MACHINE LEARNING USING PYTHON

INNOCENT CHARLES
MLOps Engineer

The greatest secret is attitude and passion.

UDOM AI Community

CONTENTS TO BE COVERED

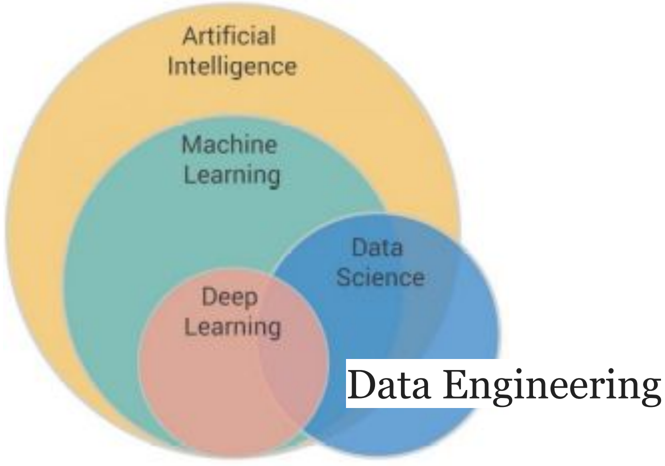
- Introduction
- Machine Learning
- Types of Machine Learning
- ML workflows
- ML tools Mostly Used in ML
- Application of AI
- Supervised Machine Learning
- Regression
- Finding the Best Fit Line
- Hands on Data (practical implementation)

INTRODUCTION

What is the difference between ML , AI , DL , DE and DS ?

Raise your hand if you’ve been caught in the confusion of differentiating artificial intelligence (AI) vs machine learning (ML) vs deep learning (DL) vs Data Science(DS) vs Data Engineering (DE)…

Bring down your hand, buddy, we can’t see it!



INTRODUCTION

Artificial Intelligence:

- Concerned with building smart machines and intelligent system capable of performing tasks that typically require human intelligence.

Machine Learning:

- Is a branch of AI based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

Deep Learning:

- Is a subset of ML that use a cascade of multiple layers of non-linear processing and artificial neural networks to deliver high accuracy for pattern recognition and feature learning , it stimulates the neurons of the human brain .

Data Science:

- Is the science of extracting useful knowledge and insights from data and apply that knowledge and actionable insights on different application domain.

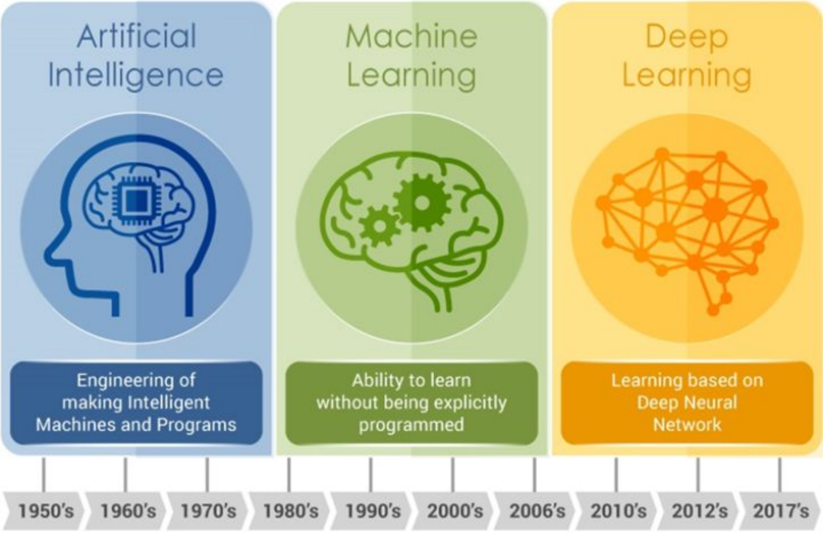
Data Engineering:

-the practice designing and building systems for collecting, storing, and analyzing data at scale and make those raw data to be used by data scientists or organization

INTRODUCTION

To make it more simple is :

1. Data engineering produces **raw data**
2. Data science produces **insights**
3. Machine learning produces **predictions**
4. Artificial intelligence produces **Actions**


Artificial Intelligence
Engineering of making Intelligent Machines and Programs
Machine Learning
Ability to learn without being explicitly programmed
Deep Learning
Learning based on Deep Neural Network
1950's
1960's
1970's
1980's
1990's
2000's
2006's
2010's
2012's
2017's

CASE STUDY ON AI ,ML AND DS

Suppose we were building a self-driving car, and were working on the specific problem of stopping at stop signs. We would need skills drawn from all three of these fields.

- **Machine learning**: The car has to recognize a stop sign using its cameras. We construct a dataset of millions of photos of streetside objects, and train an algorithm to **predict** which have stop signs in them.
- **Artificial intelligence**: Once our car can recognize stop signs, it needs to decide when to take the **action** of applying the brakes. It's dangerous to apply them too early or too late, and we need it to handle varying road conditions (for example, to recognize on a slippery road that it's not slowing down quickly enough), which is a problem of control theory.
- **Data science**: In street tests we find that the car's performance isn't good enough, with some false negatives in which it drives right by a stop sign. After analyzing the street test data, we gain the **insight** that the rate of false negatives depends on the time of day: it's more likely to miss a stop sign before sunrise or after sunset. We realize that most of our training data included only objects in full daylight, so we construct a better dataset including nighttime images and go back to the machine learning step.

MACHINE LEARNING BY USING SCIKIT-LEARN AND PYTHON

MACHINE LEARNING

What exactly Machine Machine Learning is ?

What does it mean to learn?

In Machine Learning an important concept is Generalization, the ability to **generalize.**

MACHINE LEARNING

What exactly Machine Machine Learning is ?

A computer program is said to learn from **experience E** with respect to some **task T** and some **performance P,** if its performance on T, as measured by P, improves with experience E.

- Tom Mitchell, 1997.

MACHINE LEARNING

Generally , **Machine Learning** can be defined as a subset of **artificial intelligence** that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own to identify the pattern and making prediction **without being programmed explicit.**

Example :

”You need to predict how much user “A” will like a movie that she hasn’t seen based on her ratings of movies that she has seen.”

