**PUNE INSTITUTE OF COMPUTER TECHNOLOGY**

**Subject: Machine Learning (LP-1 LAB)**

**Name: Aditya Kangune** **Roll No. : 33323**

**Batch: K11** **Academic Year: 2021-22**

---------------------------------------------------------------------------------------------

## Assignment 1
## Data Preparation

---------------------------------------------------------------------------------------------

### Problem Statement:

Perform the following operations on the given dataset:

1. Find Shape of Data
2. Find Missing Values
3. Find the data type of each column
4. Finding out Zero's
5. Find Mean age of patients
6. Now extract only Age, Sex, chest pain, RestBP, Chol. Randomly divide dataset in training (75%) and testing (25%).
7. Through the diagnosis test, I predicted 100 reports as COVID positive, but only 45 of those were actually positive. A total of 50 people in my sample were actually COVID positive. I have a total of 500 samples.
8. Create a confusion matrix based on the above data and find
   a. Accuracy
   b. Precision
   c. Recall
   d. F-1 score

### Objective:

This assignment will help the students to realize what is the need of data preparation.

## Theory:

### Data Preparation:

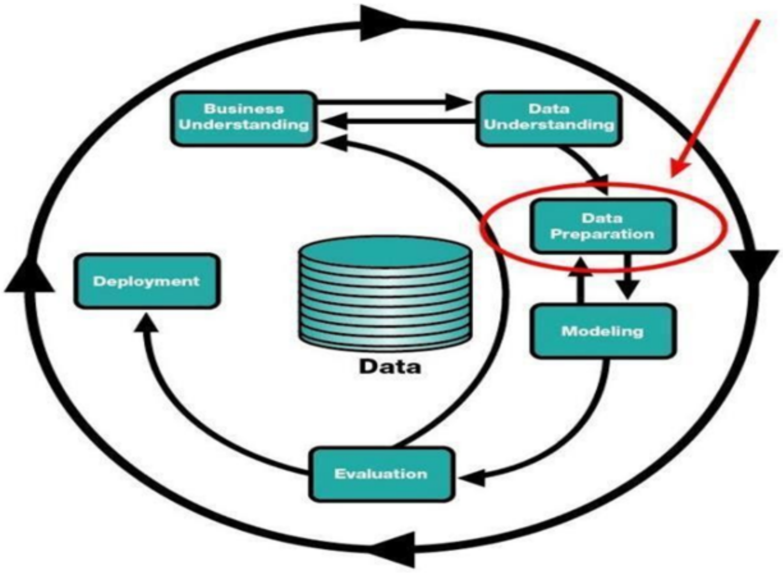
Data preparation (also referred to as “data preprocessing”) is the process of transforming raw data so that data scientists and analysts can run it through machine learning algorithms to uncover insights or make predictions.

### Why is Data Preparation Important?

1. Most machine learning algorithms require data to be formatted in a very specific way, so datasets generally require some amount of preparation before they can yield useful insights.
2. Some datasets have values that are missing, invalid, or otherwise difficult for an algorithm to process.
3. If data is missing, the algorithm can’t use it.
4. If data is invalid, the algorithm produces less accurate or even misleading outcomes.
5. Some datasets are relatively clean but need to be shaped (e.g., aggregated or pivoted) and many datasets are just lacking useful business context (e.g., poorly defined ID values), hence the need for feature enrichment.
6. Good data preparation produces clean and well-curated data which leads to more practical, accurate model outcomes.

1. It is the most required process before feeding the data into the machine learning model.
2. The reason behind that the data set needs to be different and specific according to the model so that we have to find out the required features of that data.
3. The data preparation process offers a method via which we can prepare the data for defining the project and also for the project evaluation of ML algorithms.
4. Different many predicting machine learning models are there with a different process but some of the processes are common that are

