Machine Learning

Practical File

July 20, 2022

BACHELOR OF TECHNOLOGY

Information Technology

SUBMITTED BY

MANPREET KAUR

University Roll no. 2104529

Class Roll no. 2121069

GURU NANAK DEV ENGINEERING COLLEGE

LUDHIANA-141006, INDIA

Contents

1 INTRODUCTION TO PYTHON LIBRARIES **1**

1.1 NumPy . . . 1

1.2 Pandas . . . 1

1.3 Matplotlib . . . 2

2 STUDY THE WORKNING FOR ANY TOOL FOR ARTIFICAL INTELLIGENCE APPLICATION **3**

2.1 INTRODUCTION TO DATA SCIENCE . . . 3

2.2 COMPONENTS OF DATA SCIENCE . . . 4

2.2.1 Data . . . 4

2.2.2 Big Data . . . 4

2.2.3 MACHINE LEARNING . . . 6

2.2.4 Statistics and Probability . . . 6

3 APPLICATION OF DATA SCIENCE **6**

4 INTRODUCTION TO ARTIFICAL INTELLIGENT **8**

5 APPLICATION OF ARTIFICAL INTELLIGENT **9**

6 DEFINATION OF MACHINE LEARNING **10**

7 TRADITIONAL PROGRAMMING **10**

8 MACHINE LEARNING PROGRAMMING **10**

9 SUPERVISED &UNSUPERVISED LEARNING **10**

10 Advantages of Artificial Intelligence **13**

11 REGRESSION TECHNIQUE IN MACHINE LEARNING **15**

11.1 Simple linear regression . . . 15

11.2 Multiple linear regression . . . 16

11.3 Polynomial regressionSVR . . . 18

11.4 SVR AND SVM . . . 19

12 Decision tree **20**

13 Random forest **22**

14 CLASSIFICATION TECHNIQUES IN MACHINE LEARNING **23**

14.1 K-Nearest Neighbour (KNN) . . . 23

14.2 Support Vector Machine (SVM) . . . 24

14.3 Confusion matrix . . . 25

15 Clustering Technique in machine learning **25**

15.1 K Means Clustering: . . . 25

1 INTRODUCTION TO PYTHON LIBRARIES

1. NumPy

2. Pandas

3. Matplotlib

1.1 NumPy

NumPy is a Python library used for working with arrays.

It also has functions for working in domain of linear algebra, fourier transform, and matrices.

NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.

NumPy stands for Numerical Python.

CODE

Input

```
import numpy as np
array=[23,34,45,56,67]
np.array=[array]
print(array)
```

Output

```
[23,34,45,56,67]
```

1.2 Pandas

Pandas is an open source Python package that is most widely used for data science/data analysis and machinelearning tasks. It is built on top of another package named Numpy, which provides support for multidimensionalarrays. As one of the most popular data wrangling packages, Pandas works well with th many other data science modules inside the Python ecosystem, and is typically included in every Python distribution, from those that come with your operating system to commercial vendor distributions like ActiveState's ActivePython.

CODE

Input

```
import pandas as pd
dic1={
‘‘name’’:[’ram’,’sham’,’kam’],
‘‘city’’:[’haryana’,’delhi’,’punjab’],
‘‘marks’’:[23,34,45]
}
df=pd.DataFrame[dic1]
print(df)
```

Output

1.3 Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

CODE

Input

```
import pyplot as plt
x = [5, 2, 9, 4, 7]
y = [10, 5, 8, 4, 2]
plt.bar(x,y)
plt.show()
```

Output

dJD

2 STUDY THE WORKNING FOR ANY TOOL FOR ARTIFICAL INTELLIGENCE APPLICATION

2.1 INTRODUCTION TO DATA SCIENCE

Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.

Data Science is about data gathering, analysis and decision-making.

Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

. Better decisions (should we choose A or B)

.Predictive analysis (what will happen next?)

.Pattern discoveries (find pattern, or maybe hidden information in the data)

2.2 COMPONENTS OF DATA SCIENCE

1. DATA

2. BIG DATA

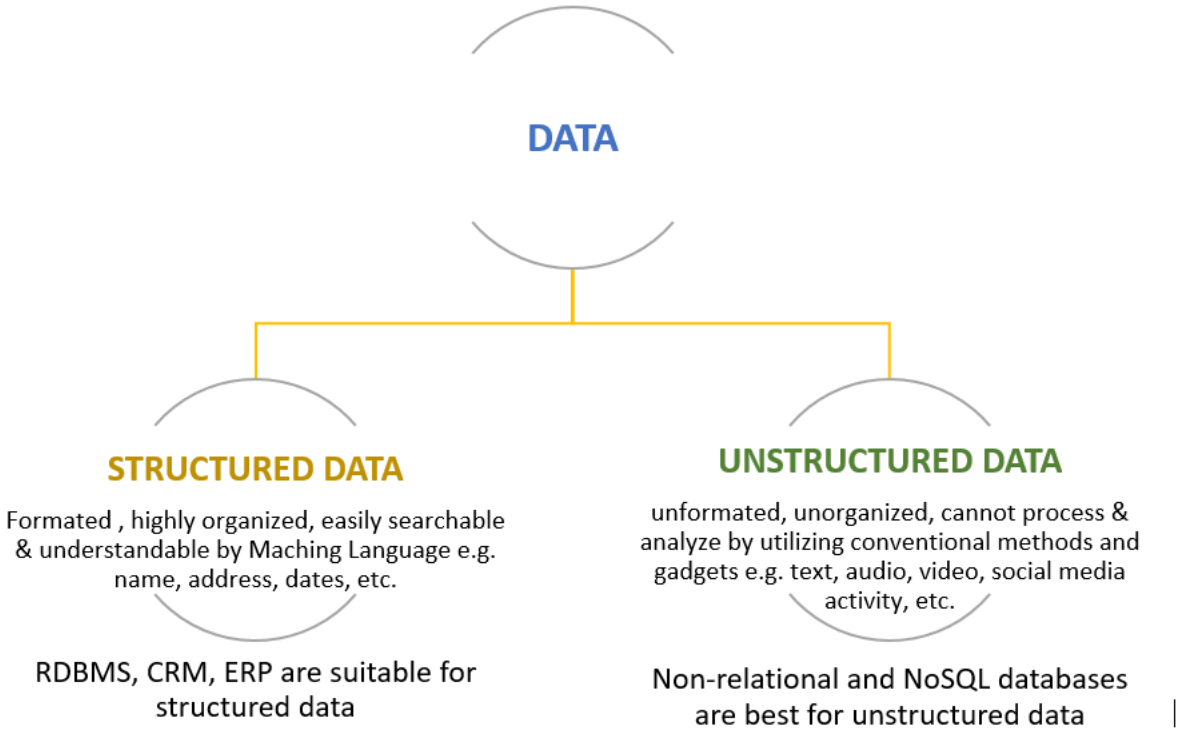
3. MACHINE LEARNING

4. STATICS

2.2.1 Data

Data is a collection of f1actual information based on numbers, words, observations, measurements which can be utilized for calculation, discussion and reasoning.