Basically;There are two ways to solve these kind of problem;

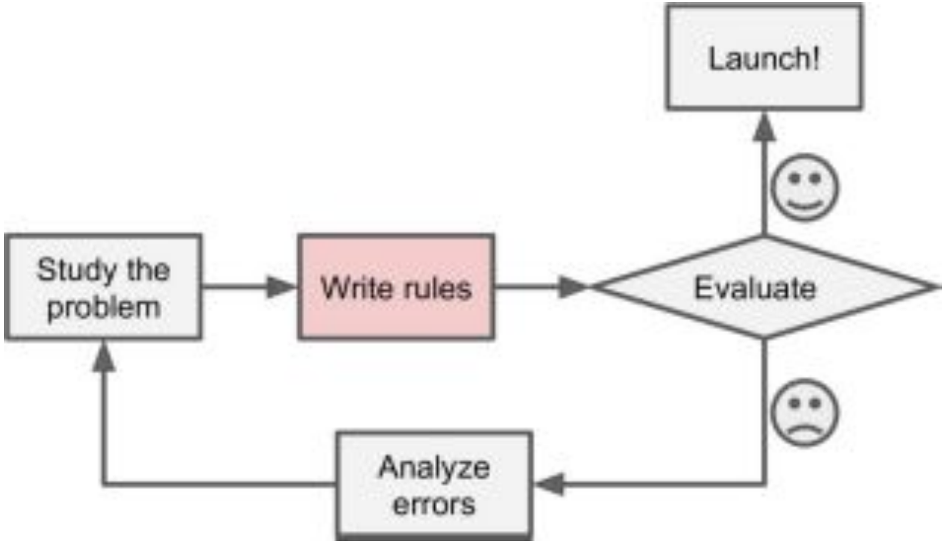
A. Traditional Programming Methods
B. Machine Learning Methods .

MACHINE LEARNING

A.Traditional Programming Methods:

-Complex

-Hard to Maintain

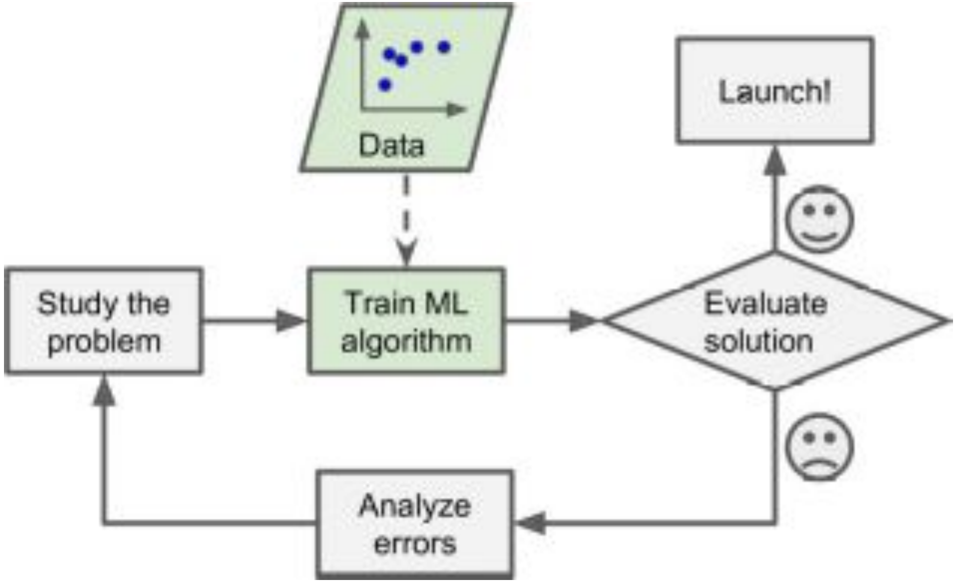


MACHINE LEARNING

A.Machine Learning Methods:

-Automatic pattern learning **- Ease to maintain** **-Adoption to changes**

-More accurate



TYPES OF THE MACHINE LEARNING

Generally : There are about Four type of the Machine Learning

1. **Supervised Learning**

2. **Unsupervised Learning**

3. **Semi-supervised Learning**

4. **Reinforcement Learning**

SUPERVISED LEARNING

Training data includes the desired solutions called labels.

SOME SUPERVISED LEARNING ALGORITHMS :

-Linear and logistic regression

-Support Vector Machines

-Decision trees

-Random forest

-Naive Bayes

-K-nearest Neighbours

-Neural Networks

UNSUPERVISED LEARNING

Training data DOES NOT includes the desired solutions called labels. Hence a machine has to find some structure in the dataset.They only extracts pattern from the provided data during learning.

SOME UNSUPERVISED LEARNING ALGORITHMS :

-k-means and hierarchical algorithms

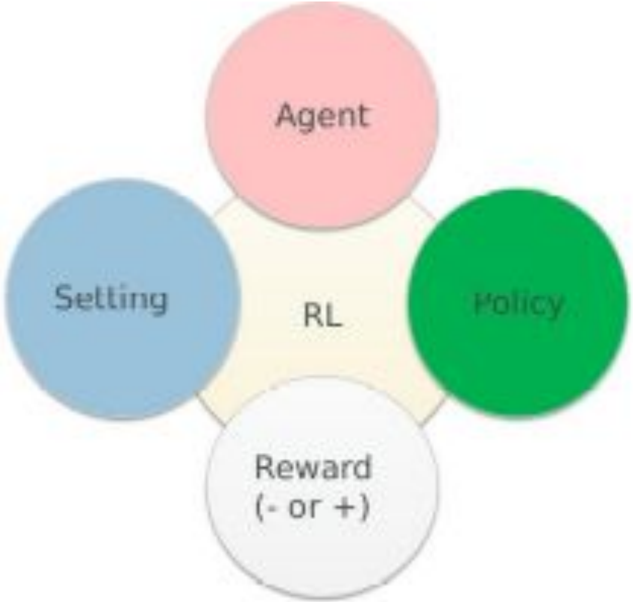
-Principal component Analysis and factor analysis , independent component Analysis , linear discriminant analysis

