PRESIDENCY UNIVERSITY

BENGALURU

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

MODULE 2

SUPERVISED MACHINE LEARNING ALGORITHMS

CONTENT

- Introduction to the Machine Learning (ML) Framework
- Types of ML,
- Types of variables/features used in ML algorithms,
- One-hot encoding,
- Simple Linear Regression,
- Multiple Linear Regression,
- Evaluation metrics for regression model

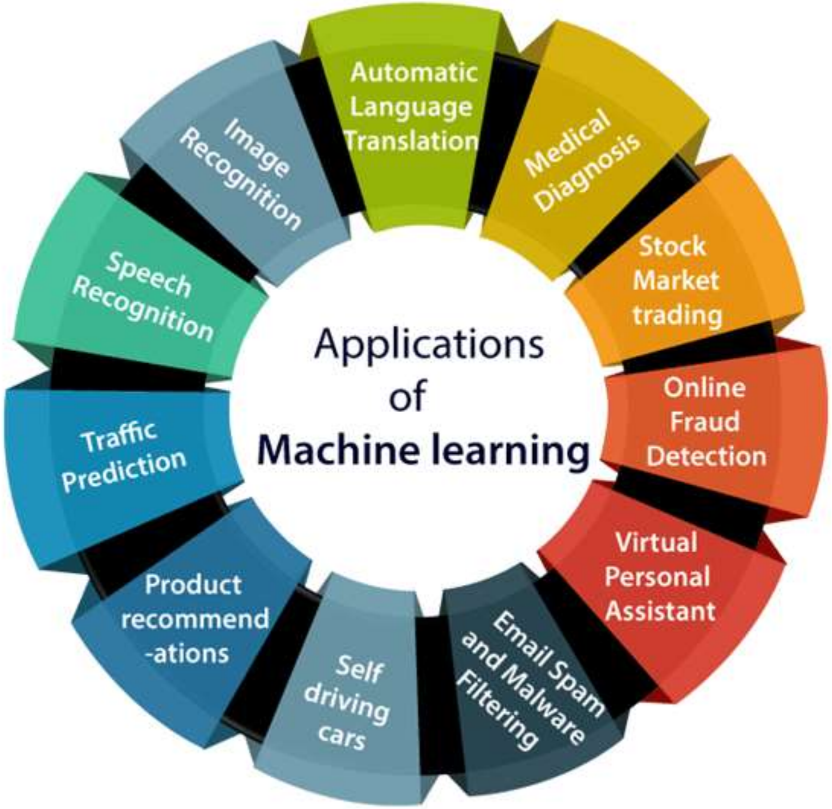
CONTENT

- Classification models
- Decision Tree algorithms using Entropy and Gini Index as measures of node impurity,
- Model evaluation metrics for classification algorithms,
- Multi-class classification
- Class Imbalance problem.
- Naïve Bayes Classifiers
- Naive Bayes model for sentiment classification

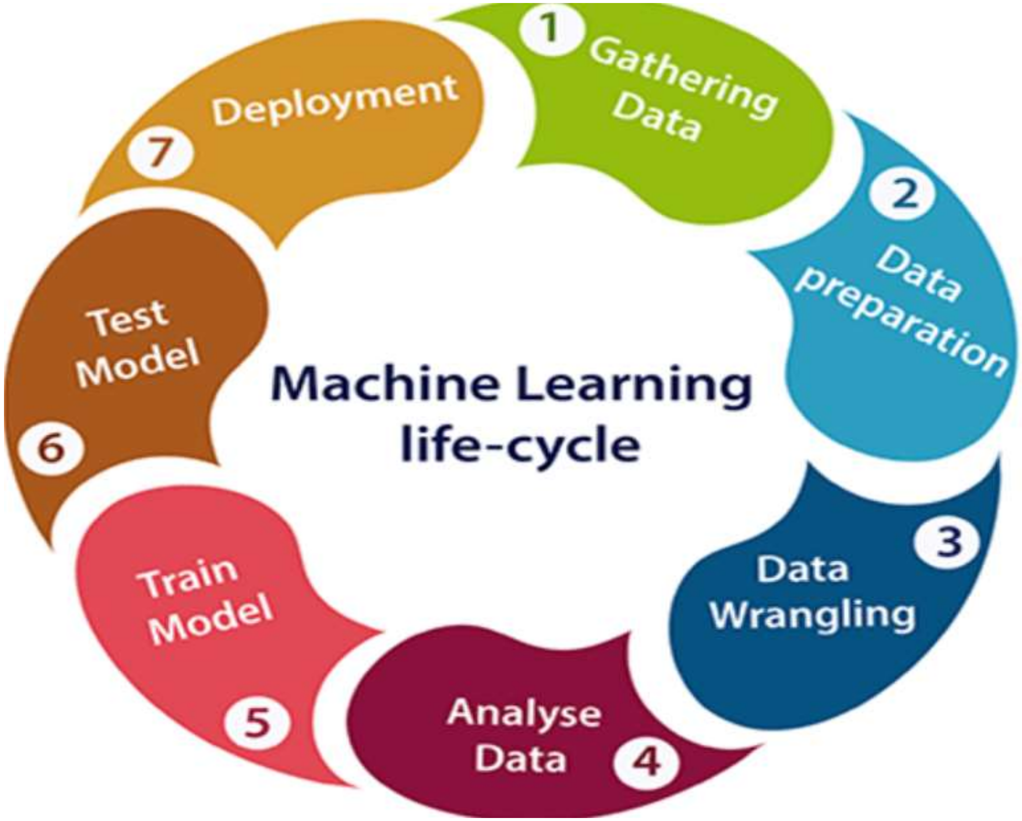
WHAT IS MACHINE LEARNING

- **Definition**
  - A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it.
  - The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.
- **Importance of machine learning**
  - Finding hidden patterns and extracting useful information from data
  - Solving complex problems and decision making in many fields (applications)
- **Feature of ML: Data-Driven Technology**
  - similar to data mining as it also deals with huge amount of data.
  - uses data to detect various patterns in a given dataset
  - learn from past data and improve automatically.

APPLICATIONS OF MACHINE LEARNING


Applications of Machine learning
Automatic Language Translation
Medical Diagnosis
Stock Market trading
Online Fraud Detection
Virtual Personal Assistant
Email Spam and Malware Filtering
Self driving cars
Product recommend -ations
Traffic Prediction
Speech Recognition
Image Recognition

Lifecycle of machine learning



LIFECYCLE OF MACHINE LEARNING

Gathering Data

- To identify the different data sources, as data can be collected from various sources such as **files**, **database** or **internet**.
- The quantity and quality of the collected data will determine the accuracy of the prediction and efficiency of the output.
- This step includes the below tasks:

➢ Identify various data sources

➢ Collect data

➢ Integrate the data obtained from different sources – This coherent set of data is called dataset

LIFECYCLE OF MACHINE LEARNING

Data preparation: This step can be further divided into two processes:

- **Data exploration:** To understand the characteristics, format, and quality of data to find Correlations, general trends, and outliers for an effective outcome.
- Characteristics are size, quality and accuracy of data analysis in data exploration.
- Outliers are incorrect data entry, measure the errors.
- **Data pre-processing:** Cleaning of data is required to address the quality issues: Missing Values, Duplicate data, Invalid data and Noise, which can be solved using filtering techniques.
- Filtering techniques are tracking pattern, classification, clustering etc.

Lifecycle of machine learning

Data Wrangling

- Reorganizing, mapping and transforming raw, unstructured data to a useable format.
- Unstructure data are information are not arranged such the text or multimedia data
- Sturture data are information arranged in the form of table in excel, data store in table which consist of row and column, and table in SQL.
- **This step involves data aggregation and data visualization.**
- Data Aggregation is process of summarizing the large pool or data for high level analysis (summarizing data multiple sources) summarization ex: sum, average, mean, median etc.
- Data Visualization: graphical representation of information and data, By using visual elements like charts, graphs, and maps, data visualization tools ...

Lifecycle of Machine Learning

Data Analysis

- The aim of this step is to build a machine learning model to analyze the data and review the outcome.

Train Model

- Datasets are used to train the model using various machine learning algorithms – to understand various patterns, rules, and, features.

Test Model

- Tests accuracy of the model with respect to the requirements of project or problem.

Deployment

- Performance of the project is checked with the available data and deployed which is similar to making the final report for a project.

