UNIT 1

Getting Started

DS Process

- Detailed predictions of what the future will hold
- The power of data science comes from a deep understanding of statistics and algorithms, programming and hacking,communication skills.
- Frame the problem
- Process the data
- Explore the data
- Communicate results of the analysis

Step1: Frame the problem

Ask the right questions depends on the Client Requirements

Step 2: Collect the right data

- Data can be collectd from the specified company CRM software which has in SQL table formats.
- Extract the required Fata from the available SQL Database that can also be Communicate results of the analysis
- Concentrate on Customer Privacy and Declaration policy while handling with their data.
- Segragate the required data as Data set as of now, Its large to hold, Extract with their data

Step3: Process the data

- Data Cleaning is the Tedious Process for Data Scientist day to day life.
- Focus and Patience must be attain to succeed in this process
- Check any Errors found in CSV file in Client requirement view.
- Decide how to choose the right data from unclear format sense for your specific problem
- FInally, Prepare for Exploring analysis after the Data Wrangling

Step4: Explore your data

- The Process of finding which parts of the data are significant in answering your questions step is called exploratory data analysis.

- Predict some information from the found data and match it with the available and unavailable Customer,
- Make a difference and set some value as how proceed with the work and group the predicted data with initial prerequites
- Eventually, This processed dataset to find a set of factors that could solve client's original problem

Step 5 : Analyze Your Data In Depth

- Data Scientist need to identify which is feature and Label, After finding it apply thode identified features in Machine Learning technique like Supervised Learning.
- Contrary to supervised learning, unsupervised learning techniques extract information from data without any labels supplied
- As last Step communicate our results to our client in a way that is compelling and comprehensible for them

Step 6 : Visualize and Communicate Your Findings

- Communication is one of the most underrated skills for a data scientist
- Present Our Result in Visual formats such as Spreadsheet to client View
- As part of our role as the interpreter of data, We will be often called upon to make recommendations about how others should use your results.
- At last We are done that exploring Technology with Business Concern
- makes the career path of data science so challenging, and so valuable.

History of Machine Learning

- Machine Learning is a sub-set of artificial intelligence where computer algorithms are used to autonomously learn from data and information

Many Scientist had involved for the birth of Machine Learning Technique. Some of them, I feels inspire and note it down

1950 — Alan Turing creates the “Turing Test” to determine if a computer has real intelligence. To pass the test, a computer must be able to fool a human into believing it is also human.

2011 — Google Brain is developed, and its deep neural network can learn to discover and categorize objects much the way a cat does.

2016 – Google’s artificial intelligence algorithm beats a professional player at the Chinese board game Go, which is considered the world’s most complex board game and is many times harder than chess. The AlphaGo algorithm developed by Google DeepMind managed to win five games out of five in the Go competition.

The Great AI Awakening

The phrase “artificial intelligence” is invoked as if its meaning were self-evident, but it has always been a source of confusion and controversy.

Mainly it evolved by the thought and approach of human Mind which can Work with Artificial Knowledge which was superior to the power the of Human Knowledge.

Part I: Learning Machine

1. The Birth of Brain

Many concepts of Machine Learning Born by the research of Google Corporation (according to the Article)

The fundamental Knowledge and Foundation has born in Machine Learning

Part II: The Language Machine

Theory Became Product

Language plays vital role in every communication. By the way of AI even it is Machine it need to learn language inorder to meet user Requirements. This Theoritical concept later converted into Product of largest Product Based Company called Google.

