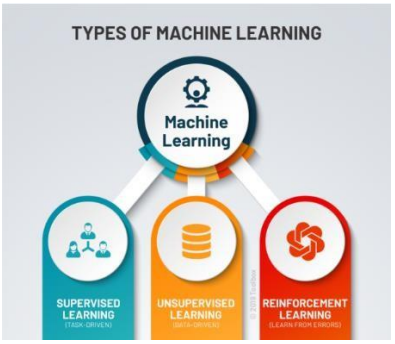
- **Machine Leaíning** is defined as the study of computeí píogíams that leveíage algoíithms and statistical models to leaín thíough infeíence and patteíns without being explicitly píogíammed.



Supervised Machine learning:

- Data is labeled and the algorithms learn to predict the output from the input data
- Learning stops when the algorithm achieves an acceptable level of performance
- It is classified as

    1. Regression
    2. Classification

Regression Algorithm	Classification Algorithm
In Regression, the output variable must be of continuous nature or real value.	In Classification, the output variable must be a discrete value.
The task of the regression algorithm is to map the input value (x) with the continuous output variable(y).	The task of the classification algorithm is to map the input value(x) with the discrete output variable(y).
Regression Algorithms are used with continuous data.	Classification Algorithms are used with discrete data.
In Regression, we try to find the best fit line, which can predict the output more accurately.	In Classification, we try to find the decision boundary, which can divide the dataset into different classes.
Regression algorithms can be used to solve the regression problems such as Weather Prediction, House price prediction, etc.	Classification Algorithms can be used to solve classification problems such as Identification of spam emails, Speech Recognition, Identification of cancer cells, etc.

ML Modelling Process Flow:

- **Get data:** Gather data from different sources
- **Clean, prepare and Manipulate data:**
    1. Handling null values and outliers
    2. Feature selection/ Variable Selection based on domain knowledge
    3. Converting Categorical data to Numerical
    4. Scaling the data
- **EDA:** Generating reports and Graphs
- **Splitting the Data:** Creating train and test data
- **Train Model:** Build a model using train data (usually 70% to 80% of whole data)
- **Evaluation/Test Model:** Test the model using the test data(20% to 30% of the whole data)
- **Model Tuning:** Optimize the model to increase its accuracy

Linear Regression

- **Linear Regression Steps:**

1. Create the dataframe properly--> pd.read_csv(), pd.read_excel()
2. Pre-processing the data:
    a. Feature selection-->domain knowledge-->drop()
    b. Handling the missing values-->isnull().sum(),fillna(),dropna()
    c. Converting the categorical data to numerical-->map()
3. Assumption 1: There should be no outliers in the data-->boxplot()
4. Assumption 2: Assumption of Linearity: Every ind var should have a linear relationship with the dep var-->pairplot(), drop()
5. Create X and Y--> X= ind vars, Y=dep var
6. Assumption 3: Assumption of Normality: The dependent variable should follow an approximate normal distribution-->distplot(),log()
7. Check and handle the skewness in the X vars-->hist(),skew(),log1p()
8. Assumption 4: Assumption of no multicollinearity: There should be no multicollinearity between the independent variables-->corr(),heatmap(),VIF(),drop()
9. Splitting the data into train and test(validation)-->train_test_split()
10. Building the model:
    a. Create the model-->obj=AlgoName()
    b. Train the model-->obj.fit(X_train,Y_train)
    c. Predict using the model-->Y_pred=obj.predict(X_test)
11. Evaluate the model:
    score, R-squared, Adj R-squared, RMSE, AIC/BIC
12. Assumption 5: There should be no auto-correlation in the data-->Durbin Watson test
13. Assumption 6: Errors should be random-->Residual v/s Fitted plot
14. Assumption 7: Errors should follow an approx normal distribution-->Normal QQ plot
15. Assumption 8: Errors should follow a constant variance(Homoskedasticity)-->Scale-Location plot
16. Tuning the model:
    a. Feature selection-->domain knowledge, p-values
    b. Regularization techniques-->Ridge(),Lasso()
    c. Stochastic Gradient Descent

Linear Regression Definition:

- Regression is a statistical technique which helps you to measure the relationship between the independent variables and dependent variables
- It helps you to understand one unit change in the independent variables is going to cause how many units change in the dependent variable
- Dependent or predicted variable is represented as 'y'

LR Approaches:

1. Statistical – Ordinary least Squares (OLS)
2. Machine Learning – Stochastic Gradient Descent (SGD)
    - **OLS is used for linear regression as it will return the best fit line which gives least errors**