The crude dataset is the basic foundation of data science and it may be of different kinds like Structured Data (Tabular structure), Unstructured Data (pictures, recordings, messages, PDF documents and so forth.) and Semi Structured.



2.2.2 Big Data

Big Data is enormously big data sets. It consists of various V's such as, volume, variety, velocit, vision, value, variability & visualization, etc. For instance, Facebook.

big data image big data image Data is contrasted and raw petroleum which is a profitable crude material, and as scientist separate the refined oil from the unrefined petroleum comparably by applying data science, scientist can remove various types of data from crude information.

2.2.3 MACHINE LEARNING

Machine Learning is the part of Data Science which enables the system to process datasets autonomously without any human interference by utilizing various algorithms to work on massive volume of data generated and extracted from numerous sources.

It makes prediction, analysis patterns and gives recommendations. Machine learning is frequently being used in fraud detection and client retention.

A social media platform i.e. Facebook is a decent example of machine learning implementation where fast and furious algorithms are used to gather the behavioral information of every user on social media and recommend them appropriate articles, multimedia files and much more according to their choice.

Machine learning is also the part of Artificial Intelligence where the requisite information is achieved after utilizing various algorithms and techniques, such as Supervised and Un-supervised Machine Learning Algorithms.

A machine learning professional must have the basic knowledge of statistics and probability, data evaluation, and technical skills of programming languages.

Types of Machine Learning

There are following three types of Machine learning:

. SUPERVISED LEARNING

. UNSUPERVISED LEARNING

. REINFORCEMENT LEARNING

2.2.4 Statistics and Probability

Data is controlled to extricate data out of it. The numerical foundation of data science is insights and likelihood as without having a reasonable learning of measurements and likelihood, there is a high plausibility of confounding the information and achieving an off base end. This is the reason that's why Statistics and Probability assume an essential job in data scienc

3 APPLICATION OF DATA SCIENCE

1. Marketing

There is a huge scope in marketing; for example, Improved Pricing strategy Companies like Uber, e-commerce companies can use data science-driven pricing, increasing their profits.

2. Healthcare

Using wearable data to prevent and monitor health problems. The data generated from the body can be used in healthcare to prevent future emergencies.

3. Banking and Finance

As we discussed the introduction to data science now, we will go ahead with applying data science uses in the banking sector for fraud detection, which can help reduce the Non-Performing Assets of banks.

4. Government Policies

The Government can use data science to prepare better policies to cater to the needs of the people and what they want using the data they can get by conducting surveys and others from other official sources

4 INTRODUCTION TO ARTIFICAL INTELLIGENT

Artificial Intelligence is an approach to make a computer, a robot, or a product to think how smart human think. AI is a study of how human brain think, learn, decide and work, when it tries to solve problems. And finally this study outputs intelligent software systems.The aim of AI is to improve computer functions which are related to human knowledge, for example, reasoning, learning, and problem-solving.

The intelligence is intangible. It is composed of

.Reasoning

.Learning

.Problem Solving

.Perception

.Linguistic Intelligence

The objectives of AI research are reasoning, knowledge representation, planning, learning, natural language processing, realization, and ability to move and manipulate objects. There are long-term goals in the general intelligence sector.Approaches include statistical methods, computational intelligence, and traditional coding AI. During the AI research related to search and mathematical optimization, artificial neural networks and methods based on statistics, probability, and economics, we use many tools. Computer science attracts AI in the field of science, mathematics, psychology, linguistics, philosophy and so on.

5 APPLICATION OF ARTIFICAL INTELLIGENT

1· Gaming

AI plays important role for machine to think of large number of possible positions based on deep knowledge in strategic games. for example, chess,river crossing, N-queens problems and etc.

2· Natural Language Processing

Interact with the computer that understands natural language spoken by humans.

3· Expert Systems

Machine or software provide explanation and advice to the users.

4· Vision Systems

Systems understand, explain, and describe visual input on the computer.

5· Speech Recognition

There are some AI based speech recognition systems have ability to hear and express as sentences and understand their meanings while a person talks to it. For example Siri and Google assistant.

6· Handwriting Recognition

The handwriting recognition software reads the text written on paper and recognize the shapes of the letters and convert it into editable text.

7· Intelligent Robots

Robots are able to perform the instructions given by a human.

6 DEFINATION OF MACHINE LEARNING

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problem

7 TRADITIONAL PROGRAMMING

Traditional programming is a manual process—meaning a person (programmer) creates the program. But without anyone programming the logic, one has to manually formulate or code rules.

In machine learning, on the other hand, the algorithm automatically formulates the rules from the dat

8 MACHINE LEARNING PROGRAMMING

Machine Learning Programming Unlike traditional programming, machine learning is an automated process. It can increase the value of your embedded analytics in many areas, including data prep, natural language interfaces, automatic outlier detection, recommendations, and causality and significance detection. All of these features help speed user insights and reduce decision bias.

9 SUPERVISED &UNSUPERVISED LEARNING

SUPERVISED LEARNING

Supervised Machine Learning Labeled dataset is used in supervised machine learning. Here, you must input variables (X) and output variables (Y) then you apply an appropriate algorithm to find the mapping function from input to output.

Y = f(X)

Supervised machine learning can be categorized into the following:-

Classification – where the output variable is a category like black or white, plus or minus.

Naïve Bayes, Support Vector Machine, Decision Tree are the most popular supervised machine learning algorithms.

Regression – where the output variable is a real value like weight, dollars, etc. Linear regression is used for regression problems.

UNSUPERVISED LEARNING

In this type of machine learning, un-labeled datasets is used. Here, you have only input variables (X) and no output variables; therefore, algorithm can be utilized to discover the inherent grouping from the input data.