LEARN PYTHON:

Here I have added all Basic Concepts of Python in 100 lines of Code.

```
#Print
print("Hello, Happy to be here as a Python Learner")
#Data Types
a=b+c-7*3//6
f="Learner"
```

```
print(a,b)
#Array of Strings
Unit1=["Basics","Flask","AI"]
print(unit1[0],unit1[-1],unit1[0:])
python=["Char","Int","Float"]
merge_Array=[unit1,python]
print(merge_Array[2][1])
#Set
Details:{'Datatypes':'Char','Array':'2D','Package':'os'}
print(Details.values(),Details.keys(),Details['package'])
#Conditional and Looping Statements
arr=[]
list1=[]
if a>2:
  for i in range(len(arr)):
     for j in range(1,len(arr)):
        if arr[i]>arr[j] and a==3 or a==4 arr[i] not in list1:
          return 0
        elif:
          return 1
        else:
          return -1
else:
   return a
#Functions
def Length(string):
  l=len(string)
  if l==0:
     return 0
  else
      return l
str="Codest"
Length(str)
#File Handling
import os
import sys
file=open("Text.txt",wb+)
print(file.mode)
file=open("Text.txt",r+)
print(file.read())
#Class
Class Python:
   __name=""
   __std=0
   __dept=""
```

```
    def init(self,name,std,dept) :
        self.__name=Yes!
        self.__std=10
        self.__dept=CSE
    def getname(self):
        return self.__name


cls1=python()
print(cls1)
python.getname(cls1)
cls1.getname()
```

#Inheritance

```
class Python1:
    __name=""
    __std=0
    __dept=""

    def init(self,name,std,dept) :
        self.__name=Yes!
        self.__std=10
        self.__dept=CSE
    def getname(self):
        return self.__name
class Python2(Python1):
    def getstd(self):
        return self.__std

cls1=python1()
cls1.getname()
cls2=python2()
cls2.getstd()
```

#Polymorphism

```
class Python1:
    def init(self,name,std,dept) :
         self.__name=Yes!
         self.__std=10
         self.__dept=CSE
     def getname(self):
         return self.__name
class Python2:
    def init(self,name,std,dept) :
         self.__name=Yes!
         self.__std=10
```

```
        self.__dept=CSE
    def getname(self):
        return self.__name

cls1=python1("Hi",2,CS)
cls1=python2("Hii",3,CSS)
```

3. DATA WRANGLING

- Data wrangling is the process of gathering, selecting, and transforming data to answer an analytical question. Also known as data cleaning or “munging”, legend has it that this wrangling costs analytics professionals as much as 80% of their time, leaving only 20% for exploration and modeling.

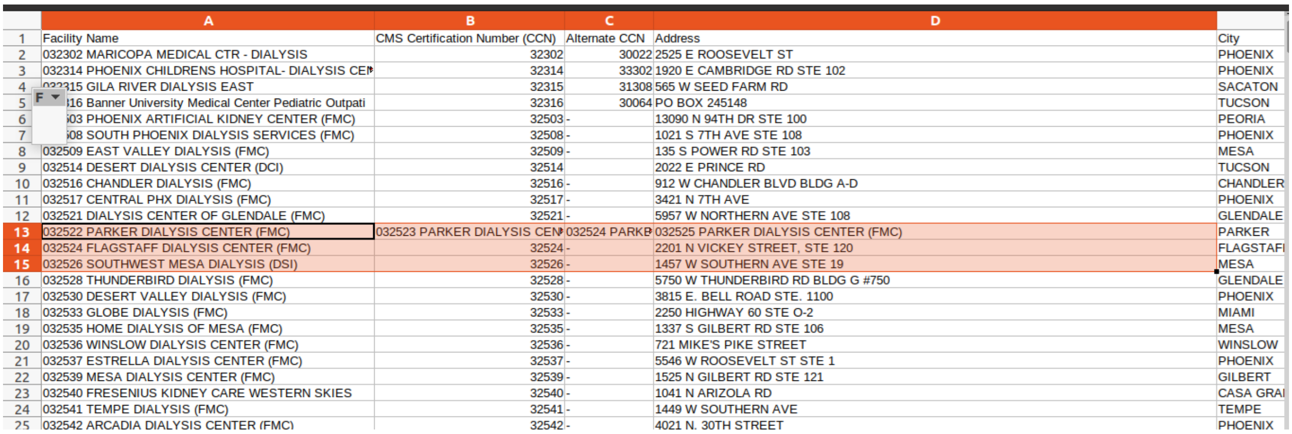
Data Quality:

Data is the most crucial and sensible tool for the Data Scientist and Analyst.Inorder to Identify the Quality of the target completes can only by confirmed through pure and clean data.