Difference between AI & ML

Artificial Intelligence	Machine learning
Artificial intelligence is a technology which enables a machine to simulate human behavior.	Machine learning is a subset of AI which allows a machine to automatically learn from past data without programming explicitly.
In AI, we make intelligent systems to perform any task like a human.	In ML, we teach machines with data to perform a particular task and give an accurate result.
AI is working to create an intelligent system which can perform various complex tasks.	Machine learning is working to create machines that can perform only those specific tasks for which they are trained.
The main applications of AI are Siri, Expert System, Online game playing, intelligent humanoid robot, etc.	The main applications of machine learning are Online recommender system and Google search algorithms

MACHINE LEARNING - DATASET

- **A dataset** is a collection of data in which data is arranged in some order. A dataset can contain any data from a series of an array to a database table.
- **Types of data in datasets**
  - **Numerical data:** Such as house price, temperature, etc.
  - **Categorical data:** Such as Yes/No, True/False, Blue/green, etc.
  - **Ordinal data:** These data are similar to categorical data but can be measured on the basis of comparison.
  - the data can be categorized while introducing an order or ranking.
  - Example of Ordinal data – Customer Feedback, Economic Status
- **Types of Datasets**
  - **Training Dataset**: This data set is used to train the model i.e.; these datasets are used to update the weight of the model.

CONTD…

- **Validation Dataset**
  - It is used to verify that the increase in the accuracy of the training dataset is actually increased if we test the model with the data that is not used in the training.
  - If the accuracy over the training dataset increases while the accuracy over the validation dataset decreases, then this results in the case of high variance i.e., overfitting.
  - Variance: average or mean, standard deviation.
  - Overfitting the machine learning model gives accurate predictions for training data but not for new data.

CONTD…

- **Test Dataset**
  - Most of the time when we try to make changes to the model based upon the output of the validation set then unintentionally, we make the model peek into our validation set and as a result, our model might get overfit on the validation set as well.
  - To overcome this issue, we have a test dataset that is only used to test the final output of the model in order to confirm the accuracy.

CONTD…

How to get the datasets / Popular sources for ML dataset

- Kaggle Dataset
- UCI Machine Learning Repository
- Datasets via AWS
- Google's Dataset Search Engine
- Microsoft Dataset
- Awesome Public Dataset Collection
- Government Datasets
- Computer Vision Datasets
- Scikit-learn dataset

MACHINE LEARNING-
DATA PREPROCESSING

- **Definition**: Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model.
- **Significance**
  - ➢ A real-world data contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models.
  - ➢ Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model.
- **Steps**

Getting the dataset

- ➢ The data is usually put in CSV file ("**Comma-Separated Values**" files; it is a file format which allows us to save the tabular data, such as spreadsheets. It is useful for huge datasets and can use these datasets in programs).

CONTD…

Importing libraries

- **Numpy:** used for including any type of mathematical operation in the code.
- **Matplotlib:** used to plot any type of charts in Python for the code.
- **Pandas:** used for importing and managing the datasets. It is an open-source data manipulation and analysis library.

Importing datasets

- **read_csv() function:** used to read a csv file
- To distinguish the matrix of features (independent variables) and dependent variables from dataset - **iloc[ ]** method is used to extract the required rows and columns from the dataset.
- To extract dependent variables, again, we will use Pandas.iloc[] method.

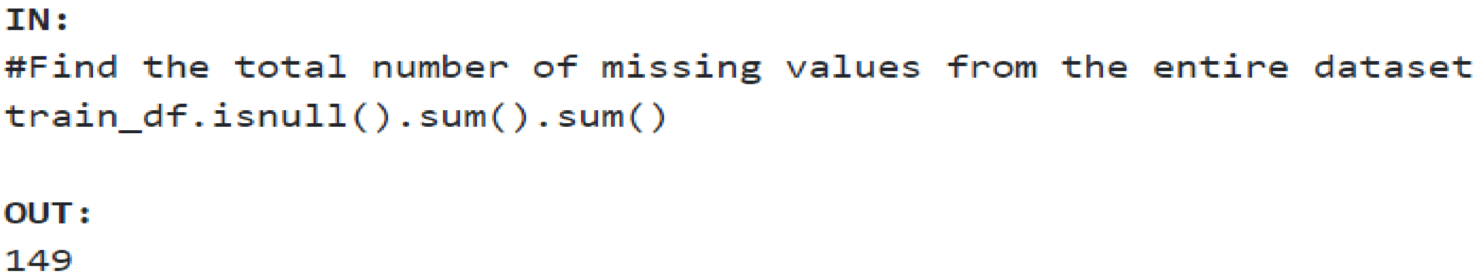
CONTD…

- Import numpy as np
- Import matplotlib as mt
- Import pandas as pd

# load the dataset

train_df = pd.read_csv("train.csv“)
train_df

# **6 columns** – Gender, Married, Dependents, Self_Employed, LoanAmount, Loan_Amount_Term, and Credit_History having missing values.



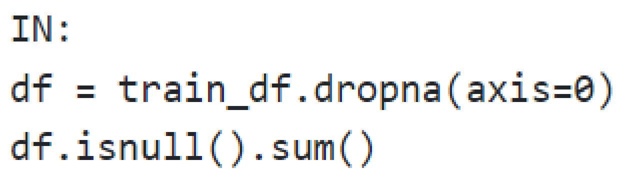
```
IN:
#Find the total number of missing values from the entire dataset
train_df.isnull().sum().sum()

OUT:
149
```

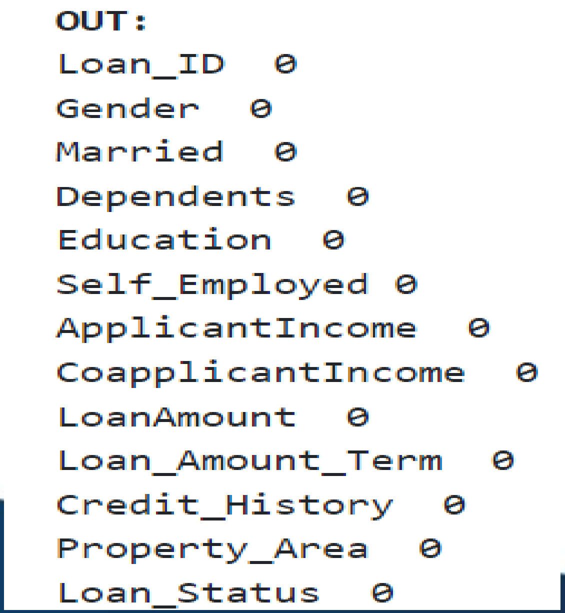
CONTD…

- *Handling Missing Data*

- **By deleting the particular row:** delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
- **Deleting the entire row (listwise deletion)** The code to drop the entire row is as follows:



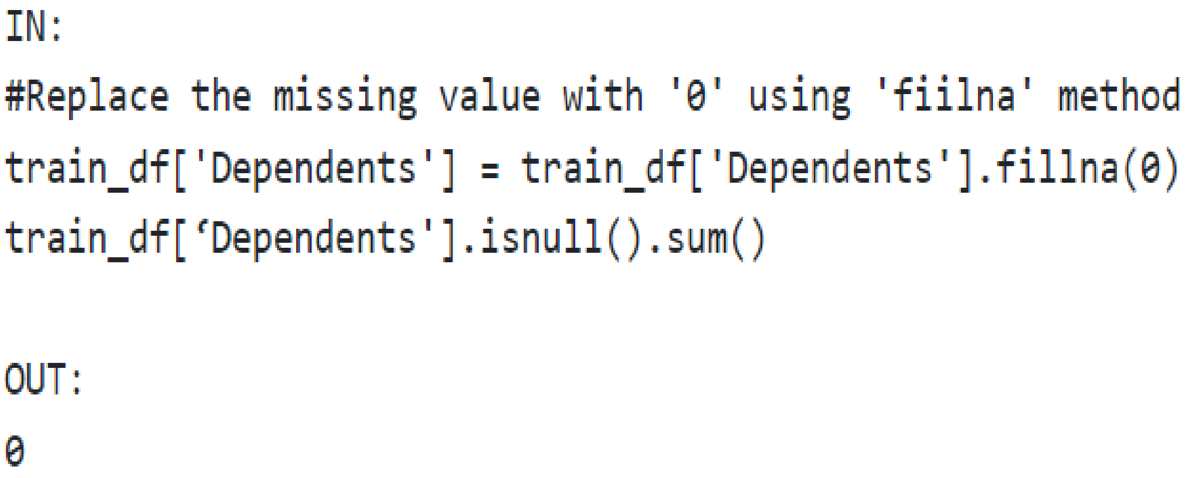
```
IN:
df = train_df.dropna(axis=0)
df.isnull().sum()
```



```
OUT:
Loan_ID  0
Gender  0
Married  0
Dependents  0
Education  0
Self_Employed 0
ApplicantIncome  0
CoapplicantIncome  0
LoanAmount  0
Loan_Amount_Term  0
Credit_History  0
Property_Area  0
Loan_Status  0
```

Imputing the Missing Value

- **Replacing with an arbitrary value**
- E.g., in the following code, we are replacing the missing values of the ‘Dependents’ column with ‘0’.