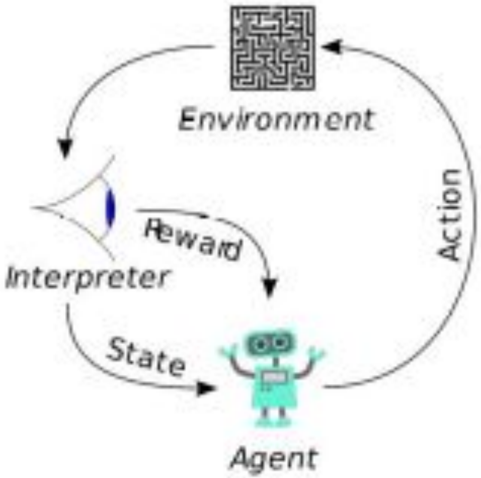
-Apriori Algorithm

-Neural Networks

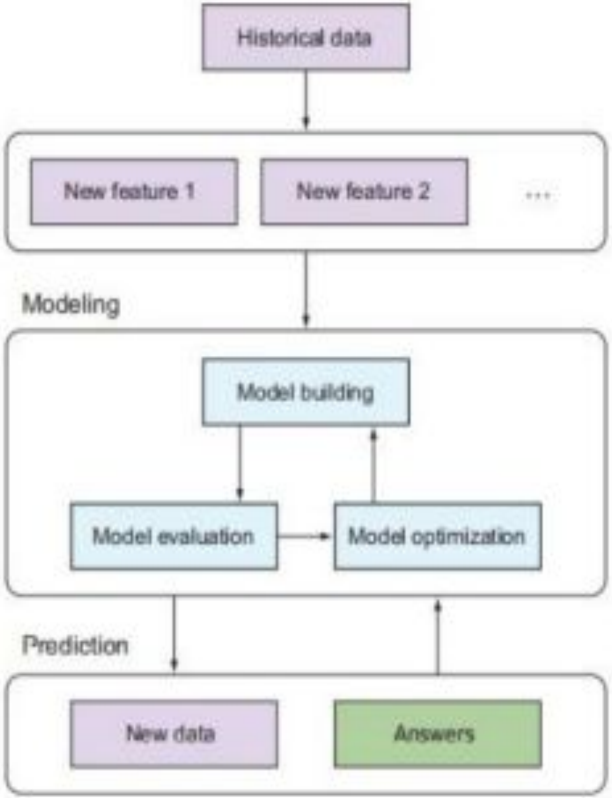
REINFORCEMENT LEARNING

Machine learning type that an agent is able to perceive and interpret its environment ,take actions ,learn through trial and error to make the right decision. Example games , Autonomous self driving car . **The agent continues doing these three things (take action,change state/remain in the same state,get feedback)**