Un-supervised machine learning can be categorized into the following: –

Clustering – where you find out the inherent groupings like grouping clients by procuring behavior.

K-means clustering, hierarchical clustering and density based spatial clustering are more popular clustering algorithms.

Association – where you find out rules that label large slices of your data.

Apriori algorithm is used for market basket analysis.

REINFORCEMENT LEARNING

Reinforcement learning is different from supervised learning, it is about to take an appropriate action in a particular situation to maximize the reward.

In supervised learning there are input as well as output variables, so, the model is trained with the correct response but in absence of training dataset, reinforcement agent learn from its experience and perform the given job efficiently.

In reinforcement learning, input should be an initial state and there are various output due to range of solutions to a specific problem but optimum solution is decided which based on maximum reward.

10 Advantages of Artificial Intelligence

1. Reduction in Human Error

One of the biggest advantages of Artificial Intelligence is that it can significantly reduce errors and increase accuracy and precision. The decisions taken by AI in every step is decided by information previously gathered and a certain set of algorithms. When programmed properly, these errors can be reduced to null.

2. Zero Risks

Another big advantage of AI is that humans can overcome many risks by letting AI robots do them for us. Whether it be defusing a bomb, going to space, exploring the deepest parts of oceans, machines with metal bodies are resistant in nature and can survive unfriendly atmospheres. Moreover, they can provide accurate work with greater responsibility and not wear out easily

3. 24x7 Availability

There are many studies that show humans are productive only about 3 to 4 hours in a day. Humans also need breaks and time offs to balance their work life and personal life. But AI can work endlessly without breaks. They think much faster than humans and perform multiple tasks at a time with accurate results. They can even handle tedious repetitive jobs easily with the help of AI algorithms.

4. Digital Assistance

Almost all the big organizations these days use digital assistants to interact with their customers which significantly minimizes the need for human resources. You can chat with a chatbot and ask them exactly what you need. Some chatbots have become so intelligent these days that you wouldn't be able to determine whether you are chatting with a chatbot or a human being.

5. Unbiased Decisions

Human beings are driven by emotions, whether we like it or not. AI on the other hand, is devoid of emotions and highly practical and rational in its approach. A huge advantage of Artificial Intelligence is that it doesn't have any biased views, which ensures more accurate decision-making.

Advantages of Machine learning

1. Easily identifies trends and patterns

Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans. For instance, for an e-commerce website like Amazon, it serves to understand the browsing behaviors and purchase histories of its users to help cater to the right products, deals, and reminders relevant to them. It uses the results to reveal relevant advertisements to them.

2. No human intervention needed

With ML, you don't need to babysit your project every step of the way. Since it means giving machines the ability to learn, it lets them make predictions and also improve the algorithms on their own. A common example of this is anti-virus software; they learn to filter new threats as they are recognized. ML is also good at recognizing spam.

3. Continuous Improvement

As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions. Say you need to make a weather forecast model. As the amount of data you have keeps growing, your algorithms learn to make more accurate predictions faster.

4. Handling multi-dimensional and multi-variety data

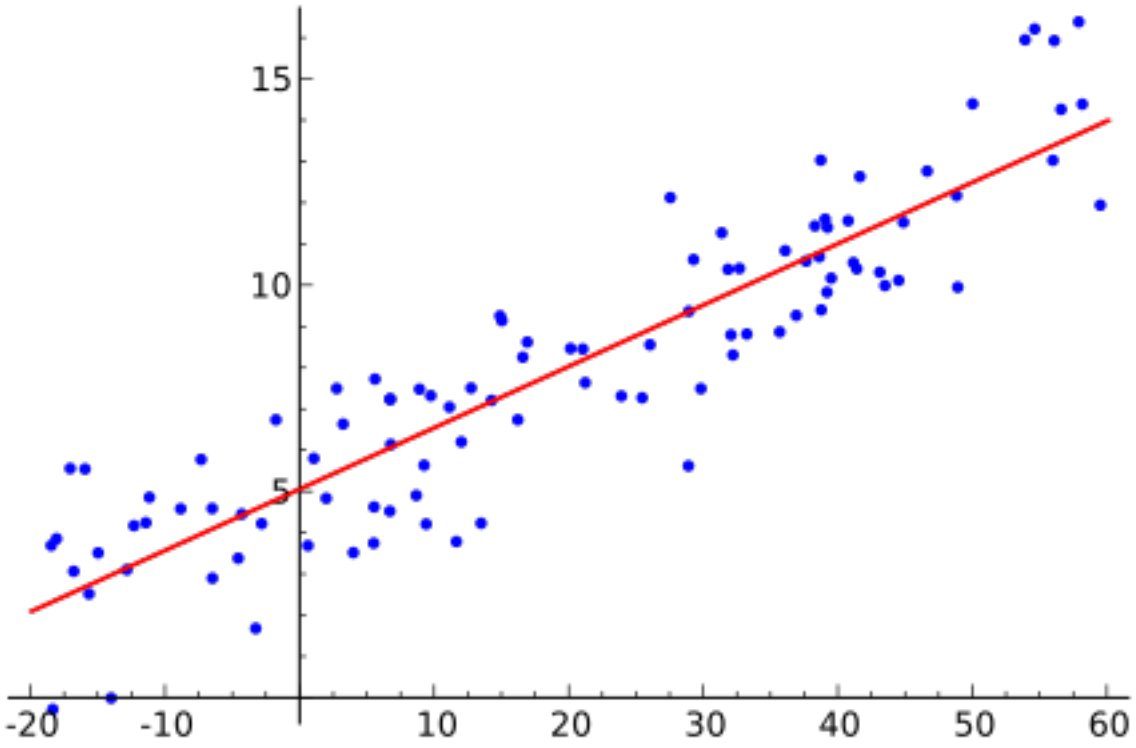
Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

5. Wide Applications

You could be an e-tailer or a healthcare provider and make ML work for you. Where it does apply, it holds the capability to help deliver a much more personal experience to customers while also targeting the right customers.

11 REGRESSION TECHNIQUE IN MACHINE LEARNING

Linear regression and logistic regression are two types of regression analysis techniques that are used to solve the regression problem using machine learning. They are the most prominent techniques of regression.