Data Wrangling With Pandas

Pandas is one of the most popular Python library for data wrangling. In this example we’ll use Pandas to learn data wrangling techniques to deal with some of the most common data formats and their transformations.

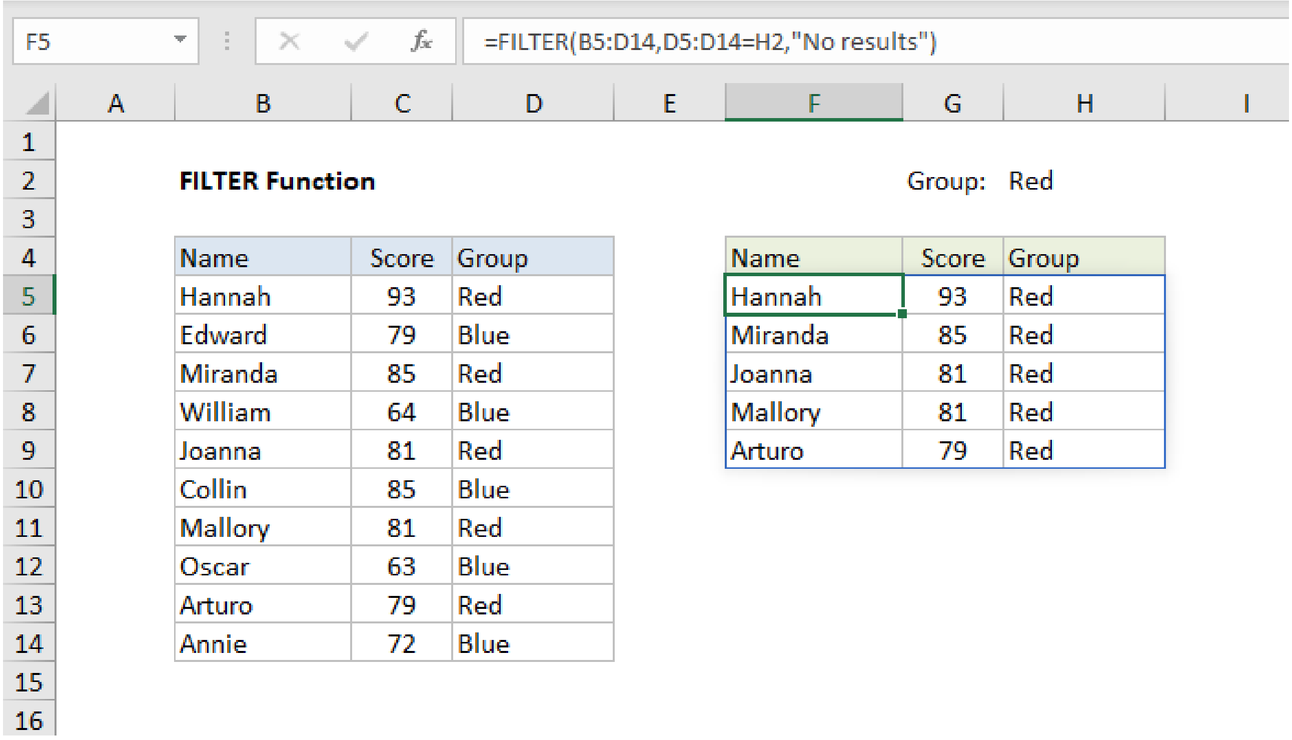
Dropping and Missing Null Values in Data Set



	A	B	C	D	
1	Facility Name	CMS Certification Number (CCN)	Alternate CCN	Address	City
2	032302 MARICOPA MEDICAL CTR - DIALYSIS	32302	30022	2525 E ROOSEVELT ST	PHOENIX
3	032314 PHOENIX CHILDRENS HOSPITAL- DIALYSIS CEN	32314	33302	1920 E CAMBRIDGE RD STE 102	PHOENIX
4	315 GILA RIVER DIALYSIS EAST	32315	31308	565 W SEED FARM RD	SACATON
5	16 Banner University Medical Center Pediatric Outpati	32316	30064	PO BOX 245148	TUCSON
6	03 PHOENIX ARTIFICIAL KIDNEY CENTER (FMC)	32503	-	13090 N 94TH DR STE 100	PEORIA
7	08 SOUTH PHOENIX DIALYSIS SERVICES (FMC)	32508	-	1021 S 7TH AVE STE 108	PHOENIX
8	032509 EAST VALLEY DIALYSIS (FMC)	32509	-	135 S POWER RD STE 103	MESA
9	032514 DESERT DIALYSIS CENTER (DCI)	32514		2022 E PRINCE RD	TUCSON
10	032516 CHANDLER DIALYSIS (FMC)	32516	-	912 W CHANDLER BLVD BLDG A-D	CHANDLER
11	032517 CENTRAL PHX DIALYSIS (FMC)	32517	-	3421 N 7TH AVE	PHOENIX
12	032521 DIALYSIS CENTER OF GLENDALE (FMC)	32521	-	5957 W NORTHERN AVE STE 108	GLENDALE
13	032522 PARKER DIALYSIS CENTER (FMC)	032523 PARKER DIALYSIS CEN	032524 PARKE	032525 PARKER DIALYSIS CENTER (FMC)	PARKER
14	032524 FLAGSTAFF DIALYSIS CENTER (FMC)	32524	-	2201 N VICKEY STREET, STE 120	FLAGSTAFI
15	032526 SOUTHWEST MESA DIALYSIS (DSI)	32526	-	1457 W SOUTHERN AVE STE 19	MESA
16	032528 THUNDERBIRD DIALYSIS (FMC)	32528	-	5750 W THUNDERBIRD RD BLDG G #750	GLENDALE
17	032530 DESERT VALLEY DIALYSIS (FMC)	32530	-	3815 E. BELL ROAD STE. 1100	PHOENIX
18	032533 GLOBE DIALYSIS (FMC)	32533	-	2250 HIGHWAY 60 STE O-2	MIAMI
19	032535 HOME DIALYSIS OF MESA (FMC)	32535	-	1337 S GILBERT RD STE 106	MESA
20	032536 WINSLOW DIALYSIS CENTER (FMC)	32536	-	721 MIKE'S PIKE STREET	WINSLOW
21	032537 ESTRELLA DIALYSIS CENTER (FMC)	32537	-	5546 W ROOSEVELT ST STE 1	PHOENIX
22	032539 MESA DIALYSIS CENTER (FMC)	32539	-	1525 N GILBERT RD STE 121	GILBERT
23	032540 FRESENIUS KIDNEY CARE WESTERN SKIES	32540	-	1041 N ARIZOLA RD	CASA GRAI
24	032541 TEMPE DIALYSIS (FMC)	32541	-	1449 W SOUTHERN AVE	TEMPE
25	032542 ARCADIA DIALYSIS CENTER (FMC)	32542	-	4021 N. 30TH STREET	PHOENIX

Filtering and Grouping Data

The Process of Optimising data set for the great Production of outcomes for the analysis inorder to fulfil Clients Requirements



=FILTER(B5:D14,D5:D14=H2,"No results")

FILTER Function

Group: Red

Name	Score	Group
Hannah	93	Red
Edward	79	Blue
Miranda	85	Red
William	64	Blue
Joanna	81	Red
Collin	85	Blue
Mallory	81	Red
Oscar	63	Blue
Arturo	79	Red
Annie	72	Blue