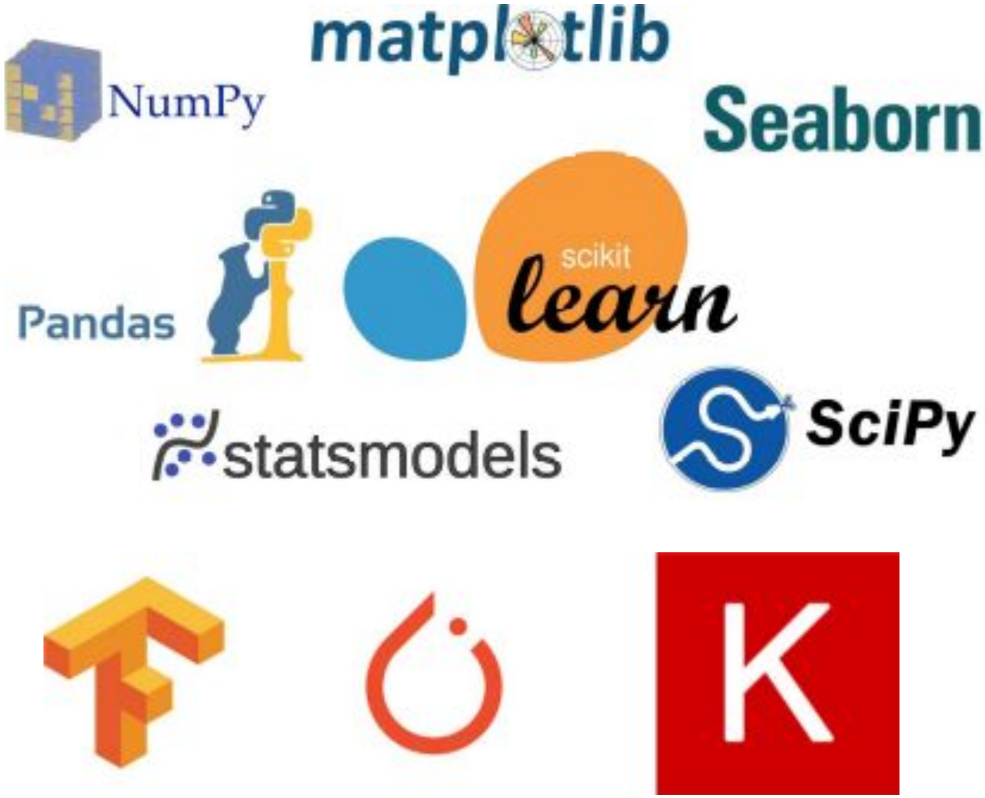




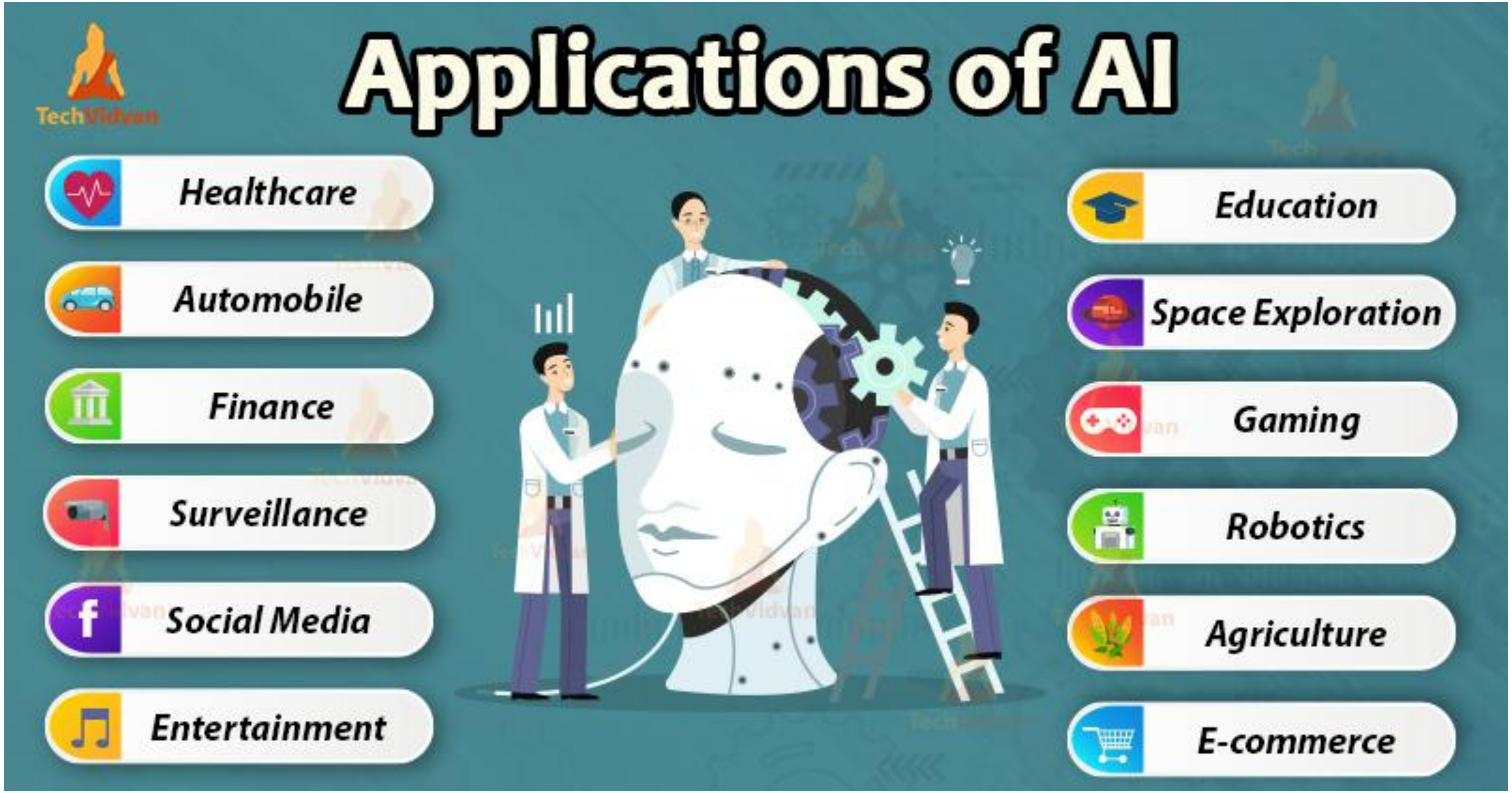
ML WORKFLOW


Historical data
New feature 1
New feature 2
...
Modeling
Model building
Model evaluation
Model optimization
Prediction
New data
Answers

ML TOOLS MOSTLY USED IN ML



APPLICATION OF ARTIFICIAL INTELLIGENCE



WHERE
DO I
START?

Linear Algebra
Statistics
Probability Theory
Calculus
Programming Languages

```
model = lm(y ~ x, train_data)

predictions = predict(model,
                      test_data)
```

ROADMAP FOR AI

A COMPLETE ROADMAP FOR AI Expert 2021-2022

SUPERVISED MACHINE LEARNING

SUPERVISED MACHINE LEARNING

is a subcategory of **machine learning** and **artificial intelligence**. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process.

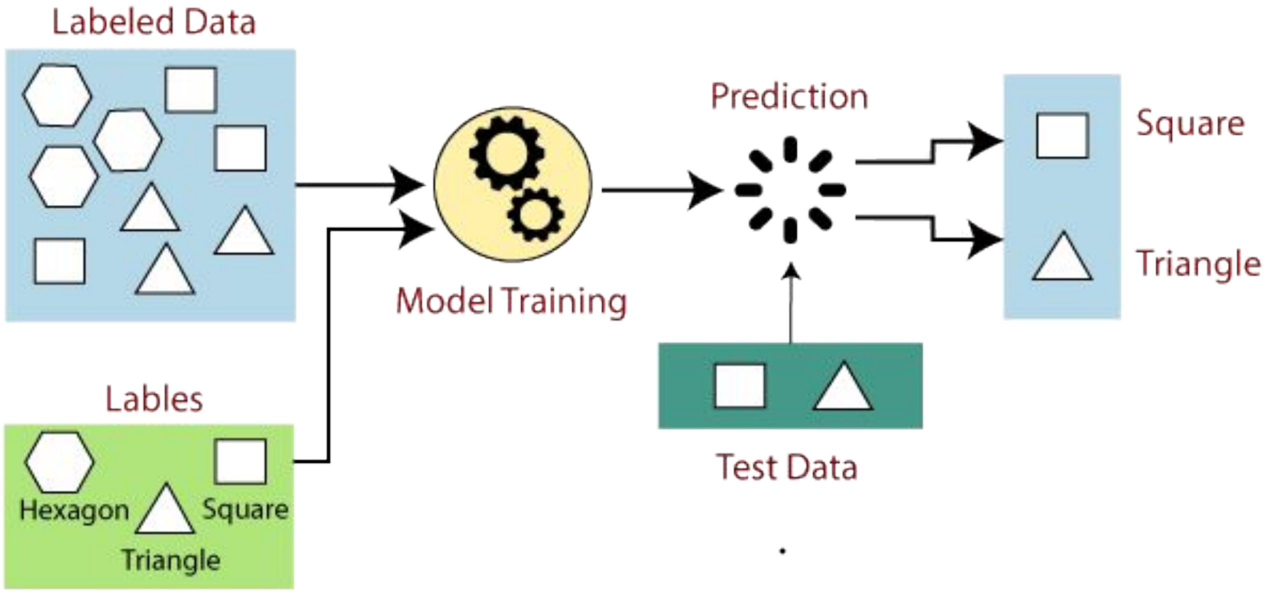
In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Supervised learning is a process of providing input data as well as correct output data to the machine learning model. The aim of a supervised learning algorithm is to **find a mapping function to map the input variable(x) with the output variable(y)**.