There are six type of machine learning

1. **Simple linear regression**
2. **Multiple linear regression**
3. **Polynomial regression**
4. **SVR**
5. **Decission tree**
6. **Random forest**

11.1 Simple linear regression

In simple linear regression, there is only one predictor variable. Since our goal is to predict the crew variable, we see from Figure 1 that the cabins variable correlates the most with the crew variable. Hence our simple regression model can be expressed in the form:

$$\hat{y}_i = m\,X_i + c$$

where m is the slope or regression coefficient, and c is the intercept. The model will be evaluated using the R2 score metric which can be calculated as follows:

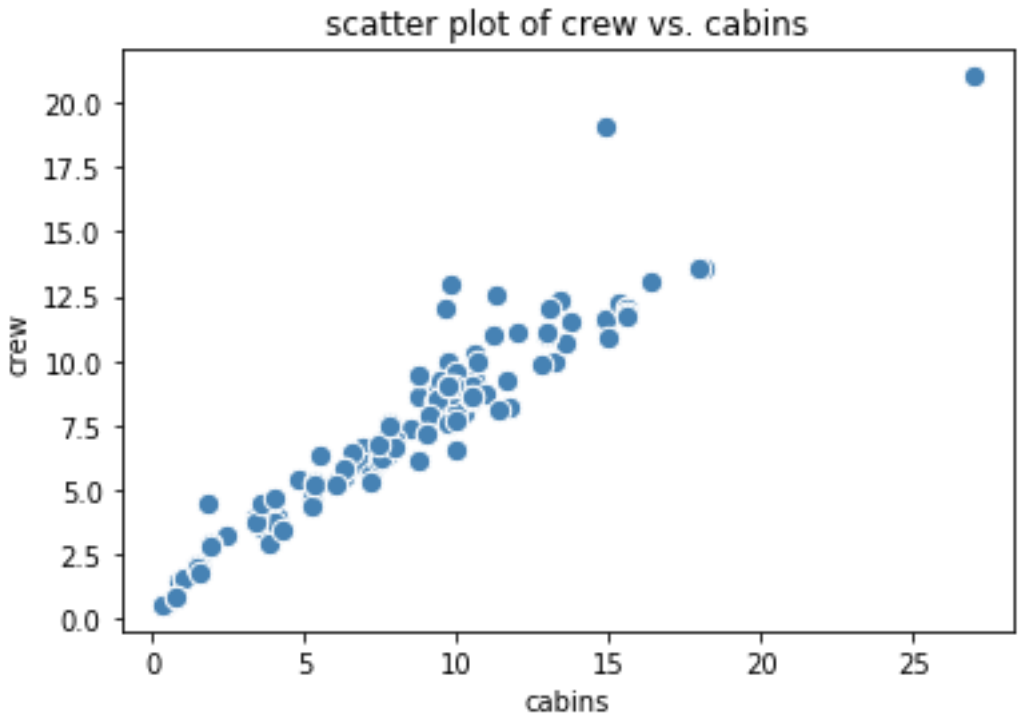
$$R2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

The R2 score takes values between 0 and 1. When R2 is close to 1, it means the predicted values agree closely with the actual values. If R2 is close to zero, then it means the predictive power of the model is very poor.

Let's now define and plot our independent and dependent variables:

```
X = df['cabins']
y = df['crew']
plt.scatter(X,y,c='steelblue', edgecolor='white', s=70)
plt.xlabel('cabins')
plt.ylabel('crew')
plt.title('scatter plot of crew vs. cabins')
plt.show()
```

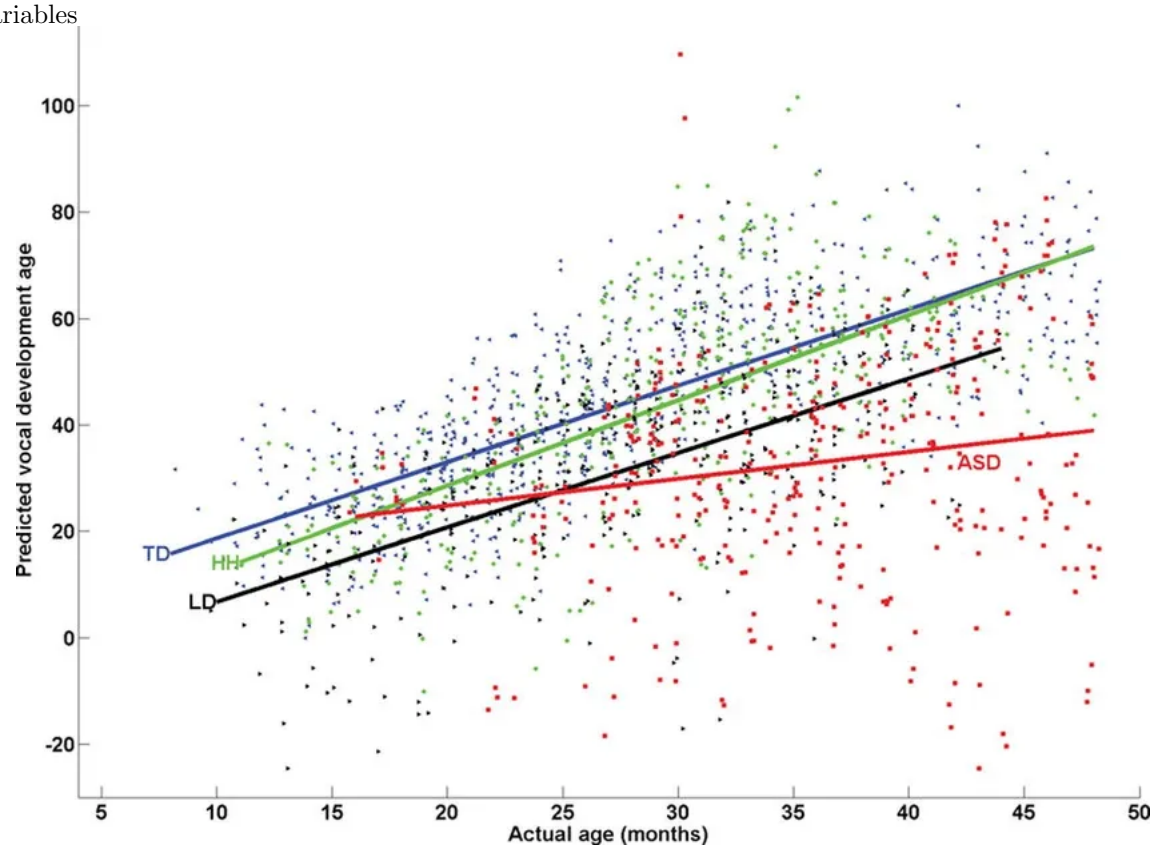
Output



11.2 Multiple linear regression

1. Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables. It is sometimes known simply as multiple regression, and it is an extension of linear regression. The variable that we want to predict is known as the dependent variable, while

the variables we use to predict the value of the dependent variable are known as independent or explanatory variables



Multiple Linear Regression Formula

$$yi = \beta 0 + \beta 1xi1 + \beta 2xi2 + ... + \beta pxip + \epsilon$$

Where:

. **yi i**s the dependent or predicted variable

. **β0** is the y-intercept, i.e., the value of y when both xi and x2 are 0.