Name	Score	Group
Hannah	93	Red
Miranda	85	Red
Joanna	81	Red
Mallory	81	Red
Arturo	79	Red

Pandas Data Wrangling Implementation

```
import pandas as pd

# Assign data
data = {'Name': ['Jai', 'Princi', 'Gaurav',
                                  'Anuj', 'Ravi', 'Natasha', 'Riya'],
                 'Age': [17, 17, 18, 17, 18, 17, 17],
                 'Gender': ['M', 'F', 'M', 'M', 'M', 'F', 'F'],
                 'Marks': [90, 76, 'NaN', 74, 65, 'NaN', 71]}

# Convert into DataFrame
df = pd.DataFrame(data)

# Display data
df
```

SUPERVISED MACHINE LEARNING

We can categories Machine Learning into two types

1 Supervised Machine Learning
2 UnSupervised Machine Learning

We will going to discuss Supervised Machine Learning in this Chapter.

Part I: LINEAR PREDICTIONS

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available

1. Simple Linear Regression

Simple Linear Regression With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

2. Ordinary Least Squares

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

3. Gradient Descent

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

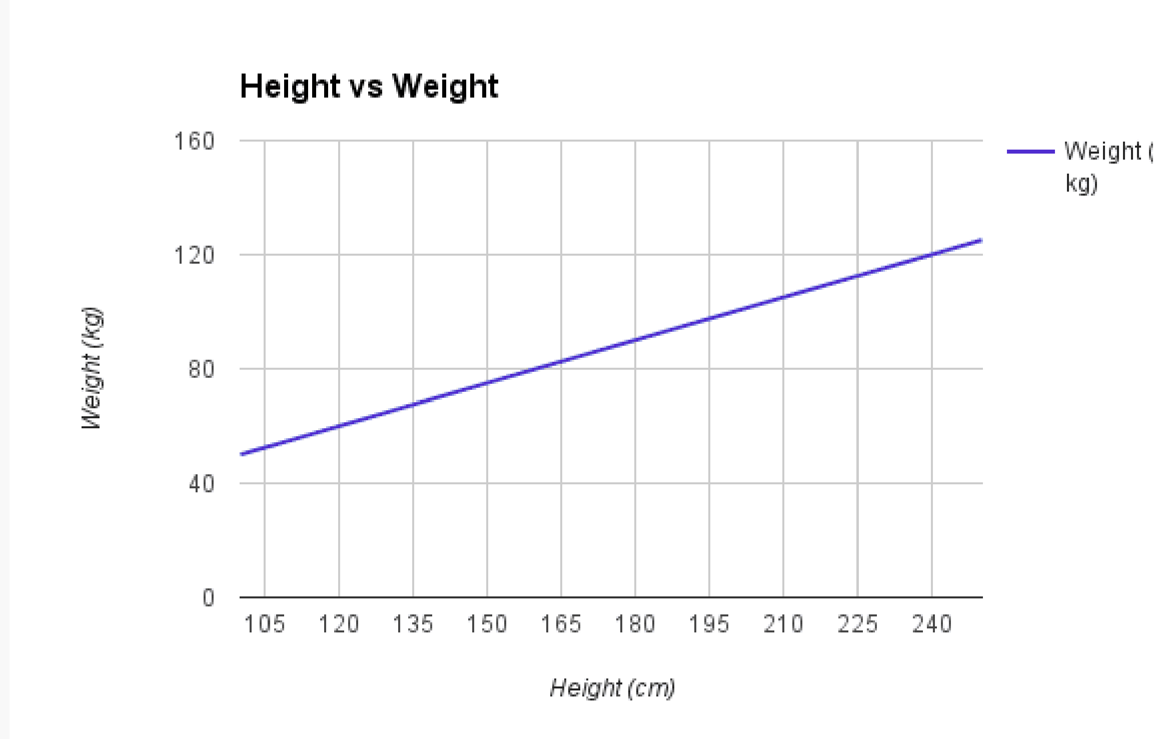
4. Regularization

There are extensions of the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).