Supervised learning helps organizations solve for a variety of real-world problems at scale such as **Risk Assessment, Image classification, Fraud Detection, spam filtering**, etc.

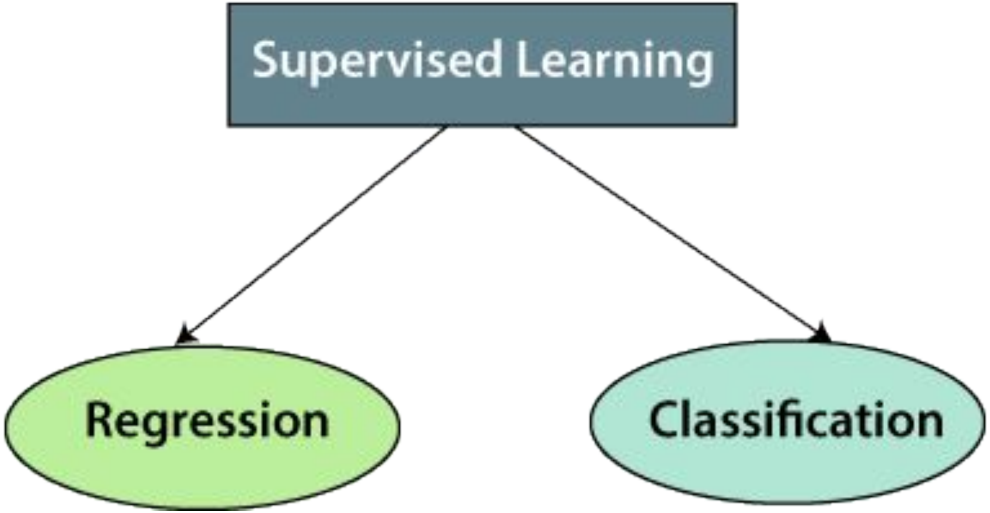
HOW SUPERVISED LEARNING WORK ?

Supervised learning uses a training set to teach models to yield the **desired output**. This training dataset includes **inputs** and **correct outputs,** which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.



TYPE OF THE SUPERVISED LEARNING

Basically ; Supervised Machine Learning is classified into Two groups:



TYPE OF THE SUPERVISED LEARNING

- **Classification** uses an algorithm to accurately assign test data into specific categories. It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined.It based on the discrete datasets.Example to classify yes or No , fraud or not Fraud . **Common classification algorithms are linear classifiers Example logistic classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, naive bayes .**

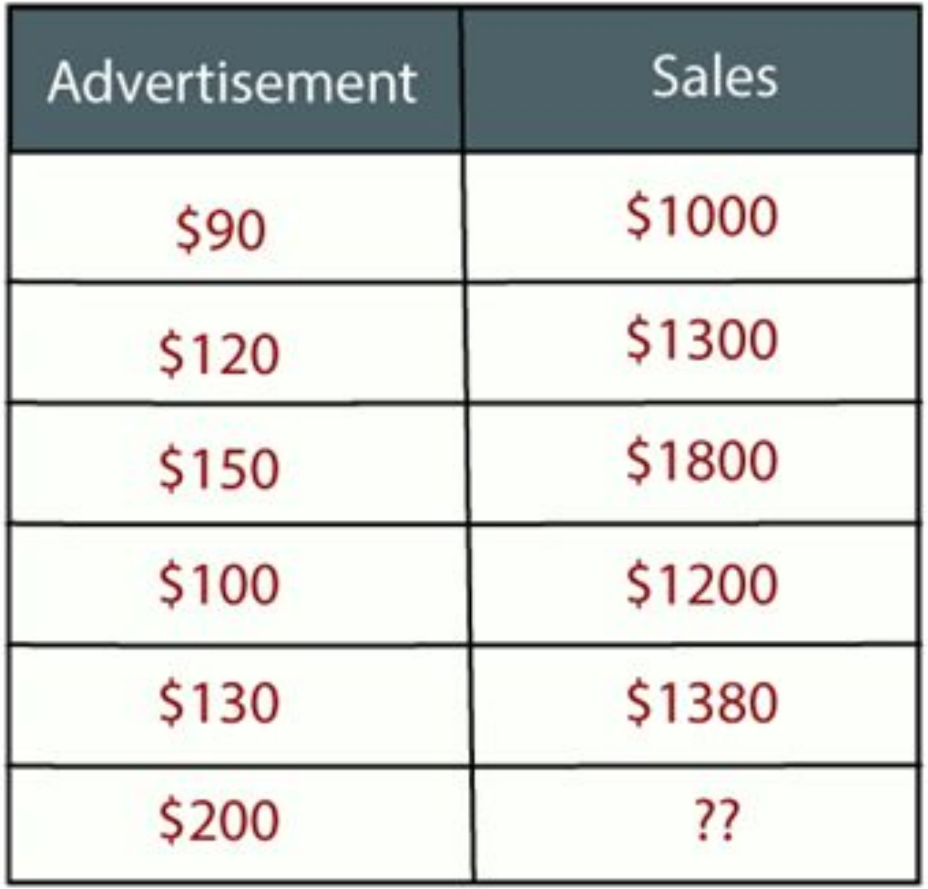
- **Regression** is used to understand the relationship between dependent and independent variables. It is commonly used to make projections and the prediction of continuous variables, such as Weather forecasting, Market Trends, sales revenue for a given business. Popular algorithms under this are **linear regression,random forest,support vector regression,decision trees,neural networks,lasso and ridge regression**

REGRESSION

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed. It predicts continuous/real values such as **temperature, age, salary, price,** etc.We can understand the concept of regression analysis using the below example:

Example: Suppose there is a marketing company A, who does various advertisement every year and get sales on that. The below list shows the advertisement made by the company in the last 5 years and the corresponding sales:

REGRESSION



Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??