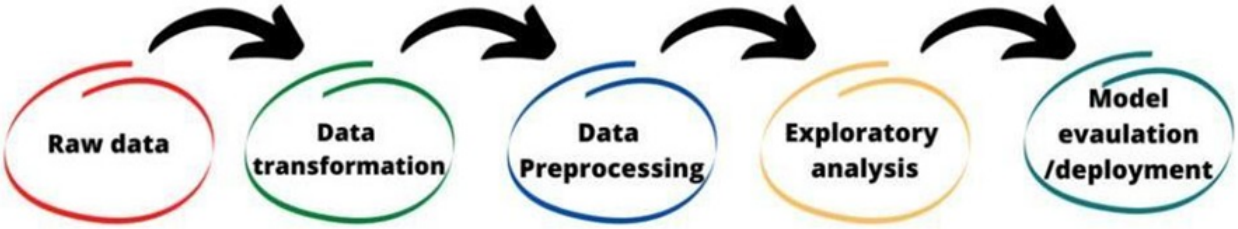
performed in every model, and also it allows us to find out the actual business problem and their solutions.



Data Preparation [3]

Some of the data preparation processes are:

1) Determine the problems
2) Data cleaning
3) Feature selection
4) Data transformation
5) Feature engineering
6) Dimensionality reduction



## 1. Determine the problems:

1. This step tells us about the learning method of the project to find out the results for future prediction or forecasting.
2. For example, which ML model is suitable for the data set regression or classification or clustering algorithms.
3. This includes data collection that is useful for predicting the result and also involving communication to project stakeholders and domain expertise. We use classification and regression models for categorical and numerical data respectively.
4. It includes determining the relevant attributes with the stied data in form of .csv, .html, .json, .doc, and many, also for unstructured data in a form for audio, video, text, images, etc for scanning and detect the patterns of data with searching and identifying the data that have taken from external repositories.

## 2. Data cleaning:

1. After collecting the data, it is very necessary to clean that data and make it proper for the ML model.
2. It includes solving problems like outliers, inconsistency, missing values, incorrect, skewed, and trends.
3. Cleaning the data is very important as the model learns from that data only, so if we feed inconsistent, appropriate data to the model it will return garbage only, so it is required to make sure that the data does not contain any unseen problem.
4. For example, if we have a data set of sales, it might be possible that it contains some features like height, age, that cannot help in the model building so we can remove it.
5. We generally remove the null values columns, fill the missing values, make the data set consistent, and remove the outliers and skewed data in data cleaning.

## 3. Feature selection:

1. Sometimes we face the problem of identifying the related features from the set of data and deleting the irrelevant and less important data without touching the target variables to get the better accuracy of the model.
2. Features selection plays a wide role in building a machine learning model that impacts the performance and accuracy of the model.
3. It is that process that contributes mostly to the predictions or output that we need by selecting the features automatically or manually.
4. If we have irrelevant data that would cause the model with overfitting and underfitting.

### The benefits of feature selection:

1. Reduce the overfitting/underfitting
2. Improves the accuracy
3. Reduced training/testing time
4. Improves performance

## 4. Data transformation:

1. Data transformation is the process that converts the data from one form to another.
2. It is required for data integration and data management. In data transformation, we can change the types of data, clear the data removing the null values or duplicate values, and get enriched data that depends on the requirements of the model.
3. It allows us to perform data mapping that determines how individual features are mapped, modified, filtered, aggregated, and joined.
4. Data transformation is needed for both structured and unstructured data but it is time-consuming, costly, slow.

## 5. Feature engineering:

1. Every ML algorithms use some input data for giving the required output and this input required some features which are in a structured form.
2. To get the proper result the algorithms required features with some specific characteristics which we find out with feature engineering.
3. We need to perform different feature engineering on different datasets and we can observe their effect on model performance.

Here I am listing out the techniques of feature engineering.

1. Imputation
2. Handling outliers
3. Binning
4. Log transform
5. one-hot encoding
6. Grouping operations
7. Feature split
8. Scaling

## 6. Dimensionality reduction:

1. When we use the dataset for building an ML model, we need to work with 1000s of features that cause the curse of dimensionality, or we can say that it refers to the process to convert a set of data.
2. For the ML model, we have to access a large amount of data and that large amount of data can lead us to a situation where we can take possible data that can be available to feed it into a forecasting model to predict and give the result of the target variable.
3. It reduced the time that is required for training and testing our machine learning model and also helps to eliminate over-fitting.
4. It is kind of zipping the data for the model.

## Conclusion:

Data preparation is recognized for helping businesses and analytics to get ready and prepare the data for operations.

## Implementation: