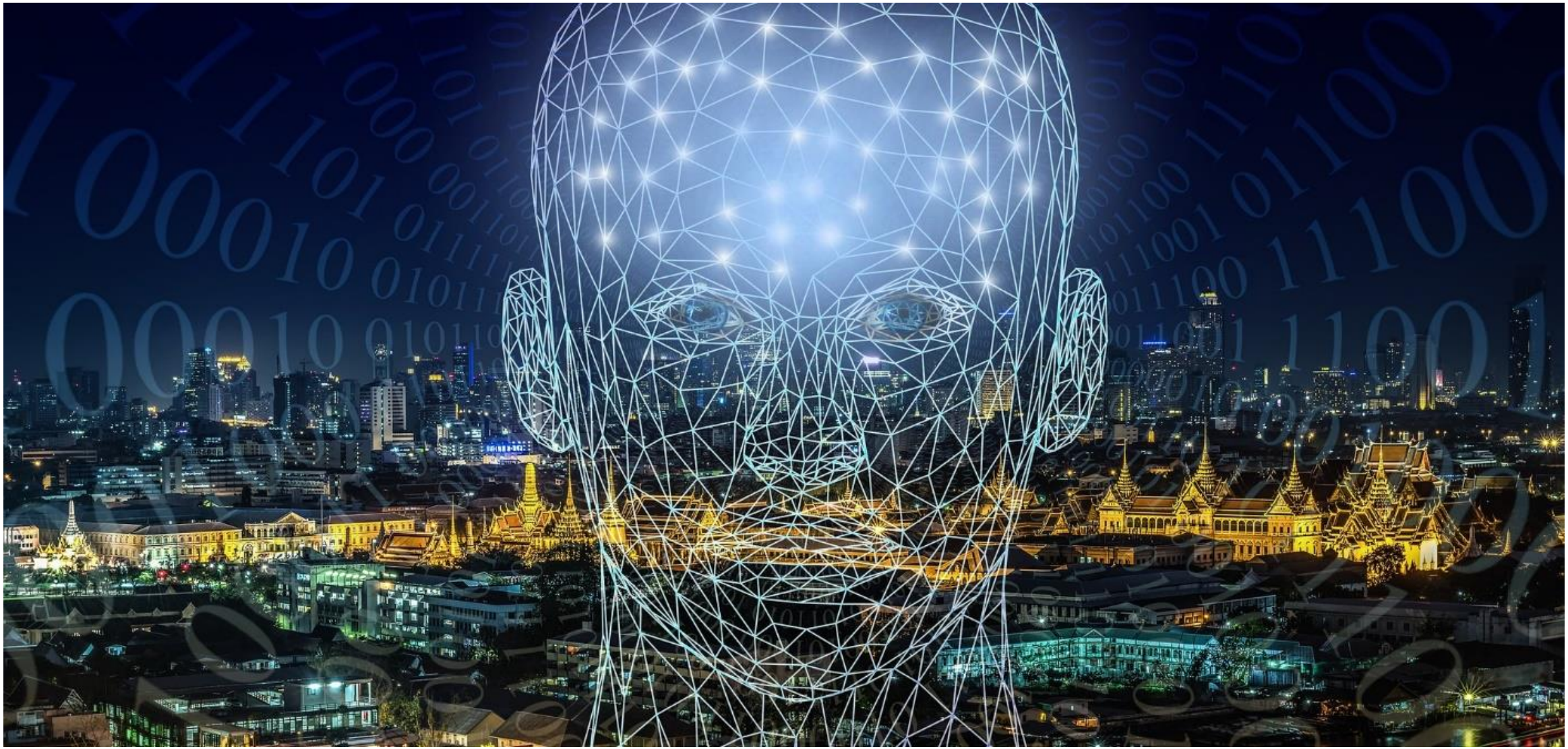
Data Preprocessing



Why Preprocessing is Required?