REGRESSION

Now, the company wants to do the advertisement of $200 in the year 2019 **and wants to know the prediction about the sales for this year**. So to solve such type of prediction problems in machine learning, we need r**egression analysis**.

Regression is a supervised learning technique

which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

In Regression, we plot a graph between the variables which best fits the given data points, using this plot, the machine learning model can make predictions about the data. In simple words, **"Regression shows a line or curve that passes through all the data points on target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum."** The distance between data points and line tells whether a model has captured a strong relationship or not.

TERMINOLOGY RELATED TO SUPERVISED LEARNING

- **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.

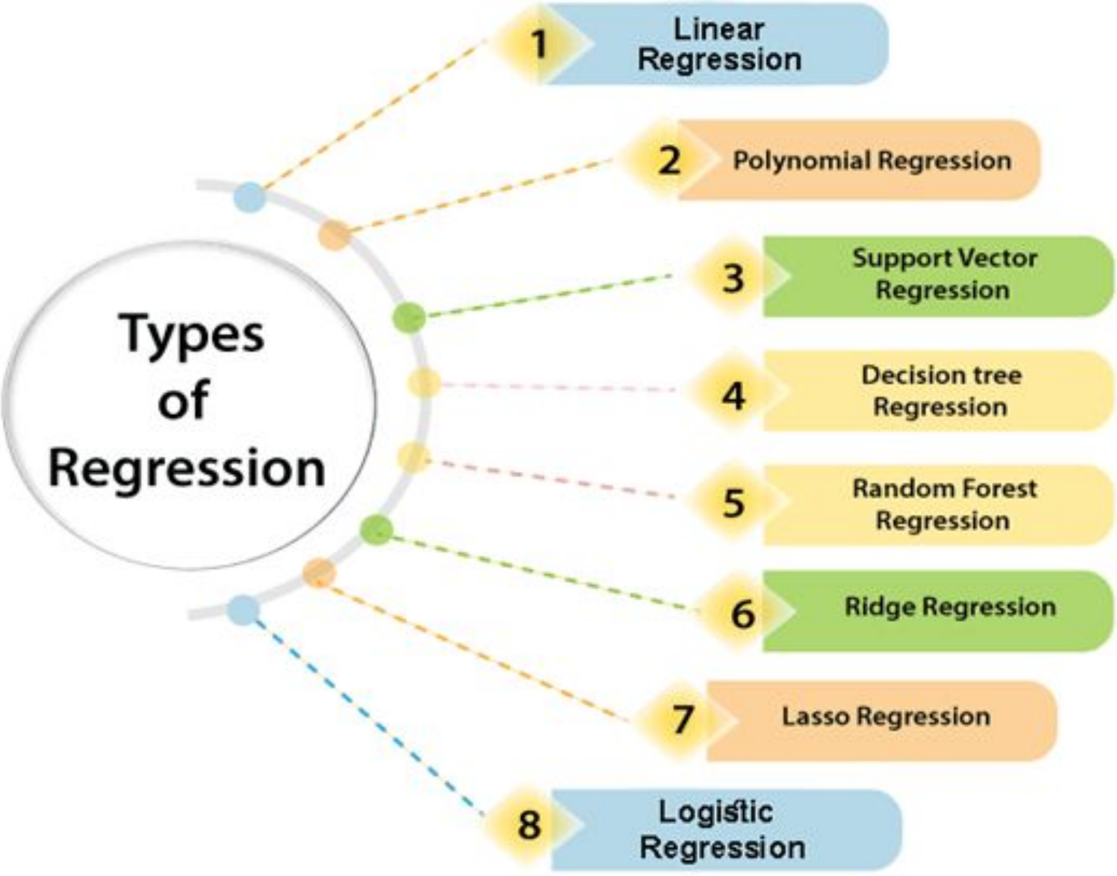
- **Independent Variable**: The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.

- **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. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.

- **Underfitting and Overfitting**: If our algorithm works well with the training dataset but not well with test dataset, then such problem is called **Overfitting**. And if our algorithm does not perform well even with training dataset, then such problem is **underfitting.**

TYPE OF THE REGRESSION ANALYSIS



LINEAR REGRESSION

- **Linear regression is** a statistical regression method which is used for predictive analysis.
- It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
- It is used for solving the regression problem in machine learning.
- Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.
- If there is only one input variable (x), then such linear regression is called **simple linear regression**. And if there is more than one input variable, then such linear regression is called **multiple linear regression**.
- The relationship between variables in the linear regression model can be explained using the below image. Here we are predicting the salary of an employee on the basis of **the year of experience**.

LINEAR REGRESSION

1. **Y= aX+b**

   y=dependent

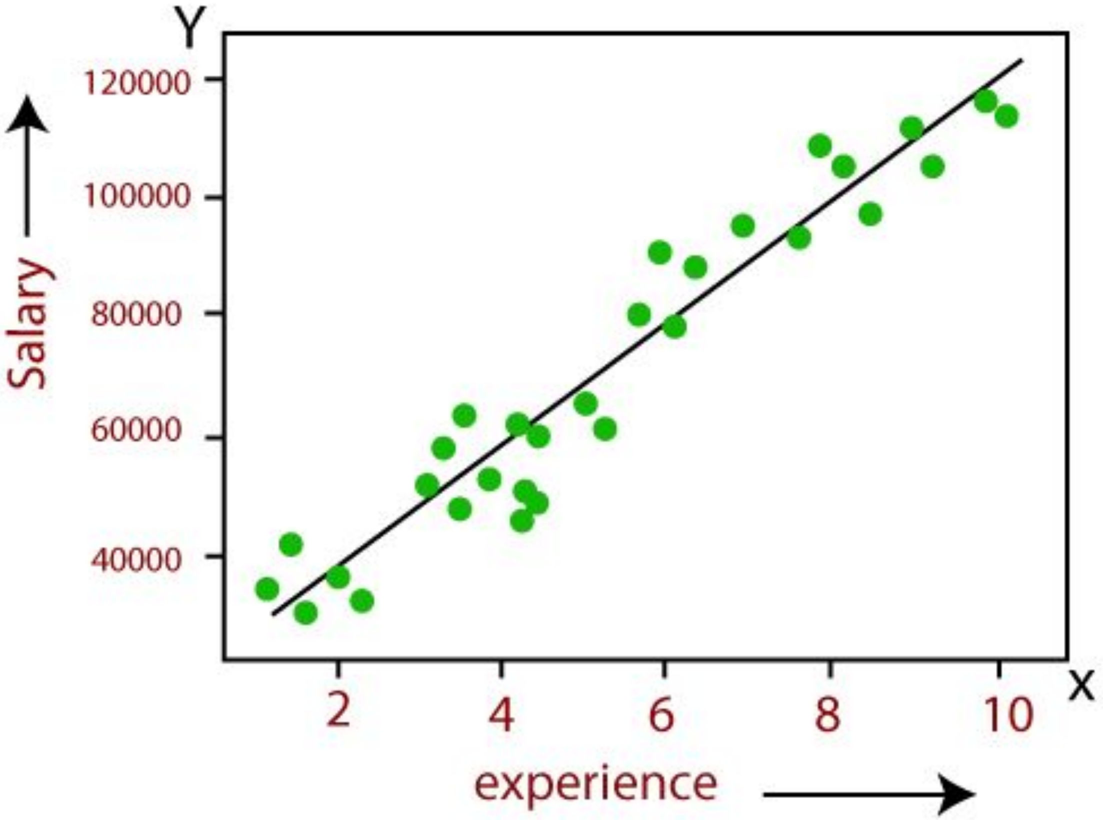
   (target)

   x=independent

   (predictors)

   a,b are

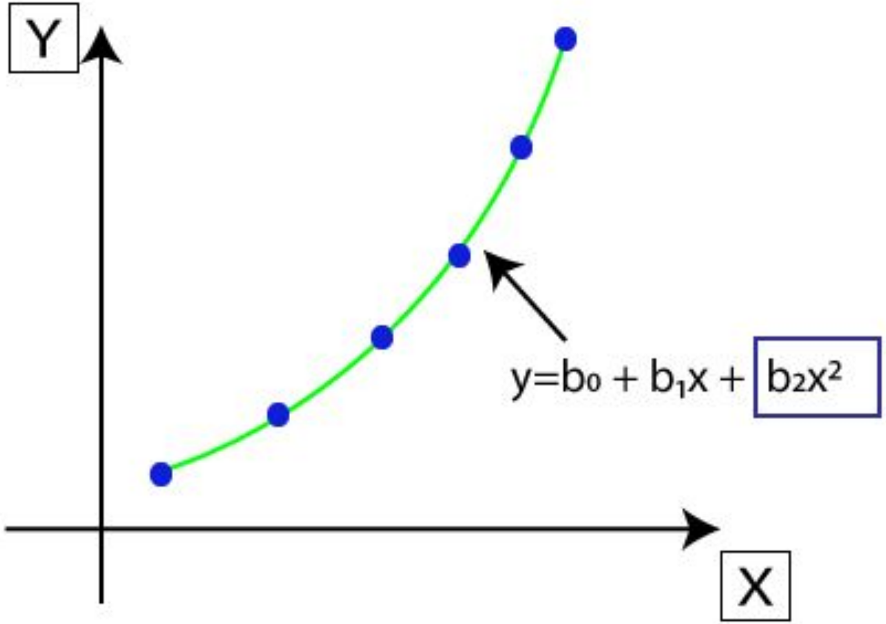
   parameters.



POLYNOMIAL REGRESSION

- Polynomial Regression is a type of regression which models the **non-linear dataset** using a linear model.
- It is similar to multiple linear regression, but it fits a non-linear curve between the value of x and corresponding conditional values of y.
- Suppose there is a dataset which consists of datapoints which are present in a non-linear fashion, so for such case, linear regression will not best fit to those datapoints. To cover such data points, we need Polynomial regression.
- **In Polynomial regression, the original features are transformed into polynomial features of given degree and then modeled using a linear model.** Which means the data points are best fitted using a polynomial line.

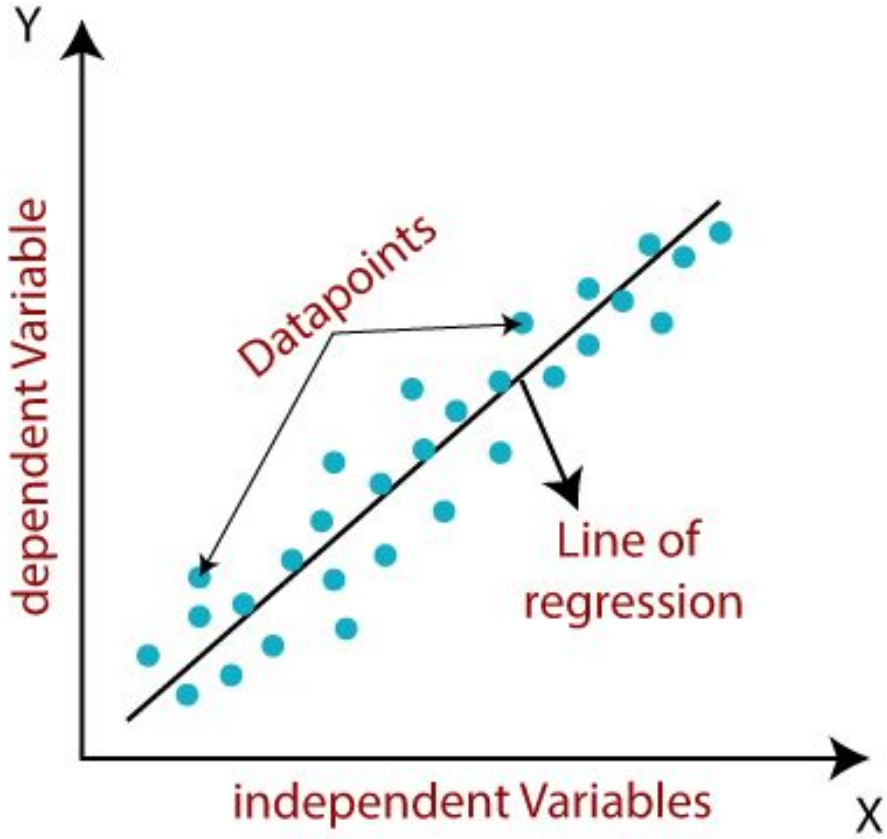
POLYNOMIAL REGRESSION



POLYNOMIAL REGRESSION

- The equation for polynomial regression also derived from linear regression equation that means Linear regression equation $Y= b_0+ b_1x$, is transformed into Polynomial regression equation $Y= b_0+b_1x+ b_2x^2+ b_3x^3+.....+ b_nx^n$.
- Here Y is the **predicted/target output, b_0, b_1,... b_n are the regression coefficients**. x is our **independent/input variable**.
- The model is still linear as the coefficients are still linear with quadratic

FINDING THE BEST FIT LINE



FINDING THE BEST FIT LINE

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

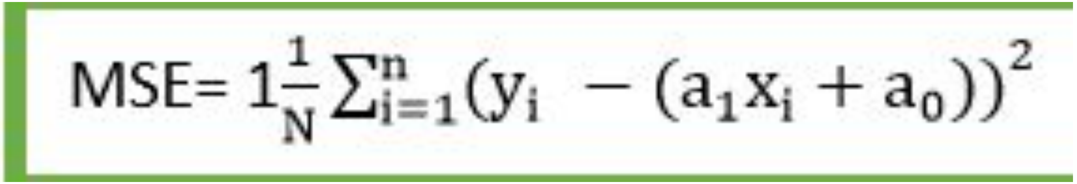
Cost function-

- The different values for weights or coefficient of lines (a_0, a_1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
- Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
- We can use the cost function to find the accuracy of the **mapping function**, which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values. It can be written as:

FINDING THE BEST FIT LINE

For the above linear equation, MSE can be calculated as:



$$\mathrm{MSE}= 1\frac{1}{N}\sum_{i=1}^{n}\left(y_i - \left(a_1 x_i + a_0\right)\right)^2$$

Where,

N=Total number of observation

Yi = Actual value

$(a1x_i+a_0)$= Predicted value.

Residuals: The distance between the actual value and predicted values is called residual. If the observed points are far from the regression line, then the residual will be high, and so cost function will high. If the scatter points are close to the regression line, then the residual will be small and hence the cost function

FINDING THE BEST FIT LINE

Gradient Descent:

- Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.
- A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.
- It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

Model Performance:

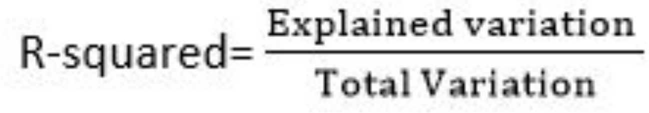
The Goodness of fit determines how the line of regression fits the set of observations. The process of finding the best model out of various models is called **optimization**. It can be achieved by below method:

FINDING THE BEST FIT LINE

R-squared

is a statistical method that determines the goodness of fit.

- It measures the strength of the relationship between the dependent and independent variables on a scale of 0-100%.
- The high value of R-square determines the less difference between the predicted values and actual values and hence represents a good model.
- It is also called a **coefficient of determination,** or **coefficient of multiple determination** for multiple regression.
- It can be calculated from the below formula:



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

ASSUMPTIONS OF LINEAR REGRESSION

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Linear relationship between the features and target:**
  Linear regression assumes the linear relationship between the dependent and independent variables.
- **Small or no multicollinearity between the features:**
  Multicollinearity means high-correlation between the independent variables. Due to multicollinearity, it may difficult to find the true relationship between the predictors and target variables. Or we can say, it is difficult to determine which predictor variable is affecting the target variable and which is not. So, the model assumes either little or no multicollinearity between the features or independent variables.
- **Homoscedasticity Assumption:**
  Homoscedasticity is a situation when the error term is the same for all the values of independent variables. With homoscedasticity, there should be no clear pattern distribution of data in the scatter plot.

ASSUMPTIONS OF LINEAR REGRESSION

Below are some important assumptions of Linear Regression. These are some formal checks while building a Linear Regression model, which ensures to get the best possible result from the given dataset.

- **Normal distribution of error terms:**
  Linear regression assumes that the error term should follow the normal distribution pattern. If error terms are not normally distributed, then confidence intervals will become either too wide or too narrow, which may cause difficulties in finding coefficients.
  It can be checked using the **q-q plot**. If the plot shows a straight line without any deviation, which means the error is normally distributed.
- **No autocorrelations:**
  The linear regression model assumes no autocorrelation in error terms. If there will be any correlation in the error term, then it will drastically reduce the accuracy of the model. Autocorrelation usually occurs if there is a dependency between residual errors.

MOSTLY TYPE OF DATASET TO WORK ON IN AI

Below are the mostly types of the dataset to work on when doing Artificial intelligence.

- Numbers
- Categoricals
- Texts
- Audio
- Videos
- Images

GENERAL STEPS TO SOLVE MOSTLY MACHINE LEARNING PROBLEMS

Below are the mostly STEPS to solve a machine learning problems:

- ➔ **Gathering data .** Example Scraping using selenium , beautifully soap , scrapy
- ➔ **Performing EDA** . Example Exploratory and Analysis of the Data
- ➔ **Preprocessing of the Data** . Example feature engineering , data imputation , data cleaning and data transformation .
- ➔ **Cross validation** . Example Hold-out ,k-fold,stratified k-fold,repeated stratified k-fold.
- ➔ **Feature scaling (optional)** due to depend on some algorithms and normalization of the data.
- ➔ **Data modelling**
- ➔ **Model training**
- ➔ **Predictions.**
- ➔ **Model optimizations and regulation** . Example Ensembling techniques such as stacking ,bagging , boosting , hyperparameters tuning also using searching algorithms like grid and random search .

BOOKS AND USEFULL LINKS

1. AI LEARNING MATERIALS
2. Krish Naiki Youtube
3. Total Data science
4. https://www.youtube.com/user/pantechsolutions

IMPLEMENTATION OF LINEAR REGRESSION

THANKS !!!! …IT IS TIME FOR HANDS ON A LINEAR REGRESSION SUPERVISED MACHINE LEARNING