Making Predictions with Linear Regression

Let’s make this concrete with an example. Imagine we are predicting weight (y) from height (x). Our linear regression model representation for this problem would be:

$$\mathbf{y = B0 + B1 * x1}$$



Sample Height vs Weight Linear Regression

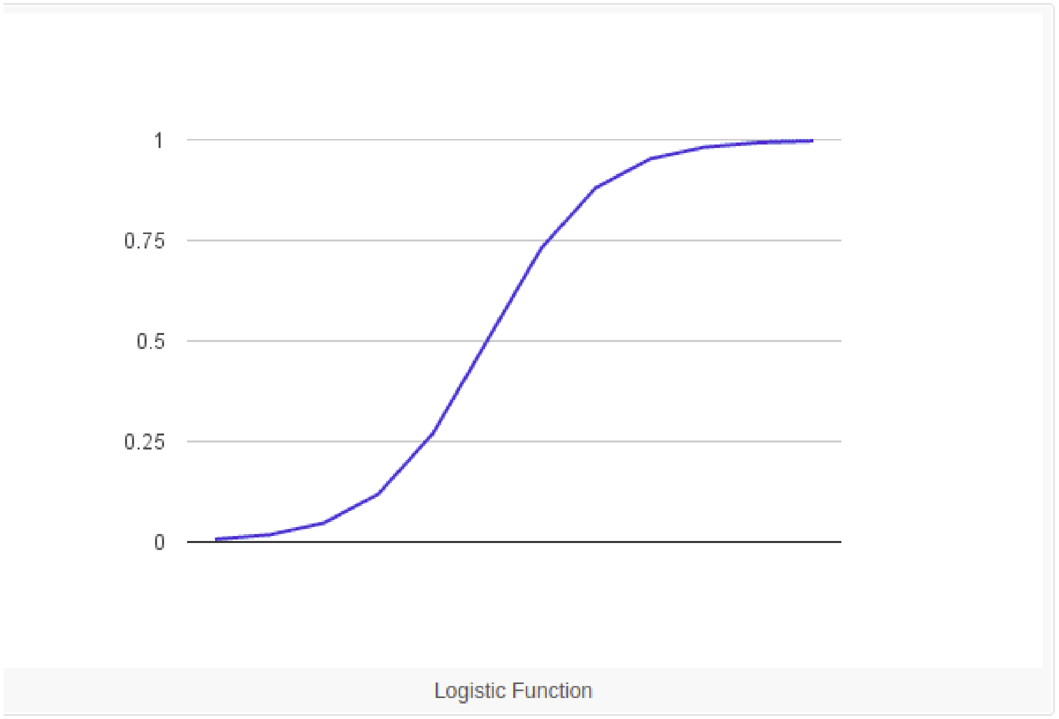
Logistic regression

Logistic regression is another technique borrowed by machine learning from the field of statistics.

The logistic function, also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

1 / (1 + e^-value)