. **β1** and **β2** are the regression coefficients representing the change in y relative to a one-unit change in xi1 and xi2, respectively.

. **βp** is the slope coefficient for each independent variable

. ϵ is the model's random error (residual) term.

Understanding Multiple Linear Regression

Simple linear regression enables statisticians to predict the value of one variable using the available information about another variable. Linear regression attempts to establish the relationship between the two variables along a straight line.

Multiple regression is a type of regression where the dependent variable shows a linear relationship with two or more independent variables. It can also be non-linear, where the dependent and independent variables do not follow a straight line.

Both linear and non-linear regression track a particular response using two or more variables graphically. However, non-linear regression is usually difficult to execute since it is created from assumptions derived from trial and error.

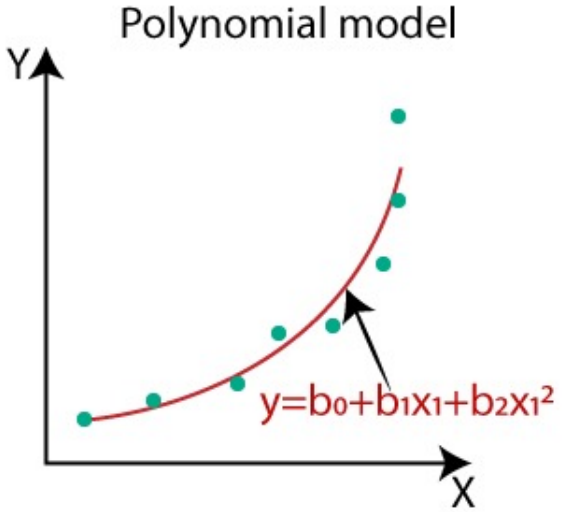
Assumptions of Multiple Linear Regression

Multiple linear regression is based on the following assumptions:

1. A linear relationship between the dependent and independent variables
2. The independent variables are not highly correlated with each other
3. The variance of the residuals is constant
4. Independence of observation
5. Multivariate normality

11.3 Polynomial regressionSVR

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x)



Understanding polynomial Regression

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In simple linear regression, the model

$y = \beta_0 + \beta_1 x + \varepsilon$

is used, where ε is an unobserved random error with mean zero conditioned on a scalar variable x. In this model, for each unit increase in the value of x, the conditional expectation of y increases by β1 units.

In many settings, such a linear relationship may not hold. For example, if we are modeling the yield of a chemical synthesis in terms of the temperature at which the synthesis takes place, we may find that the yield improves by increasing amounts for each unit increase in temperature. In this case, we might propose a quadratic model of the form

$y=\beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

In this model, when the temperature is increased from x to x + 1 units, the expected yield changes by $\beta_1 + \beta_2(2x + 1)$.(This can be seen by replacing x in this equation with x+1 and subtracting the equation in x from the equation in x+1.) For infinitesimal changes in x, the effect on y is given by the total derivative with respect to x: $\beta_1 + 2\beta_2 x$.The fact that the change

in yield depends on x is what makes the relationship between x and y nonlinear even though the model is linear in the parameters to be estimated.

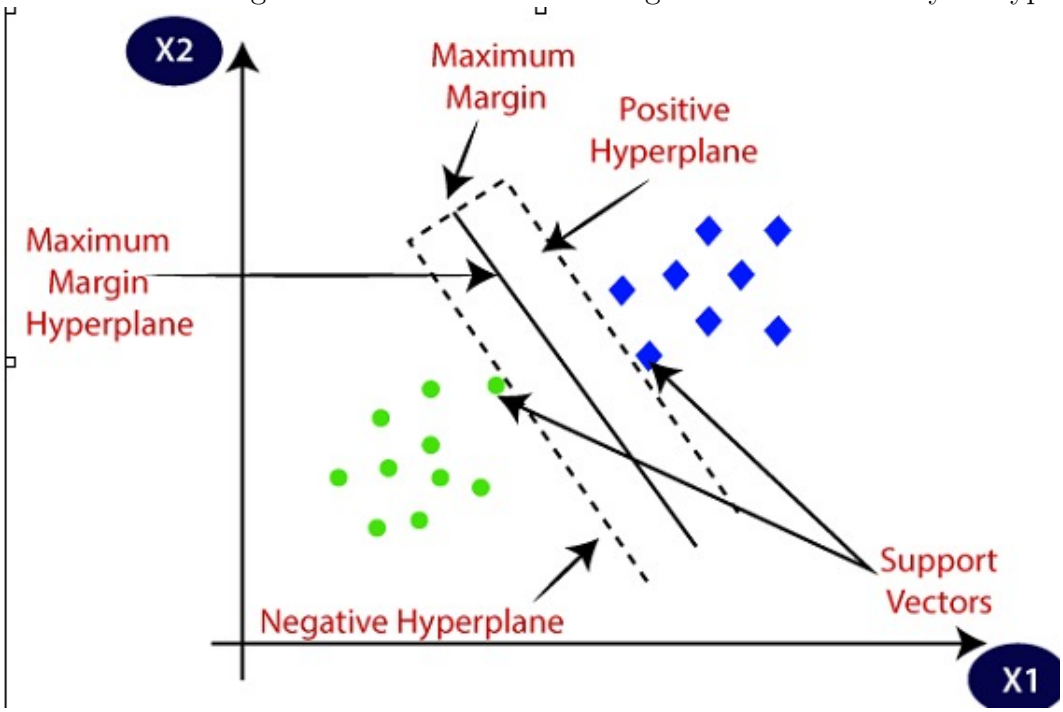
In general, we can model the expected value of y as an nth degree polynomial, yielding the general polynomial regression model

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \varepsilon$

Conveniently, these models are all linear from the point of view of estimation, since the regression function is linear in terms of the unknown parameters β0, β1, ... Therefore, for least squares analysis, the computational and inferential problems of polynomial regression can be completely addressed using the techniques of multiple regression. This is done by treating x, x2, ... as being distinct independent variables in a multiple regression model.