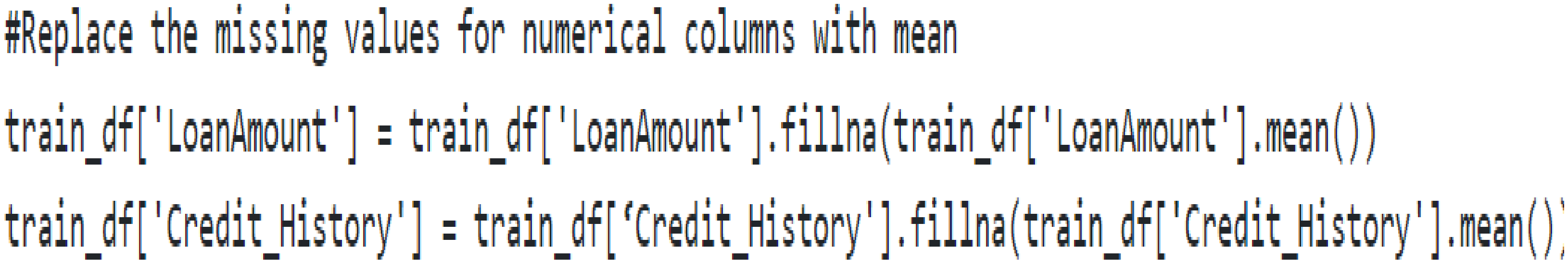


```
IN:
#Replace the missing value with '0' using 'fiilna' method
train_df['Dependents'] = train_df['Dependents'].fillna(0)
train_df[‘Dependents'].isnull().sum()

OUT:
0
```

CONTD…

- **By calculating the mean:** In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.
- Ex: use the ‘fillna’ method for imputing the columns ‘LoanAmount’ and ‘Credit_History’ with the mean of the respective column values.



```
#Replace the missing values for numerical columns with mean

train_df['LoanAmount'] = train_df['LoanAmount'].fillna(train_df['LoanAmount'].mean())

train_df['Credit_History'] = train_df[‘Credit_History'].fillna(train_df['Credit_History'].mean()
```

CONTD…

- To handle missing values, we will use **Scikit-learn** library in our code, which contains various libraries for building machine learning models. Here we will use **Import** class of **sklearn.preprocessing** library.

CONTD…

- ***Encoding Categorical Data***

○ **Label Encoder()** class from **pre-processing** library is used for encoding the variables into digits.

○ Categorical variables usually have strings for their values. Many machine learning algorithms do not support string values for the input variables. Therefore, we need to replace these string values with numbers. This process is called **categorical variable encoding**.

○ Types of encoding:

  ❖ **One-hot encoding**

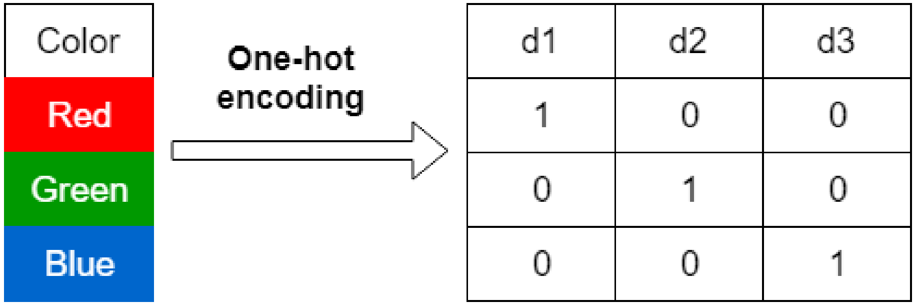
  ❖ **Dummy encoding**

❖ **One-hot encoding**

- In one-hot encoding, we create a new set of dummy (binary) variables that is equal to the number of categories (k) in the variable.

CONTD…

- For example, let’s say we have a categorical variable ***Color*** with three categories called “Red”, “Green” and “Blue”, we need to use three dummy variables to encode this variable using one-hot encoding. A dummy (binary) variable just takes the value 0 or 1 to indicate the exclusion or inclusion of a category.



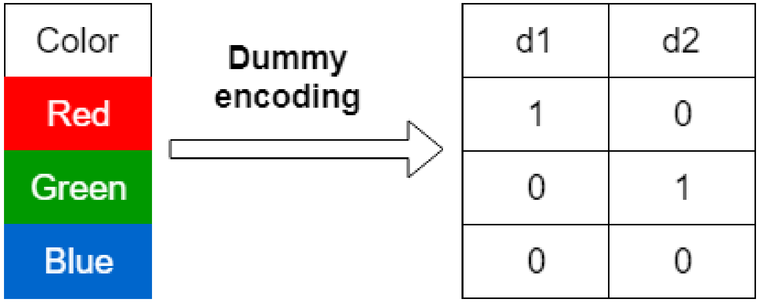
Color
Red
Green
Blue

d1	d2	d3
1	0	0
0	1	0
0	0	1

- In one-hot encoding,
- “Red” color is encoded as [1 0 0] vector of size 3.
- “Green” color is encoded as [0 1 0] vector of size 3.
- “Blue” color is encoded as [0 0 1] vector of size 3.

CONTD…

- Dummy encoding
  - Dummy encoding also uses dummy (binary) variables. Instead of creating a number of dummy variables that is equal to the number of categories (k) in the variable, dummy encoding uses k-1 dummy variables.
  - To encode the same Color variable with three categories using the dummy encoding, we need to use only two dummy variables



Color
Red
Green
Blue

d1	d2
1	0
0	1
0	0

  - In dummy encoding, “Red” color is encoded as [1 0] vector of size 2, “Green” color is encoded as [0 1] vector of size 2, “Blue” color is encoded as [0 0] vector of size 2.
  - Dummy encoding removes a duplicate category present in the one-hot encoding.

CONTD…

- **Splitting dataset into training, validation and test set**

- **Feature scaling**
  - Feature scaling is the final step of data pre-processing in machine learning.
  - several common techniques for feature scaling, including standardization(mean and standard deviation), normalization, and min-max scaling. These methods adjust the feature values while preserving their relative relationships and distributions.
  - **Normalization** is a scaling technique in which values are shifted and rescaled so that they end up ranging between 0 and 1. It is also known as Min-Max scaling.
  - It is a technique to standardize the independent variables of the dataset in a specific range.

Feature Selection Techniques in Machine Learning

FEATURE SELECTION

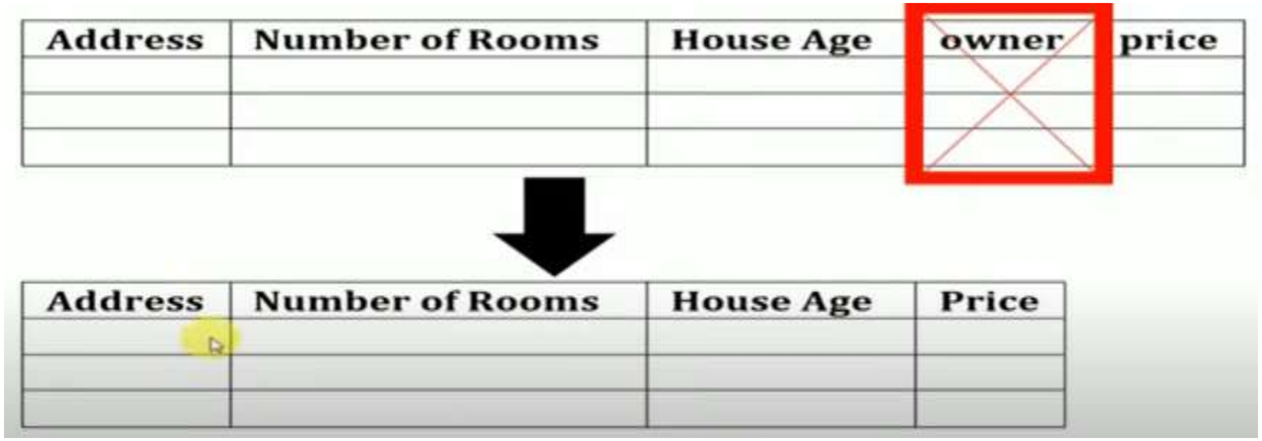
- A **feature** is an attribute that has an impact on a problem or is useful for the problem, and choosing the important features for the model is known as feature selection.
- Ex: Dataset is usually represented in a tabular form: rows and columns are features

Address	Number of Rooms	House Age	Owner	price

- To predict the price of the house then address, no.of rooms, house age, owner are features.
- **Definition**: Feature selection is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features.
- The goal of feature selection techniques in machine learning is to find the best set of features that allows one to build optimized models
- Applications like object detection, NLP, image retrieval etc.

