**

What is Data Cleaning?

- **Data Cleaning means fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.**

CO2 emissions (metric tons per capita)

Country Name	2000	2001	2002
United States	20.17875	19.63651	NAN
United Kingdom	9.199549	9.233175	8.904123
India	0.97987	NAN	0.967381
China	2.696862	2.742121	3.007083
Russian Federation	NAN	10.6696	10.7159
Australia	17.20061	16.73337	17.37045

CO2 emissions (metric tons per capita)

Country Name	2000	2001	2002
United States	20.17875	19.63651	19.6134
United Kingdom	9.199549	9.233175	8.904123
India	0.97987	0.971698	0.967381
China	2.696862	2.742121	3.007083
Russian Federation	10.62712	10.6696	10.7159
Australia	17.20061	16.73337	17.37045

What is Data Integration?

- **Data Integration** is a technique to merges data from multiple sources into a coherent data store, such as a data warehouse.

Country Name	2000	2001
United States	20.17875	19.63651
United Kingdom	9.199549	9.233175
India	0.97987	0.971698
China	2.696862	2.742121
Russian Federation	10.62712	10.6696
Australia	17.20061	16.73337

Country Code	Indicator Name
USA	CO2 emissions (metric tons per capita)
GBR	CO2 emissions (metric tons per capita)
IND	CO2 emissions (metric tons per capita)
CHN	CO2 emissions (metric tons per capita)
RUS	CO2 emissions (metric tons per capita)
AUS	CO2 emissions (metric tons per capita)

CO2 emissions (metric tons per capita)

Country Name	Country Code	Indicator Name	2000	2001	2002
United States	USA	CO2 emissions (metric tons per capita)	20.17875	19.63651	19.6134
United Kingdom	GBR	CO2 emissions (metric tons per capita)	9.199549	9.233175	8.904123
India	IND	CO2 emissions (metric tons per capita)	0.97987	0.971698	0.967381
China	CHN	CO2 emissions (metric tons per capita)	2.696862	2.742121	3.007083
Russian Federation	RUS	CO2 emissions (metric tons per capita)	10.62712	10.6696	10.7159
Australia	AUS	CO2 emissions (metric tons per capita)	17.20061	16.73337	17.37045

What is Data Reduction?

- **Data Reduction** is a technique use to reduce the data size by aggregating, eliminating redundant features, or clustering, for instance.

CO2 emissions (metric tons per capita)

Country Name	Country Code	Indicator Name	2000	2001	2002
United States	USA	CO2 emissions (metric tons per capita)	20.17875	19.63651	19.6134
United Kingdom	GBR	CO2 emissions (metric tons per capita)	9.199549	9.233175	8.904123
India	IND	CO2 emissions (metric tons per capita)	0.97987	0.971698	0.967381
China	CHN	CO2 emissions (metric tons per capita)	2.696862	2.742121	3.007083
Russian Federation	RUS	CO2 emissions (metric tons per capita)	10.62712	10.6696	10.7159
Australia	AUS	CO2 emissions (metric tons per capita)	17.20061	16.73337	17.37045

CO2 emissions (metric tons per capita)

Country Code	2000	2002
USA	20.17875	19.6134
GBR	9.199549	8.904123
IND	0.97987	0.967381
CHN	2.696862	3.007083
RUS	10.62712	10.7159
AUS	17.20061	17.37045

What is Data Transformation?

- **Data Transformation means data are transformed or consolidated into forms appropriate for ML model training, such as normalization, may be applied where data are scaled to fall within a smaller range like 0.0 to 1.0.**
  - **Aggregation**
  - **Feature type conversion:**
  - **Normalization**
  - **Attribute/feature construction**

What is Data Discretization?

- **Data Discretization technique transforms numeric data by mapping values to interval or concept labels.**
- **It can be used to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals.**
- **Discretization techniques include -**
  - **Binning**
  - **Histogram analysis**
  - **Cluster analysis**
  - **Decision-tree analysis**
  - **Correlation analysis**

Ex-

Age: 1,2,3,4,5,6,7,8,9

Output>>>

Age: 1-3, 4-6, 7-9

Prerequisites for Data Preprocessing

- **Python Libraries**
  - NumPy
  - Pandas
  - Matplotlib
  - Seaborn
  - Scikit Learn
- **Mathematics**
  - Statistics
  - Probability
  - Calculus
  - Linear Algebra

- **Software**
  - Anaconda
    - Jupyter Notebook & Spyder
- **Learning Platform**
  - www.YouTube.com/IndianAIProduction