11.4 SVR AND SVM

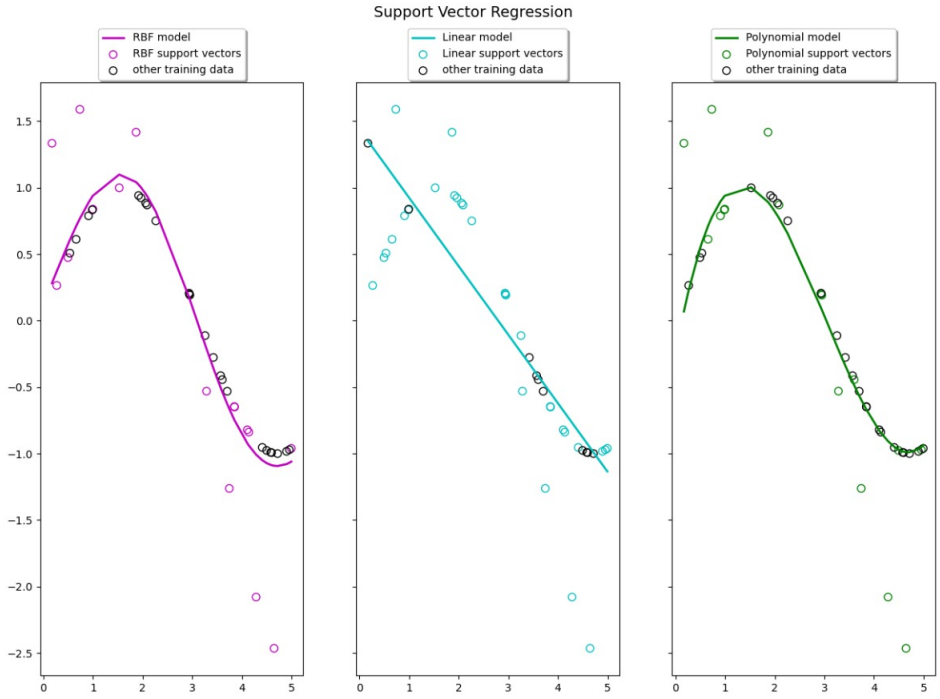
SVM:-Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:



SVR:-Support Vector Regression is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that

has the maximum number of points.Unlike other Regression models that try to minimize the error between the real and predicted value, the SVR tries to fit the best line within a threshold value. The threshold value is the distance between the hyperplane and boundary line. The fit time complexity of SVR is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples.

For large datasets, Linear SVR or SGD Regressor is used. Linear SVR provides a faster implementation than SVR but only considers the linear kernel. The model produced by Support Vector Regression depends only on a subset of the training data, because the cost function ignores samples whose prediction is close to their target.



12 Decision tree

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

The decision tree may not always provide a clear-cut answer or decision. Instead, it may present options so the data scientist can make an informed decision on their own. Decision trees imitate human thinking, so it's generally easy for data scientists to understand and interpret the

Decision Tree Work

Before we dive into how a decision tree works

. **Root node:** The base of the decision tree.

. **Splitting:** The process of dividing a node into multiple sub-nodes.

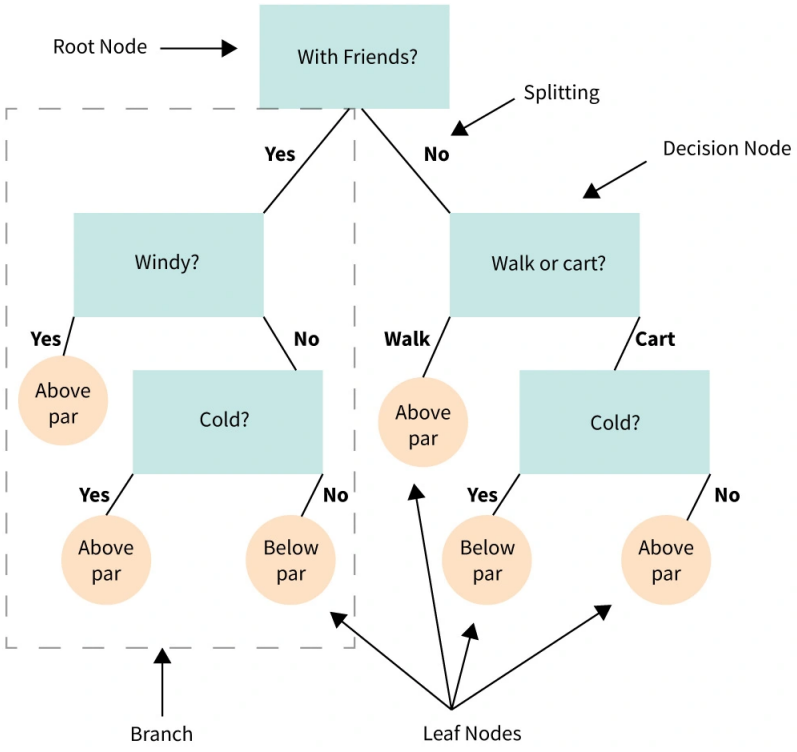
. **Decision node:** When a sub-node is further split into additional sub-nodes.

. **Leaf node:** When a sub-node does not further split into additional sub-nodes; represents possible outcomes.

. **Pruning:** The process of removing sub-nodes of a decision tree.

. **Branch:** A subsection of the decision tree consisting of multiple nodes.

A decision tree resembles, well, a tree. The base of the tree is the root node. From the root node flows a series of decision nodes that depict decisions to be made. From the decision nodes are leaf nodes that represent the consequences of those decisions. Each decision node represents a question or split point, and the leaf nodes that stem from a decision node represent the possible answers. Leaf nodes sprout from decision nodes similar to how a leaf sprouts on a tree branch. This is why we call each subsection of a decision tree a "branch." Let's take a look at an example for this. You're a golfer, and a consistent one at that. On any given day you want to predict where your score will be in two buckets: below par or over par.



While you are a consistent golfer, your score is dependent on a few sets of input variables. Wind speed, cloud cover and temperature all play a role. In addition, your score tends to deviate depending on whether or not you walk or ride a cart. And it deviates if you are golfing with friends or strangers.