- **Significance** of Feature Selection:
  - ➢ It helps in the simplification of the model so that it can be easily interpreted by the researchers.
  - ➢ It reduces the training time.
  - ➢ It reduces overfitting hence enhance the generalization.

- Why Do we need feature selection?
  - If there are too many features, our model can may become weak or generate some misleading patterns.



Address	Number of Rooms	House Age	~~owner~~	price

Address	Number of Rooms	House Age	Price

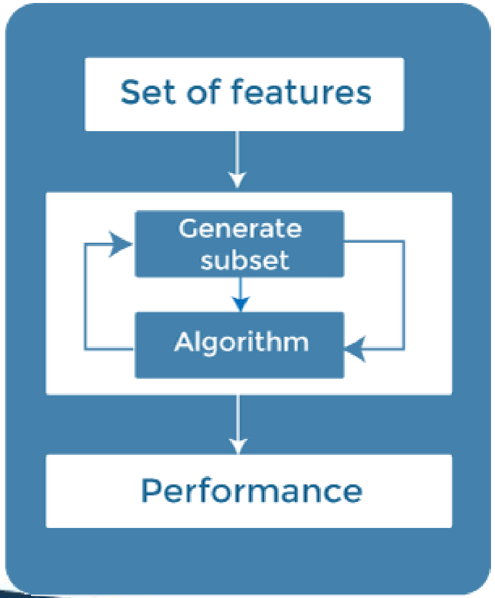
feature selection techniques

- Wrapper method
  - Forward Feature selection
  - Backward Feature Selection
- Filter Method
- Embedded Method

SUPERVISED FEATURE SELECTION TECHNIQUES

- **Wrapper Methods**
  - In wrapper methodology, selection of features is done by considering it as a search problem, in which different combinations are made, evaluated, and compared with other combinations.
  - It trains the algorithm by using the subset of features iteratively.
  - On the basis of the output of the model, features are added or subtracted, and with this feature set, the model has trained again.

Each subset of feature trains the model iteratively then finally Will get the maximum performance of the subset of features



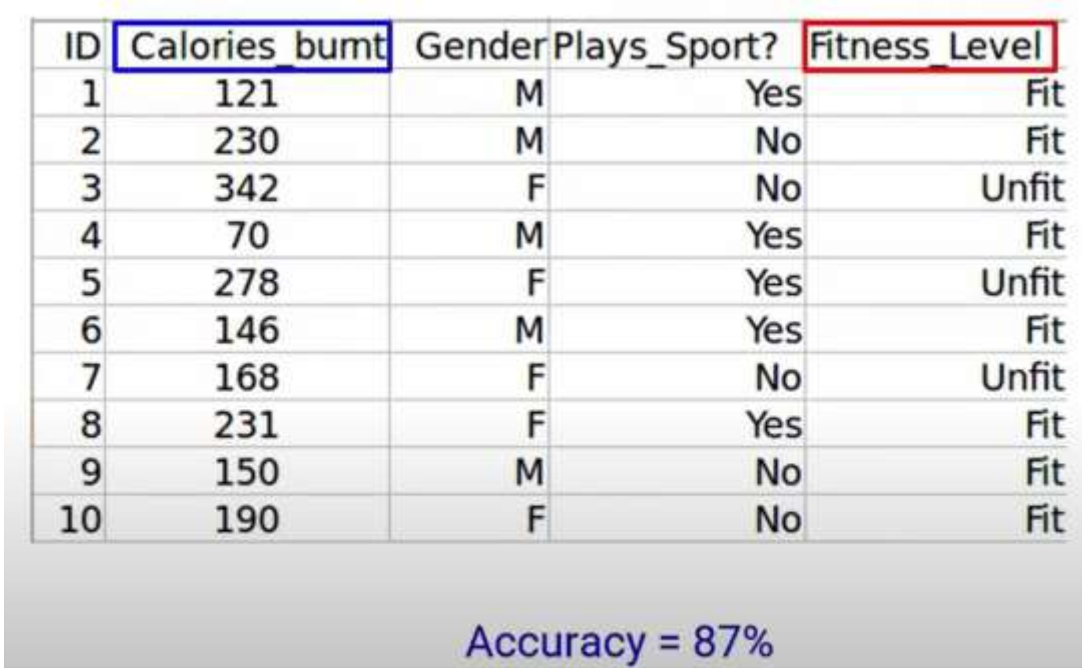
Generate feature subsets

- In **Forward wrapper methods**, we start from an empty feature set and add the feature maximizing the performance in each step until no substantial improvement is observed.
- So it there are **n features**, we build n ML model in the first iteration. Then we select the feature corresponding to the model with the best performance.
- In the second iteration, we repeat the process with the remaining **n-1 features**.
- We continue like this as long as there’s significant performance improvement between model with which we end successive iterations.

Ex: Forward Feature Selection

Step to perform forward feature selection

- Train n model using each feature (n) individually and check the performance

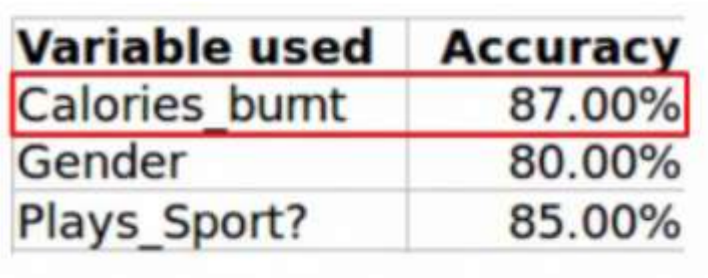


ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 87%

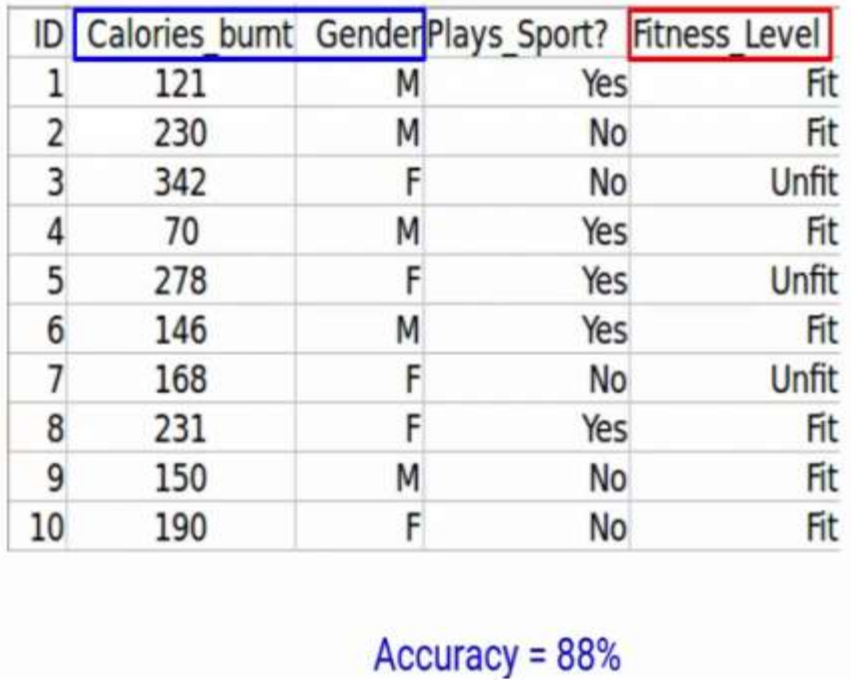
CONTD...

- Next, well train the model using the Gender feature, we get an accuracy of 80%
- Similarly, the plays_sport variable gives us a accuracy of 85%
- Now we will choose the variable, which gives us the best performance.



