Pre-processing and EDA (Exploratory Data Analysis)



Preprocessing:

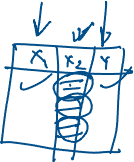


1. Taking care of redundant data – Missing Values, Outliers, Duplicates etc.



Handling Missing Values:

df.isna().sum()



Imputing:



From sklearn.impute import SimpleImputer

Imputer = SimpleImputer(fill\_value = np.nan, strategy = ‘median’)

X = imputer.fit\_transform(df)



B. df.dropna()



C. Predict for the missing values



Duplicates:



Df.duplicated(keep=’first’)

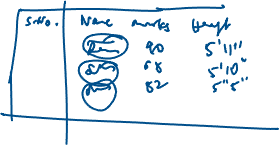
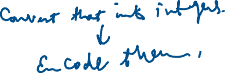


1. Train Test Split:

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y, test\_size = 0.25)

1. Handling Categorical Variables:



1. Label Encoding – convert cat variables into numerical labels.

Eg : (India -1, China -0, US -2, UK - 4, Italy – 3)

Drawback: it gives the highest priority to any one category due to its label is high.

From sklearn.preprocessing import LabelEncoder

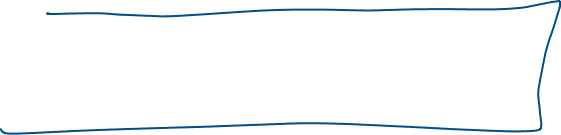
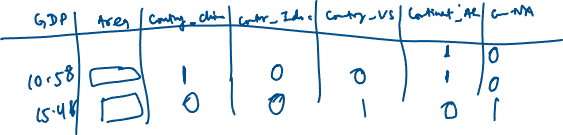
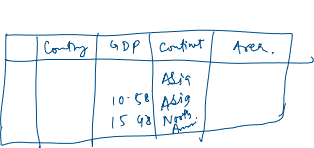
Le = LabelEncoder()

Le.fit(catDf[‘Country’])

catDf.Country=Le.transform(catDf.Country)

print(catDf)

1. One Hot Encoding:



1st Method:

Pd.get\_dummies(data=catDf)

2nd Method:

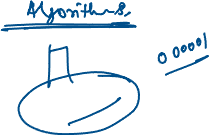
From sklearn.preprocessing import OneHotEncoder

ohe = OneHotEncoder()

df1 = pd.DataFrame(ohe.fit\_transform(catDf.iloc[:,[0,3]]))

pd.concat([catDf,df1], axis = 1)

Normalize/Standardize the data



>>>>>>>EOF