In this example, there are two leaf nodes: below par or over par. Each of the input variables will determine decision nodes. Was it windy? Cold? Did you golf with friends? Did you walk or take a cart? With enough data on your golfing habits (and assuming you are a consistent golfer), a decision tree could help predict how you will do on the course on any given day.

Types of Decision Trees

There are two main types of decision trees

. **Categorical Variable Decision Tree**

. **Continuous Variable Decision Tree**

Advantages

. Works for numerical or categorical data and variables.
. Models problems with multiple outputs.
. Tests the reliability of the tree.
. Requires less data cleaning than other data modeling techniques.
. Easy to explain to those without an analytical background.

Disadvantages

. Affected by noise in the data.
. Not ideal for large datasets.
. Can disproportionately value, or weigh, attributes.
. The decisions at nodes are limited to binary outcomes, reducing the complexity that the tree can handle.
. Trees can become very complex when dealing with uncertainty and numerous linked outcomes.

13 Random forest

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems.

Random forest algorithm

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees. Feature randomness, also known as feature bagging or "the random subspace method"(link resides outside IBM) (PDF, 121 KB), generates a random subset of features, which ensures low correlation among decision trees. This is a key difference between decision trees and random forests. While decision trees consider all the possible feature splits, random forests only select a subset of those features.

If we go back to the "should I surf?" example, the questions that I may ask to determine the prediction may not be as comprehensive as someone else's set of questions. By accounting for all the potential variability in the data, we can reduce the risk of overfitting, bias, and overall variance, resulting in more precise predictions.

How it works

Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve for regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote—i.e. the most frequent categorical variable—will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.

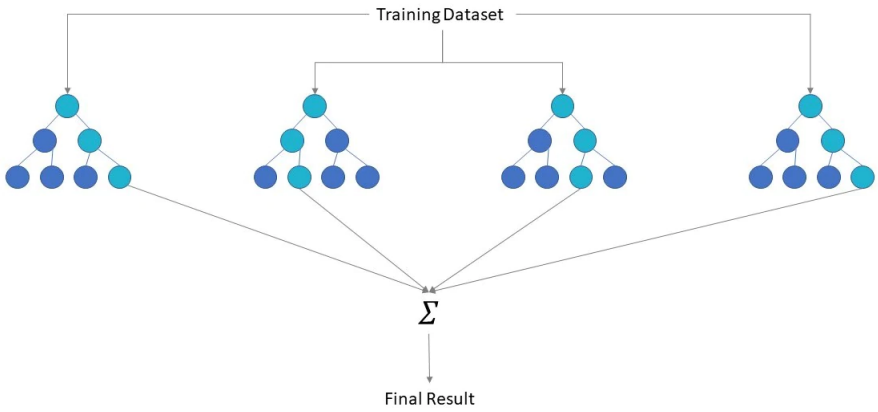


Diagram of random forest algorithm

Benefits

. Reduced risk of overfitting:

. Provides flexibility:

. Easy to determine feature importance:

Challenges

. Time-consuming process:

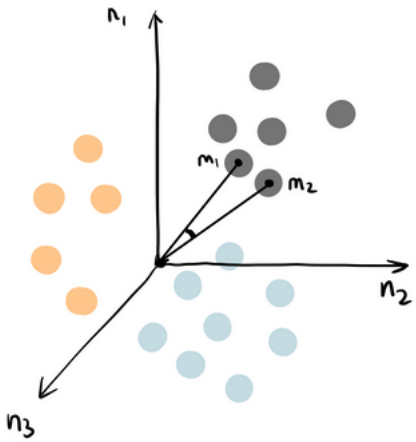
. Requires more resources:

. More complex:

14 CLASSIFICATION TECHNIQUES IN MACHINE LEARNING

There are many techniques for solving classification problems: classification trees, logistic regression, discriminant analysis, neural networks, boosted trees, random forests, deep learning methods, nearest neighbors, support vector machines, etc,

14.1 K-Nearest Neighbour (KNN)



You can think of k nearest neighbour algorithm as representing each data point in a n dimensional space — which is defined by n features. And it calculates the distance between one point to another, then assign the label of unobserved data based on the labels of nearest observed data points. KNN can also be used for building recommendation system, check out my article on "Collaborative Filtering for Movie Recommendation" if you are interested in this topic.

```
from sklearn.neighbors import KNeighborsClassifier
knn = KNeighborsClassifier()
knn.fit(X_train, y_train)
y_pred = knn.predict(X_test)
```

Advantages

. It is simple to implement.

. It is robust to the noisy training data

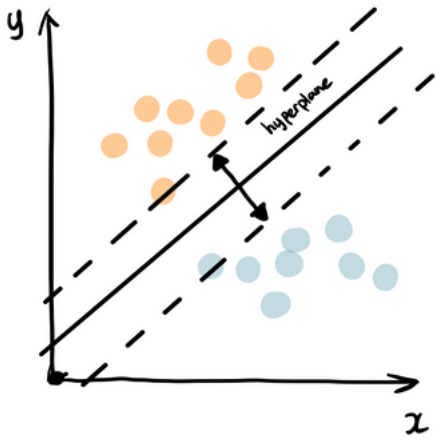
. It can be more effective if the training data is large.

Disadvantages :

. Always needs to determine the value of K which may be complex some time.

. The computation cost is high because of calculating the distance between the data points for all the training samples.

14.2 Support Vector Machine (SVM)



support vector machine (image by author) Support vector machine finds the best way to classify the data based on the position in relation to a border between positive class and negative class. This border is known as the hyperplane which maximize the distance between data points from different classes. Similar to decision tree and random forest, support vector machine can be used in both classification and regression, SVC (support vector classifier) is for classification problem.

```
from sklearn.svm import SVC
svc = SVC()
svc.fit(X_train, y_train)
y_pred = svc.predict(X_test)
```