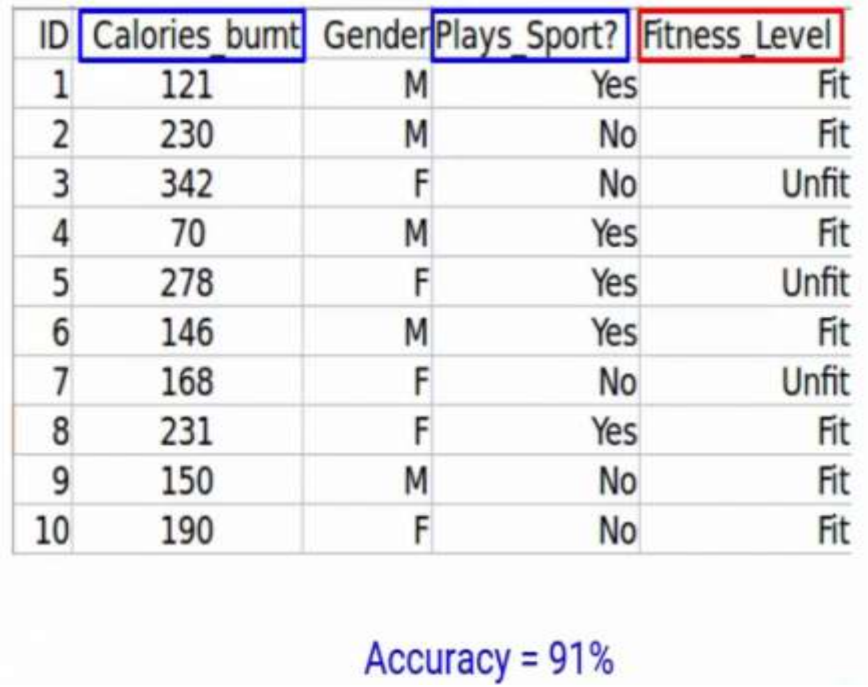
Variable used	Accuracy
Calories_bumt	87.00%
Gender	80.00%
Plays_Sport?	85.00%

- Next we will repeat this process and add one variable at a time.



ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

Accuracy = 88%



ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

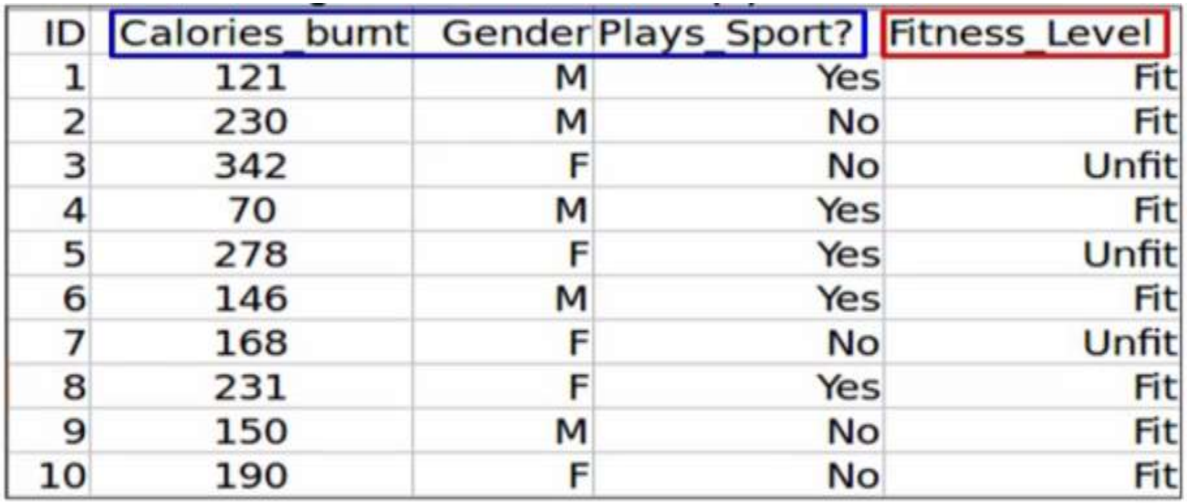
Accuracy = 91%

CONTD…

- **Backward methods:** work the opposite way.
- They start from the full feature set and remove them one by one.
- So in each iteration, they can remove a feature previously added as well as add a feature discarded in a previous step.

Ex. Backward feature selection

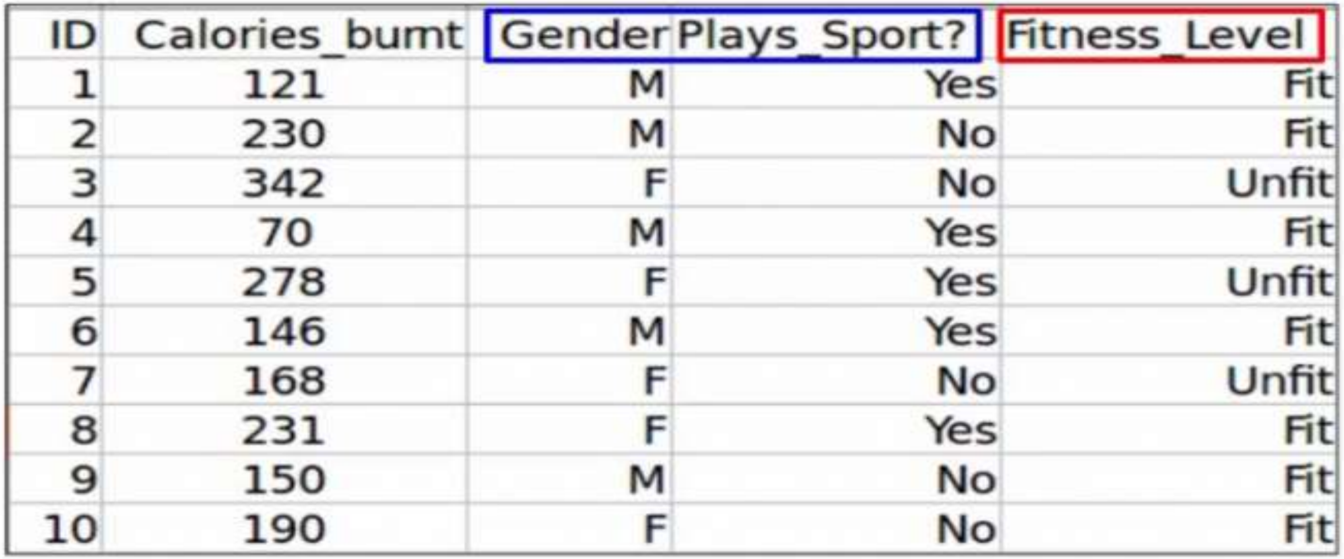
- The first step is to train the model using all the variables.



ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

- Accuracy 92%

- Next we will eliminate a variable and train the model on the remaining variable.

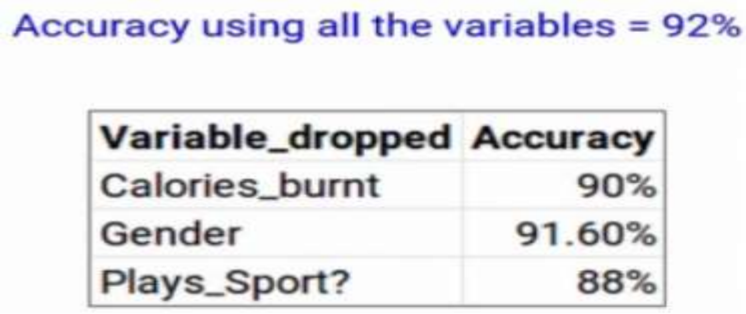


ID	Calories_bumt	Gender	Plays_Sport?	Fitness_Level
1	121	M	Yes	Fit
2	230	M	No	Fit
3	342	F	No	Unfit
4	70	M	Yes	Fit
5	278	F	Yes	Unfit
6	146	M	Yes	Fit
7	168	F	No	Unfit
8	231	F	Yes	Fit
9	150	M	No	Fit
10	190	F	No	Fit

- After drop one variable calories-burnt accuracy is 90%

CONTD…

- Similarly, we will drop each variable at a time and train the model on remaining variable.



Accuracy using all the variables = 92%

Variable_dropped	Accuracy
Calories_burnt	90%
Gender	91.60%
Plays_Sport?	88%

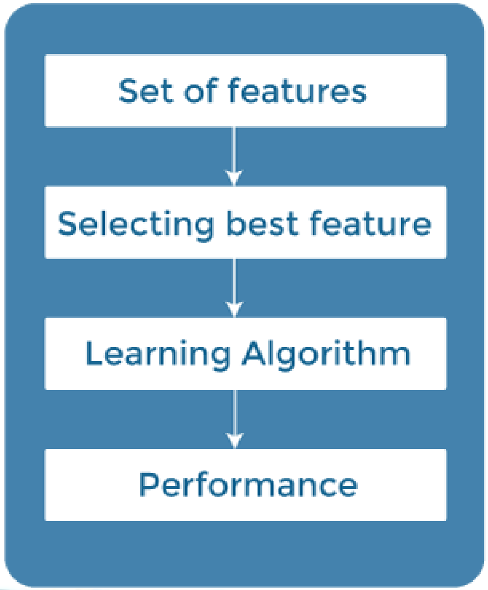
- Finally, we will repeat all these steps until no more variables can be dropped.