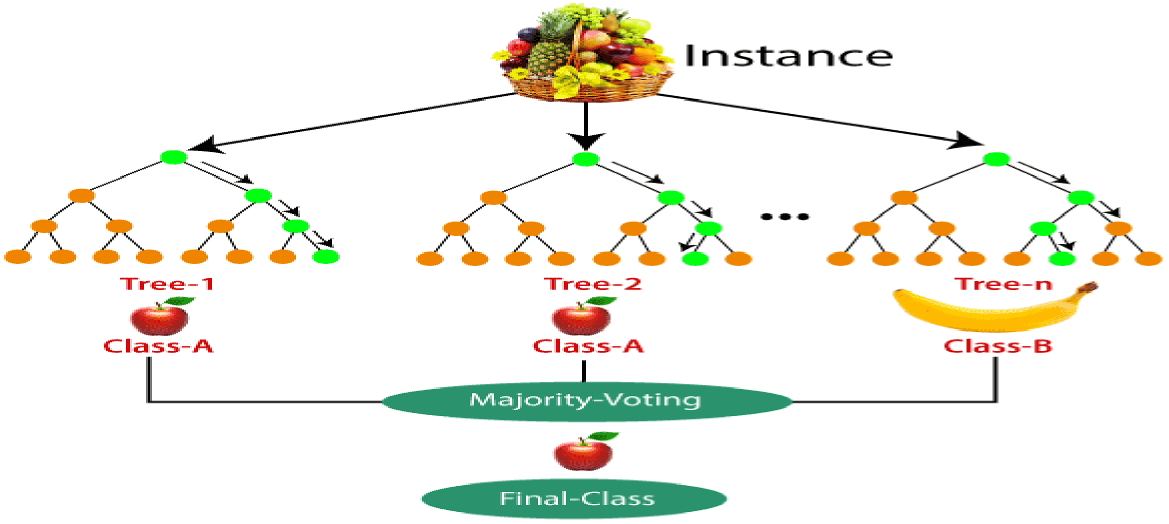
ution.



NON-LINEAR PREDICTIONS

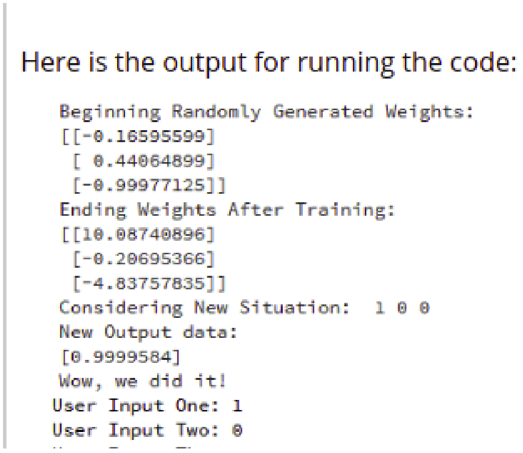
Random Forests

- Random Forest is a robust machine learning algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions
- We need to approach the Random Forest regression technique like any other machine learning technique
- Design a specific question or data and get the source to determine the required data.
- Make sure the data is in an accessible format else convert it to the required format.
- Specify all noticeable anomalies and missing data points that may be required to achieve the required data.
- Create a machine learning model
- Set the baseline model that you want to achieve
- Train the data machine learning model.
- Provide an insight into the model with test data
- Now compare the performance metrics of both the test data and the predicted data from the model.
- If it doesn’t satisfy your expectations, you can try improving your model accordingly or dating your data or use another data modeling technique.
- At this stage we interpret the data you have gained and report accordingly.



Neural Network

The process of fine-tuning the weights and biases from the input data is known as training the Neural Network.



Here is the output for running the code:

```
Beginning Randomly Generated Weights:
[[-0.16595599]
 [ 0.44064899]
 [-0.99977125]]
Ending Weights After Training:
[[10.08740896]
 [-0.20695366]
 [-4.83757835]]
Considering New Situation:  1 0 0
New Output data:
[0.9999584]
Wow, we did it!
User Input One: 1
User Input Two: 0
```

The neuron began by allocating itself some random weights. Thereafter, it trained itself using the training examples.

Consequently, if it was presented with a new situation [1,0,0], it gave the value of 0.9999584.

Advantages:

1. ANNs have the ability to learn and model non-linear and complex relationships, which is really important because in real-life, many of the relationships between inputs and outputs are non-linear as well as complex.

2. ANNs can generalize, After learning from the initial inputs and their relationships, it can infer unseen relationships on unseen data as well, thus making the model generalize and predict on unseen data.