- Preprocessing refers to transformation before feeding to machine learning
- Quality of data is important to train the model
- Source – Government databases, professional or company data sources(twitter), your company, etc
- Data will never be in the format you need – Pandas Dataframe for reformatting

Why Preprocessing is Required? Cont…

- Columns to remove – No values, duplicate(correlated column, e.g: house size in ft and metres)
- Learning algorithms understands only number, converting text image to number is required
- Unscaled or unstandardized data might have unacceptable prediction

Types of Data Preprocessing

- Checking for Null Values
- Correlated Feature Check
- Data Molding (Label Encoding)
- Splitting the Data
- Impute Missing Values
- Data Standardization(Feature Scaling)
- Label Encoding
- One-Hot Encoding

Checking for Null Values

- Check for Null values
- Remove or Impute
- df.isnull().values.any()
- Drop row

Correlated Feature Check

- Sometimes two features that are meant to measure different characteristics of a model are influenced by common mechanism and they move together.

- **How to Handle Correlation**:
- Remove one of the feature
- Apply Principal Component Analysis(PLA)

Data Molding (Encoding)

- Adjusting Data Types - Inspect data types to see if there are any issues. Data should be numeric.
- If required create new columns

Splitting the Data

- **Variance** is the amount that the estimate of the target function will change given different training data.
- Less training data -> your parameter estimates have greater variance
- Less testing data -> your performance statistic will have greater variance
- **Divide data such that neither variance is too high**
- Less data -> chances of no satisfactory variance
- More data -> split doesn't really matter
- X = feature, independent, predictor Y = predicted, dependent

Impute Missing Values

- **Missing Data**
- Drop rows
- Replace values (Impute)

Data Standardization(Feature Scaling)

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

$$z = \frac{(x-\mu)}{\sigma}$$

- **Disadvantage**:
- Without Feature Scaling a machine learning algorithm tends to weigh greater values -> higher and consider smaller values as the lower values, regardless of the unit of the values.

Label Encoding & Ordinal Encoding

Convert text values to numbers. These can be used in the following situations:

- There are only two values for a column in your data. The values will then become 0/1 - effectively a binary representation
- The values have relationship with each other where comparisons are meaningful (e.g. low<medium<high)

One-Hot Encoding

- Use when there is no meaningful comparison between values in the column
- Creates a new column for each unique value for the specified feature in the data set

Dummy Variables

- Dummy Variable Trap

Profit	Admin Exp	R&D	Ad Spend	City
230	4	43	2	Delhi
423	3	12	6	Bangalore
324	7	45	4	Delhi

City_Delhi	City_Bangalore
1	0
0	1
1	0

- y = b0 + b1x1 + b2x2 + b3x3 + b4x4 + b5x5
- Here x4,x5 are dummy variable
- x5 = 1 – x4
- Multicollinearity -> that's why its called as dummy variable
- for 2 -> 1 & 0
- For: > 2 -> column

Quiz

- **Which of the following correlation coefficients value shows the strongest relationship?**

A. 0.91

B. 0.12

C. -0.12

D. - 0.91

Quiz

- **What is the full range of Correlation Coefficient? Choose the best answer**

A. 0 to 1

B. -1 to 0

C. -1 to 1

D. -2 to 2

Quiz

- **Which one of the below can not be used to handle Missing Value in a data?**

A. Mean Value

B. Most Frequent Value

C. Maximum Value

D. Remove missing observation, if dataset is large and no. of missing observations are less

Quiz

- **What are some examples of data quality problems?**

A. Duplicate Data

B. Correlation between features

C. Missing values

Quiz

- **Which Method is used for encoding the categorical variables**

A. LabelEncoder

B. OneHotEncoder

C. None of the Above

D. All of the Above

Quiz

- **Which of the below is valid for Imputation**

A. Imputation with mean/median

B. Imputing with random numbers

C. Imputing with one

D. All of the Above

Quiz

- **What's the purpose of feature scaling**

A. Reducing the training time

B. Getting better accuracy

C. Both A and B

D. None