CONTD…

- **Filter Methods**
  - In Filter Method, features are selected on the basis of statistics measures. This method does not depend on the learning algorithm and chooses the features as a pre-processing step.
  - The filter method filters out the irrelevant feature and redundant columns from the model by using different metrics through ranking.
  - The advantage of using filter methods is that it needs low computational time and does not overfit the data.

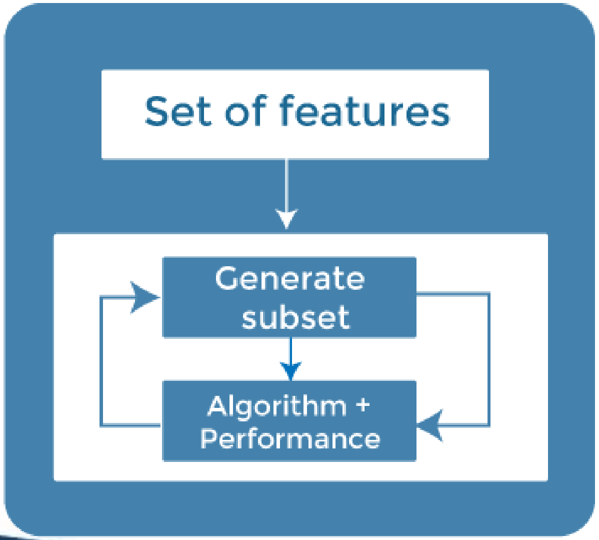
To Compute the correlation and estimate the strength of the relationship using statistical tools

variance thresholding, redundant column removal, Pearson correlation



CONTD…

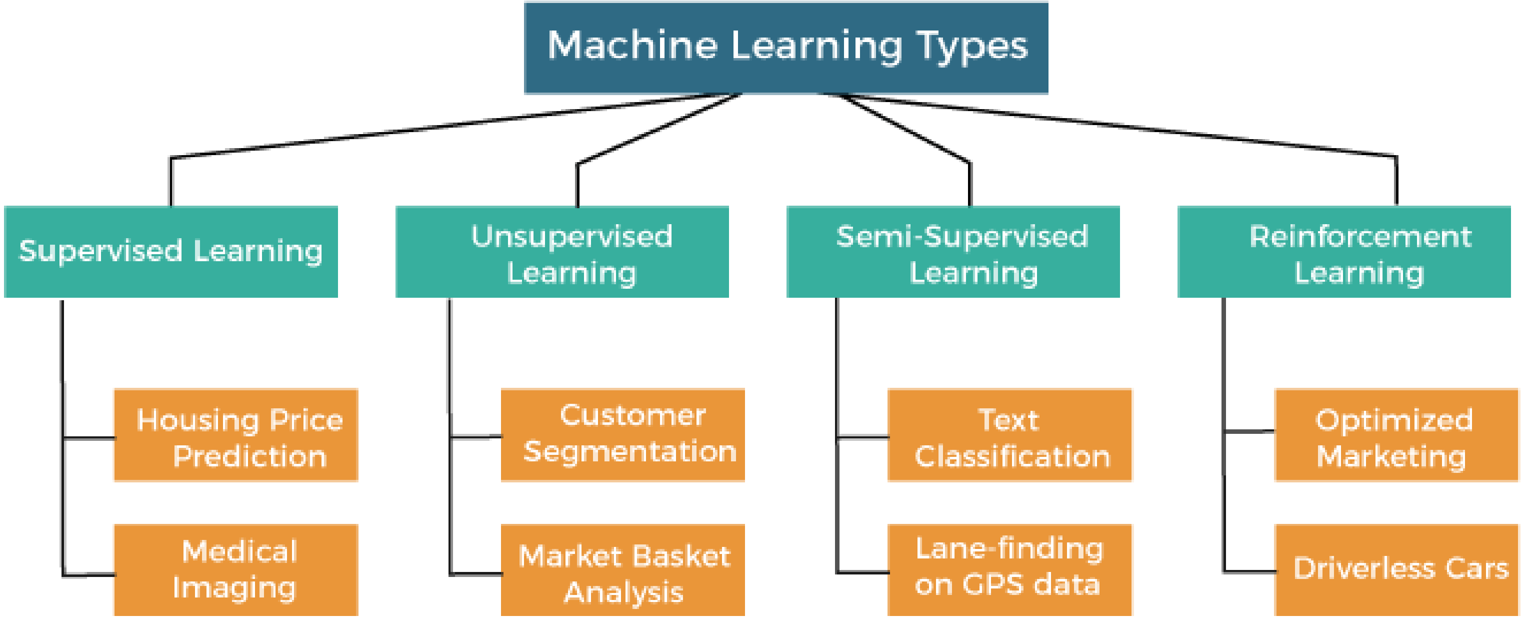
- **Embedded Methods**
  - Embedded methods combined the advantages of both filter and wrapper methods by considering the interaction of features along with low computational cost. These are fast processing methods similar to the filter method but more accurate than the filter method.
  - These methods are also iterative, which evaluates each iteration, and optimally finds the most important features that contribute the most to training in a particular iteration.



FEATURE ENGINEERING FOR MACHINE LEARNING

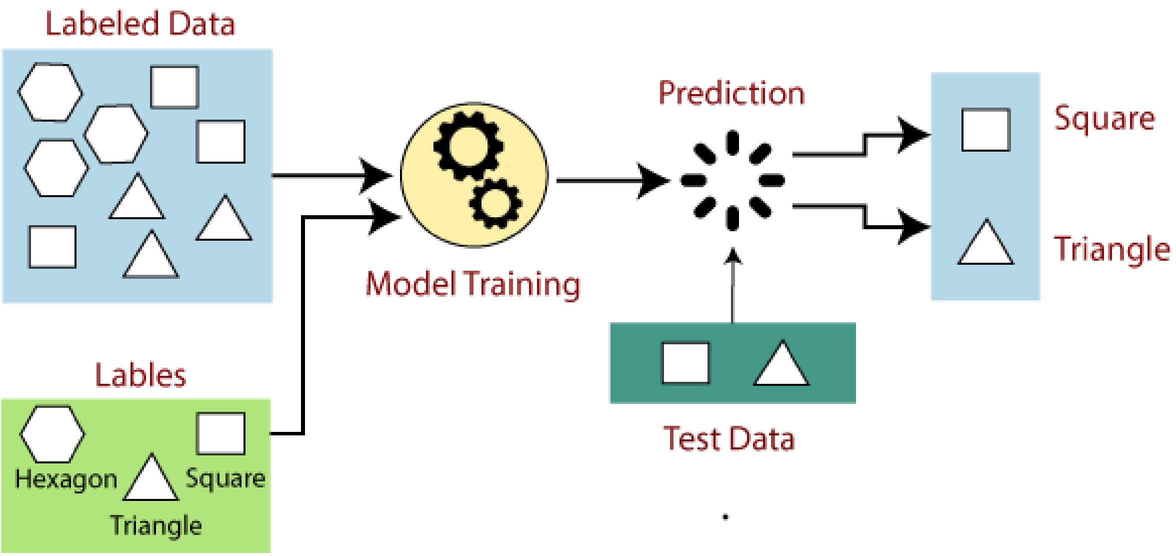
- **Feature engineering is the pre-processing step of machine learning, which extracts features from raw data**.
- Feature engineering in ML contains mainly four processes:
  - **Feature Creation:** finding the most useful variables to be used in a predictive model.,
  - **Transformations:** This step of feature engineering involves adjusting the predictor variable to improve the accuracy and performance of the model.
  - **Feature Extraction:** Is an automated feature engineering process that generates new variables by extracting them from the raw data
  - **Feature Selection:** Is a way of selecting the subset of the most relevant features from the original features set by removing the redundant, irrelevant, or noisy features

MACHINE LEARNING - TYPES



SUPERVISED LEARNING

- Supervised learning is the types of machine learning in which **machines are trained using well "labelled" training data**, and on basis of that data, machines predict the output. The labelled data means some **input data is already tagged with the correct output.**



Types of Supervised Learning

Regression Algorithms

- Are used if there is a relationship between the input variable and the output variable. Example: Weather forecasting, Market Trends, etc.
- Regression algorithms under supervised learning: Linear Regression, Non-Linear Regression, Polynomial Regression, Ridge Regression and Lasso Regression.