Applications

1. Image Processing and Character recognition
2. Forecasting

5. UNSUPERVISED MACHINE LEARNING

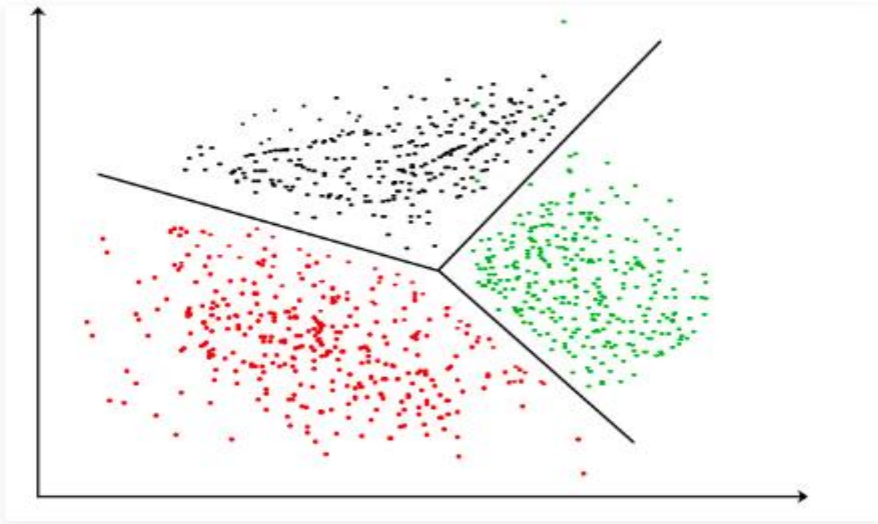
As we Discussed earlier. This is another types of Machine Learning Technique

K-means clustering with scikit-learn

- Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Why Clustering ?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present



Applications

- Marketing
- Biology
- Libraries

- Insurance

Hierarchical clustering

1. Agglomerative Hierarchical Clustering.
2. Divisive Hierarchical Clustering

1. Agglomerative Hierarchical Clustering

In Agglomerative Hierarchical Clustering, Each data point is considered as a single cluster making the total number of clusters equal to the number of data points.

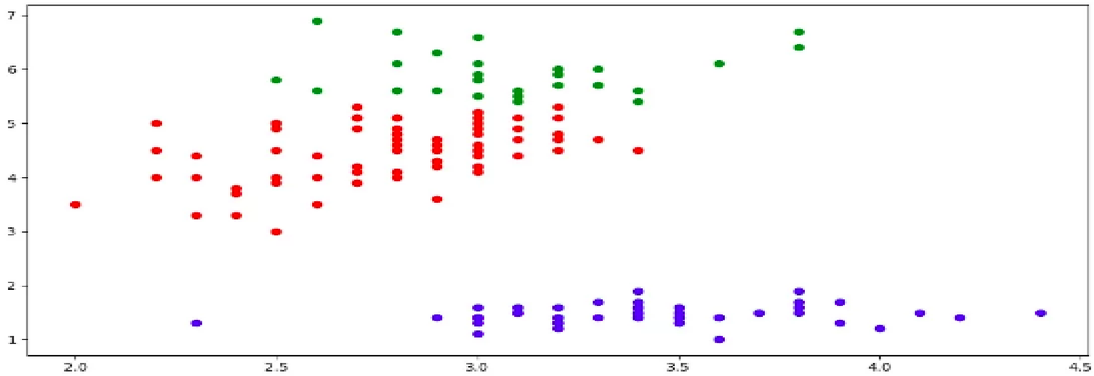
we keep grouping the data based on the similarity metrics, making clusters as we move up in the hierarchy. This approach is also called a bottom-up approach.

2. Divisive hierarchical clustering

Divisive hierarchical clustering is opposite to what agglomerative HC is. Here we start with a single cluster consisting of all the data points.

With each iteration, we separate points which are distant from others based on distance metrics until every cluster has exactly 1 data point.

The Prioritised or hierarchical data shown in the below Graph


7
6
5
4
3
2
1
2.0
2.5
3.0
3.5
4.0
4.5