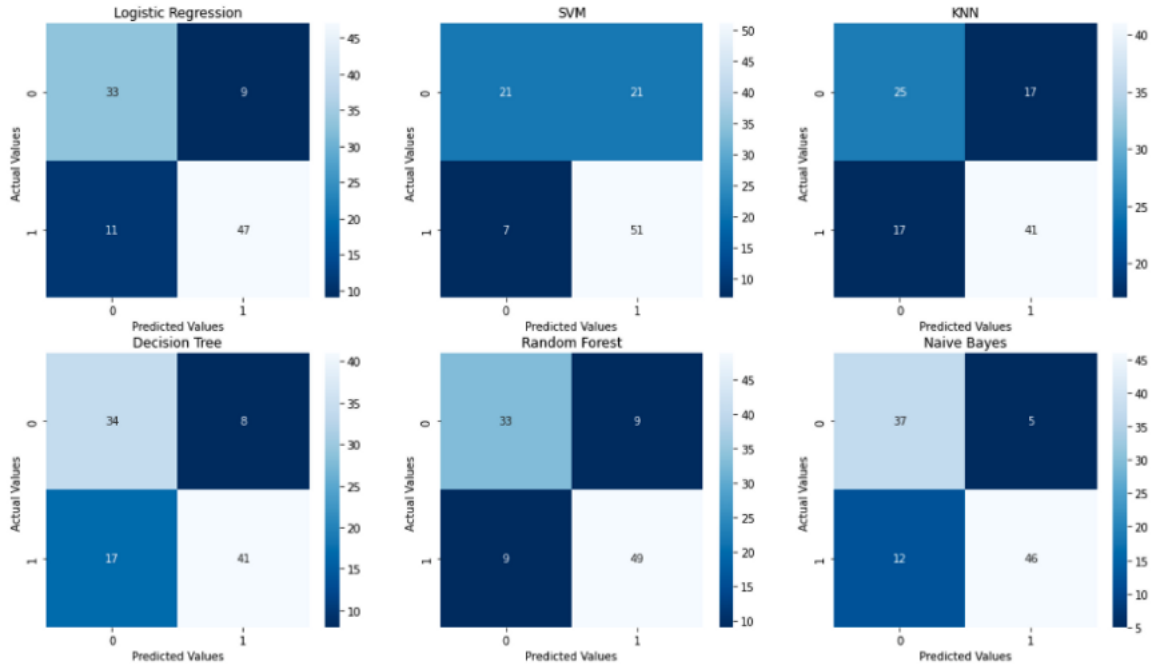
14.3 Confusion matrix

Confusion matrix indicates the actual values vs. predicted values and summarize the true negative, false positive, false negative and true positive values in a matrix format.

True Negative	False Positive
False Negative	True Positive

Then we can use seaborn to visualize the confusion matrix in a heatmap.

```
## plot confusion matrix
fig = plt.figure(figsize = (18,10))
for i in range(len(cm_list)):
    cm = cm_list[i]
    model = model_list[i]
    sub = fig.add_subplot(2, 3, i+1).set_title(model)
    cm_plot = sns.heatmap(cm, annot=True, cmap = 'Blues_r')
    cm_plot.set_xlabel('Predicted Values')
    cm_plot.set_ylabel('Actual Values ')
```



15 Clustering Technique in machine learning

15.1 K Means Clustering:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this

algorithm is to minimize the sum of distances between the data point and their corresponding clusters

The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

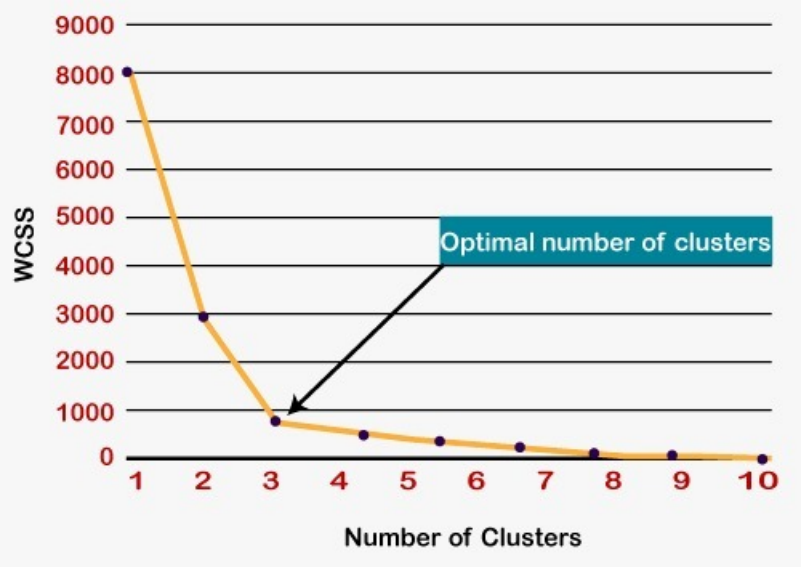
Determines the best value for K center points or centroids by an iterative process. Assigns each data point to its closest k-center. Those data points which are near to the particular k-center, create a cluster. Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

Choosing Right No. of f K's:

$$\textbf{WCSS}=\sum dist(p_1.c_1)^2 + \sum dist(p_2.c_2)^2 + \sum dist(p_1.c_1)^2 + \ldots\ldots\ldots$$

Minm. WCSS=0(if we have n points and have n clusters so dist.=0)

Max WCSS is when cluster is 1.



Code:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
dataset= pd.read_csv(‘‘Mall_customers.csv’’)
x=dataset.loc[:,[3:4]].values
from sklearn cluster import K means
wcss = [ ]
for i in range(1,11)
k means = k means(n_cluster=i,init=’k_means++’,random_state=42)
k means.fit(x)
wcss,append(k means.inertia)
plt.plot(range(1,11),wcss)
plt.title(’The Elbow method’)
plt.xlabel(’The no.of clusters’)
plt.ylabel(’wcss’)
```

```
plt.show()
kmeans = KMeans(n_clusters=5, init='k-means++', random_state= 42)
y_predict= kmeans.fit_predict(x)
mtp.scatter(x[y_predict == 0, 0], x[y_predict == 0, 1], s = 100, c = 'blue', label = 'Cluster
mtp.scatter(x[y_predict == 1, 0], x[y_predict == 1, 1], s = 100, c = 'green', label = 'Cluster
mtp.scatter(x[y_predict== 2, 0], x[y_predict == 2, 1], s = 100, c = 'red', label = 'Cluster 3'
mtp.scatter(x[y_predict == 3, 0], x[y_predict == 3, 1], s = 100, c = 'cyan', label = 'Cluster
mtp.scatter(x[y_predict == 4, 0], x[y_predict == 4, 1], s = 100, c = 'magenta', label = 'Clust
mtp.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s = 300, c = 'yellow
mtp.title('Clusters of customers')
mtp.xlabel('Annual Income (k$)')
mtp.ylabel('Spending Score (1-100)')
mtp.legend()
mtp.show()
```