Classification Algorithms

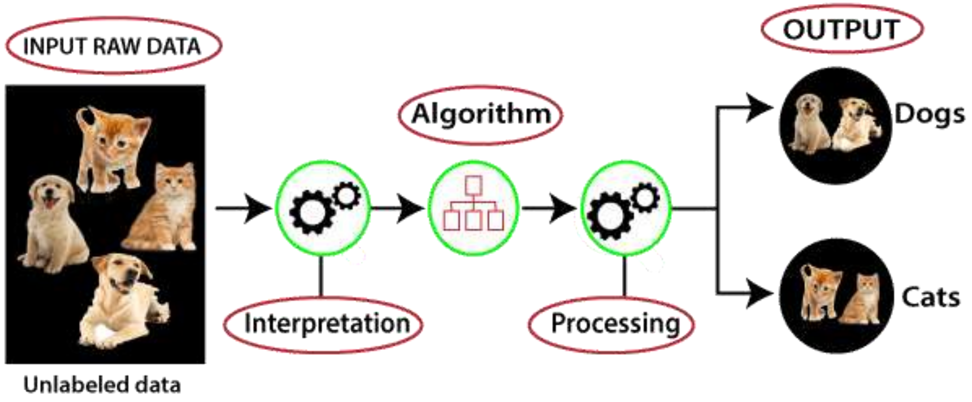
- Classification algorithms are used when the output variable is categorical, which means there are two classes such as Yes-No, Male-Female, True-false, etc. Example: Spam Filtering.
- Classification algorithms under supervised learning: Random Forest, Decision Trees, Logistic Regression, Support vector Machines

IMPORTANT TERMINOLOGIES

- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called **target variable**.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a **predictor**.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.
- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity.
- **Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**.
- **Underfitting:** If our algorithm does not perform well even with training dataset, then such problem is called **underfitting**.

UNSUPERVISED MACHINE LEARNING

- **Unsupervised learning** is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.



- The goal of unsupervised learning is to find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.

UNSUPERVISED MACHINE LEARNING - TYPES

- **Types:**
  - **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group.
  - **Association**: An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset.
- **Unsupervised learning algorithms**: K-means clustering, Hierarchal clustering, Anomaly detection, Neural Networks, Principle Component Analysis, Apriori algorithm
- **Advantage of Unsupervised Learning**: “Preferable” as it is easy to get unlabeled data in comparison to labeled data.
- **Disadvantages of Unsupervised Learning**: The result might be less accurate as input data is not labeled, and algorithms do not know the exact output in advance.

SEMI-SUPERVISED LEARNING

- **Semi-Supervised learning** is a type of Machine Learning algorithm that lies between Supervised and Unsupervised machine learning.
- The main aim of semi-supervised learning is to effectively use all the available data, rather than only labelled data like in supervised learning.
- **Advantages:** It is highly efficient and is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.
- **Disadvantages**
  - ➢ Iterations results may not be stable.
  - ➢ We cannot apply these algorithms to network-level data.
  - ➢ Accuracy is low.

Reinforcement Learning

- Reinforcement learning works on a feedback-based process, in which an AI agent (A software component) automatically explore its surrounding by hitting & trail, taking action, learning from experiences, and improving its performance.
- Agent gets rewarded for each good action and get punished for each bad action; hence the goal of reinforcement learning agent is to maximize the rewards.
- In reinforcement learning, there is no labelled data like supervised learning, and agents learn from their experiences only.
- Due to its way of working, reinforcement learning is employed in different fields such as **Game theory, Operation Research, Information theory, multi-agent systems.**
- A reinforcement learning problem can be formalized using **Markov Decision Process(MDP).**

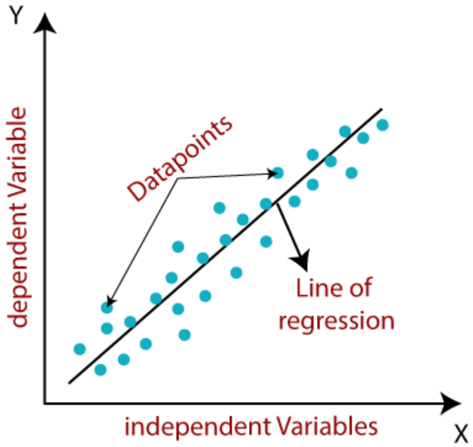
REINFORCEMENT LEARNING

- **Categories of Reinforcement Learning**
  - **Positive Reinforcement Learning:** Specifies increasing the tendency that the required behavior would occur again by adding something.
  - **Negative Reinforcement Learning:** It increases the tendency that the specific behavior would occur again by avoiding the negative condition.
- **Applications**: Robotics, Text Mining, Resource Management, Video Games.
- **Advantages**
  - The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
  - Helps in achieving long term results.
- **Disadvantages**
  - RL algorithms require huge data and computations.
  - Too much reinforcement learning can lead to an overload of states which can weaken the results.

LINEAR REGRESSION ANALYSIS

- It is a statistical method that is used for predictive analysis.
- Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.
- Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression.
- Mathematically, we can represent a linear regression as:

  $y= a_0+a_1x+ \varepsilon$



LINEAR REGRESSION ANALYSIS

- **Types of Linear Regression**
  - **Simple Linear Regression:** If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
  - **Multiple Linear regression:** If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.
- **Model Performance: R-squared method:**
  - R-squared is a statistical method that determines the goodness of fit.
  - The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
  - It can be calculated from the below formula:

$$\text{R-squared}=\frac{\text{Explained variation}}{\text{Total Variation}}$$

SIMPLE LINEAR REGRESSION

- Models the relationship between a dependent variable and a single independent variable. The relationship shown by a Simple Linear Regression model is linear or a sloped straight line.
- Simple Linear regression algorithm has mainly two objectives:
  - Model the relationship between the two variables. Eg: Income and expenditure, experience and Salary, etc.
  - Forecasting new observations. Such as Weather forecasting according to temperature, Revenue of a company according to the investments in a year, etc.
- The Simple Linear Regression model can be represented using the below equation:

$$y= a0+a1x+ \varepsilon$$

a0= It is the intercept of the Regression line (can be obtained putting x=0)

a1= It is the slope of the regression line, which is either increasing or decreasing.

ε = The error term. (For a good model it will be negligible)

Multiple Linear Regression

- Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and more than one independent variable.
- For MLR, the dependent or target variable(Y) must be the continuous/real, but the predictor or independent variable may be of continuous or categorical form.
- Each feature variable must model the linear relationship with the dependent variable.
- MLR tries to fit a regression line through a multidimensional space of data-points.
- **Example:** Prediction of CO_2 emission based on engine size and number of cylinders in a car.

MULTIPLE LINEAR REGRESSION

- **MLR equation:**
  - In Multiple Linear Regression, the target variable(Y) is a linear combination of multiple predictor variables x_1, x_2, x_3, ...,x_n.

Y= b<sub>0</sub>+b<sub>1</sub>x<sub>1</sub>+ b<sub>2</sub>x<sub>2</sub>+ b<sub>3</sub>x<sub>3</sub>+...... bnxn

where, Y= Output/Response variable, b_0, b_1, b_2, b_3 , b_n....= Coefficients of the model, x_1, x_2, x_3, x_4,...= Various Independent/feature variable

- **Assumptions for Multiple Linear Regression:**
  - A **linear relationship** should exist between the Target and predictor variables.
  - The regression residuals must be **normally distributed**.
  - MLR assumes little or **no multicollinearity** (correlation between the independent variable) in data.

EVALUATION METRICS FOR REGRESSION MODEL

- In regression problems, the prediction error is used to define the model performance. The prediction error is also referred to as residuals and it is defined as the difference between the actual and predicted values.
- Residuals are important when determining the quality of a model.
- **Residual = actual value — predicted value**

  $error(e) = y — \hat{y}$
- We can technically inspect all residuals to judge the model's accuracy, but this does not scale if we have thousands or millions of data points. That's why we have summary measurements that take our collection of residuals and condense them into a *single* value representing our model's predictive ability.

EVALUATION METRICS FOR REGRESSION MODEL - Mean Absolute Error (MAE)

- It is the average of the absolute differences between the actual value and the model's predicted value.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

where, N = total number of data points, Yi = actual value, Ŷi = predicted value.

- A small MAE suggests the model is great at prediction, while a large MAE suggests that your model may have trouble in certain areas. MAE of 0 means that your model is a perfect predictor of the outputs.
- **Advantages of MAE**: It is most Robust to outliers.
- **Disadvantages of MAE**: The graph of MAE is not differentiable so we have to apply various optimizers like Gradient descent which can be differentiable.