Types of Linear Regression:

1. Simple Linear Regression: 1 X- variable and Y- Variable
   Example -predicting salary on the basis of experience

   Equation:

   $$Y = \beta 0 + \beta 1 \, . X$$

   were,

   $\beta 0 = Point\ on\ Y\ axis\ where\ best\ fit\ line\ intersect\ it,$
   $the\ value\ of\ Y\ when\ X\ is\ zero$

   Y = Dependent variable

   X = Independent variable

   $\beta 1 = Slope\ coefficient\ of\ X - variable$

   It indicates one unit change in X cause how many units change in Y.

Positive Slope Coefficient → X increases Y increases

Negative Slope Coefficient → X increases Y decreases

0 or close to 0 → No Relation

1. Multiple Linear Regression: Multiple X- variable and 1 Y- Variable
   Example: Predicting Salary on the basis of Department, age, domain etc.

   Equation:

   $$Y = \beta 0 + \beta 1 \, . X1 + \beta 2 . X2 + \cdots . \beta n Xn$$

   where,

   $\beta 0 = Point\ on\ Y\ axis\ where\ best\ fit\ line\ intersect\ it,$
   $the\ value\ of\ Y\ when\ X\ is\ zero$

   Y = Dependent variable

   X1, X2, Xn = Independent variables

   $\beta 1, \beta 2, \beta n = Slope\ coefficients\ of\ X - variables$

   It indicates one unit change in X cause how many units change in Y.

Positive Slope Coefficient → X increases Y increases

Negative Slope Coefficient → X increases Y decreases

0 or close to 0 → No Relation

Multicollinearity

Multicollinearity is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model.

How to Deal with Multicollinearity

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.

No Auto-Correlation:

Auto-Correlation measures the relationship between a variable's current value and its past value, mean errors are assumed to be uncorrelated i.e. randomly spread around the regression line

No patterns in Graph indicates no Auto - Correlation

Durbin Watson Test for Auto-Correlation:

It takes a value between 0 – 4

- If value is close to 0 → No Auto-correlation
- If value is close to 2 → Positive Auto-correlation
- If value is close to 4 → Negative Auto-correlation

<u>Decomposition Variability/ Evaluation matrices:</u>

- It is a determinant for a good regression
- It includes
  1. Sum of Squares total
  2. Sum of Squares regression
  3. Sum of Squares Error
  4. R – Squared
  5. Adjusted R – Squared
  6. AIC & BIC
  7. RMSE

Sum of Squares total (SST):

The square of the difference between the observed dependent variable and its mean

$$SST = \sum (y - \hat{y})^2$$

Sum of Squares regression (SSR):

The sum of the difference between the predicted value and the mean of dependent variable

$$SSR = \sum (y' - \hat{y}')^2$$

Sum of Squares Error (SSE):

The sum of the difference between the observed value and predicted value of the dependent variable

$$SSE = \sum (y - y')^2$$

Relation between SST, SSR, and SSE:

SST = SSR + SSE

Total Variability = Explained Variability + Unexplained Variability

AIC & BIC:

- It is used to compare different models with same algorithms but with different number of Independent Variables

1. **AIC**
   - Akaike Information criteria to penalize the inclusion of additional variables to the model
   - It adds penalty to those variables that increases the error
   - The model with lowest AIC is selected
   - The penalty of AIC is less compared to BIC, causing AIC to pick More complex models
2. **BIC**
   - Bayesian Information Criterion is a variant of AIC with a stronger penalty for including additional variables to the model
   - It adds penalty to those variables that increases the error
   - The model with lowest BIC is selected
   - As compared to AIC, it penalizes model complexity more heavily

RMSE:

- Root Mean Square error is an absolute measure of the goodness for the fit
- It gives an absolute number on how much your predicted results deviate from the actual number
- Low the RMSE better the model

R Squared:

- It tells you how well the regression model is predicting as compared to the mean model
- Lies between (0-1)
- If R squared is close to 1 → very good model
- If R squared is close to 0.5 → Needs tuning
- If R squared is close to 0 → Not a good model
- If R squared is less than 0 → Mean model is better than the regression model

Adjusted – R Squared:

- Penalized R Squared Value
- It will always be lower than R Squared
- As more data is added R Squared goes on increasing whereas Adj -R Squared will increase only when significant data is added to the model hence it is more reliable to quote the adjusted R Squared to the client