EVALUATION METRICS FOR REGRESSION MODEL – Mean Squared Error

- It is the average of the squared differences between the actual and the predicted values. Lower the value, the better the regression model.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

where, **n** = total number of data points, **yi** = actual value, **ŷi** = predicted value

- If you have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger.
- **Advantages of MSE** - The graph of MSE is differentiable, so you can easily use it as a loss function.
- **Disadvantages of MSE** - If you have outliers in the dataset then it penalizes the outliers most and the calculated MSE is bigger. So, in short, It is not Robust to outliers which were an advantage in MAE.

EVALUATION METRICS FOR REGRESSION MODEL – Route Mean Squared Error

- It is the average root-squared difference between the real value and the predicted value.
- lower the RMSE value, the better the model is with its predictions.
- A Higher RMSE indicates that there are large deviations between the predicted and actual value.

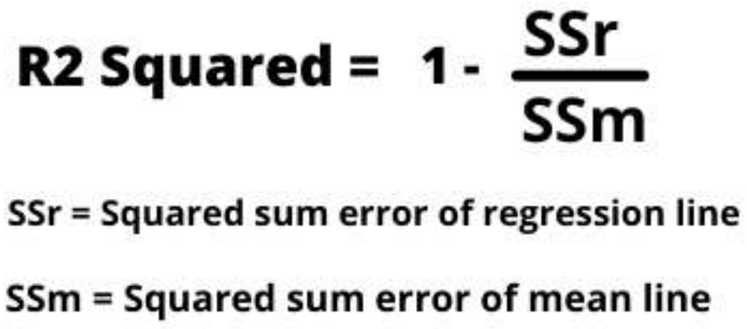
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

where, **n** = total number of data points, **yj** = actual value, **ŷj**= predicted value

- **Advantages of RMSE**: The output value is in the same unit as the required output variable which makes interpretation of loss easy.
- **Disadvantages of RMSE:** It is not that robust to outliers as compared to MAE.

EVALUATION METRICS FOR REGRESSION MODEL – R Squared

- R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.
- So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides.
- The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically R2 squared calculates how must regression line is better than a mean line.



$$\text{R2 Squared} = 1 - \frac{SSr}{SSm}$$

SSr = Squared sum error of regression line

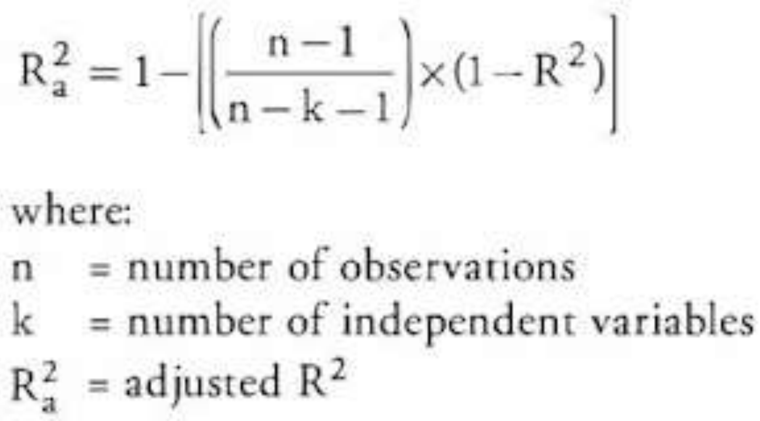
SSm = Squared sum error of mean line

EVALUATION METRICS FOR REGRESSION MODEL – R Squared

- Now, how will you interpret the R2 score? suppose If the R2 score is zero then the above regression line by mean line is equal means 1 so 1-1 is zero.
- So, in this case, both lines are overlapping means model performance is worst, It is not capable to take advantage of the output column.
- Now the second case is when the R2 score is 1, it means when the division term is zero and it will happen when the regression line does not make any mistake, it is perfect. In the real world, it is not possible.
- So we can conclude that as our regression line moves towards perfection, R2 score move towards one. And the model performance improves.
- The normal case is when the R2 score is between zero and one like 0.8 which means your model is capable to explain 80 per cent of the variance of data.

EVALUATION METRICS FOR REGRESSION MODEL – Adjusted R Squared

- The disadvantage of the R2 score is while adding new features in data the R2 score starts increasing or remains constant but it never decreases because It assumes that while adding more data variance of data increases.
- But the problem is when we add an irrelevant feature in the dataset then at that time R2 sometimes starts increasing which is incorrect.
- Hence, To control this situation Adjusted R Squared came into existence.



$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \times (1 - R^2) \right]$$

where:

n = number of observations

k = number of independent variables

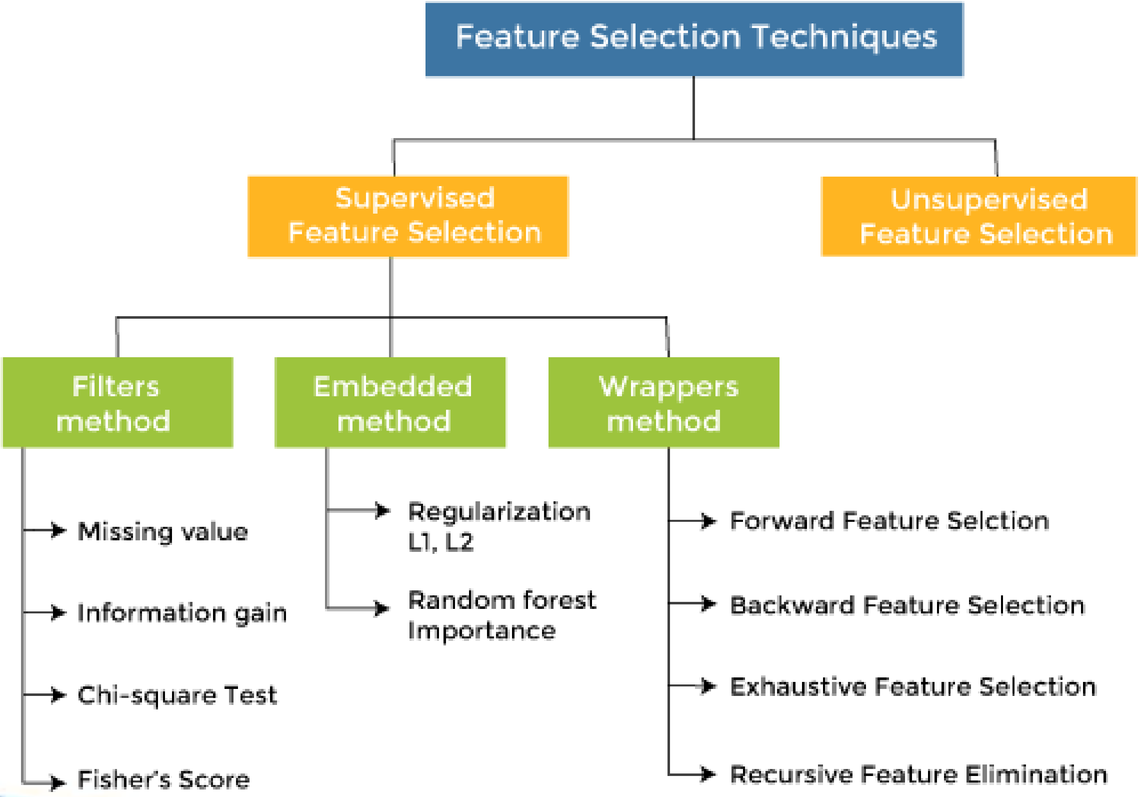
R_a^2 = adjusted R^2

EVALUATION METRICS FOR REGRESSION MODEL – Adjusted R Squared

- Now as K increases by adding some features so the denominator will decrease, n-1 will remain constant.
- R2 score will remain constant or will increase slightly so the complete answer will increase and when we subtract this from one then the resultant score will decrease.
- So, this is the case when we add an irrelevant feature in the dataset.
- And if we add a relevant feature then the R2 score will increase and 1-R2 will decrease heavily and the denominator will also decrease so the complete term decreases, and on subtracting from one the score increases.

Thank You

FEATURE SELECTION TECHNIQUES



FEATURE ENGINEERING TECHNIQUES FOR ML

- **Imputation**: Imputation is responsible for handling irregularities within the dataset.
- **Handling Outliers**: Standard deviation can be used to identify the outliers. Z-score can also be used to detect outliers.
- **Log Transform**: helps in handling the skewed data, and it makes the distribution more approximate to normal after transformation.
- **Binning**: used to normalize the noisy data.
- **Feature Split**: is the process of splitting features intimately into two or more parts and performing to make new features.
- **One hot encoding**: It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms and hence can make a good prediction.