**The value chain in ai ml**

This chain begins with collecting data, which is the raw material.

There are two ways of accessing data

1**) Primary**

2**)Secondary**

In secondary data the data already exists and we need to figure out where the data can be downloaded. If the data does not exist then we need to collect it. We can create a Google form to collect the data, once the data is collected then we need to crunch the data to get certain information out of it and this information is needed for making decisions that are more accurate, precise, and scientific. This is also known as statistical analysis of data.

**Components of statistical analysis of data:**

1. **Descriptive Analysis**

Under this analysis, we calculate 4 measures.

1. Measure of central tendency
2. Measure of Dispersion
3. Measure of Position
4. Measure of Distribution
5. **Visual Interpretation of Data**

This is done by making a variety of graphs and plots

(i)Pie chart

(ii) Bar plot

(iii)Clustered bar plot

(iv)Scatter plot

(v)Heat Map

(vi)Box Plot etc..

1. **Models in AI ML**

**(i)Predictive Model**

There are a variety of outcomes, which are informed by some predictions.

Important algorithms for making predictions:

(i)linear regression

(ii)logistic regression

(iii)linear discriminant analysis

(iv)CART classification and regression tree

(v)random forest

(vi)artificial neural network

For a good predictive model, we need to make a model which gives us the most accurate results.

**The difference between traditional programming and machine learning**

Data(input)

Program …………………………>Traditional Programming………………..>Output

Data(input)

Output………………………….>Machine Learning………………>Program

In traditional programming, the data is inputted on a prewritten program which gives a certain output while in machine learning the machine learns from the inputted data and keeps on learning from the output at every step till the model becomes accurate.

**Steps of a Machine Learning Model**

**1)Gathering the data**

**2)Preparing that data:** in this step, the unclean data is cleaned so that the machine learns good

**3) Choosing a model:**

**4)Training:** train the machine with the data and the algorithm and this process of training will continue till the machine learns it well.

**5)Evaluation:** we evaluate how well the machine has learned.

**6)Hyperparameter Tuning:** in this step, we tune the parameters of the model to make its learning more accurate.

**7)Prediction:** predict the model.

To make these ML models the two most popular open source tools are **R** and **Python.**

How to install python

[www.anaconda.com](http://www.anaconda.com)

anaconda is a distribution package for statistical analysis of data

double click on the .exe file to start the installation

once the file is installed we will get an anaconda 3 folder with jupyer notebook

jupyter notebook is an IDE(Integrated Design Environment) for writing the code in python.

Tools are of 2 types…..1) License Tool

Eg power bi, tableau, ibm spss, matlab, lisrel,

2)Open Source Tool R,python

The popular ide for working on python are the following:

!)Jupyter notebook

2)PyCharm

3)Spyder

For R IDEs are

1)R Studio

2)R Console

3)R Commander

Library/Package

Python has a lot of libraries and these libraries consist of a variety of functions that takes an input and you can get multiple outputs. Some of the popular libraries in python are: 1)pandas: this library is used for handling a data frame.

2)numpy: also known as numerical python,is used for all the numerical calculations like mean, median, mode, standard deviation etc..

3)SciPy: Also known as scientific python, is used for statistical values like skewness, kyrtosis, correlation etc..

4)Stats models: this library is used for creating variety of statistical models like linear regression,t-test,anoma etc…

5)Sci-kit learn: This library is used for ML models like CART, Random Forest,ANN, NB,Marcov,etc…

Other libraries : Keras,opencv etc

**How to use the libraries:**

Step 1)

Install the libraries in the machine which is a one time process. If the IDE is Jupyter notebook then these libraires are preinstalled

Step 2)

Load the library using the import function in the current notebook of jupyter. Once the library is loaded successfully we can use the functions inside the library

**How to use Jupyter notebook for writing the code in python**

Step 1) To start a new notebook click on new button and then select python3.

A new notebook will give you access to an empty cell in which you can write any number of lines of code.

Leave a single space between any two words in the code and the coding is case sensitive.

In import pandas as pd , pd is called alias

Step 2)how to access the data in jupyter notebook.

To create a variable ID

The name of the data frame is data.

To create ID for the first 10 students.

Create a data frame df1 for the variable gender for the first 10 students.

Create a data frame Df2 for the first 10 students for the variable race

Data dictionary for the file hsb2.csv

Id is the roll no of the student

Gender is whether the student is boy or a girl

What is the data type of the variable gender?

0 is a boy and 1 is a girl

Create a variable df2 for the variable race for the first 10 students.

Create a variable df3 for the variable ses for the first 10 students and report the data type of the variable.

Create a data frame df4 for the first 10 students having the marks in the subject read and report its data type

Create a data frame df5 for the marks in the subject rite and then report its data type.

Create a data frame df6 for the marks in the subject maths for the first 10 students and report its data type

How to save a jupyter notebook

To create a data frame with multiple variables

Create a data frame df7 with two variables id and gender

Ceate a data frame df8 with three variables id , gender and race

How to generate the frequency the table for a variable

freq\_table = pd.crosstab(df[‘id’], ‘count’)

freq\_table

Geneate the fre table for the variable read and then find which marks is scored by the maximum number of students, repeat the same for the other two var wirte and maths

Calculate x+2y and x-y . again use the frequency tablr to calculate the new area this should be greater than equal to 95.5%.

If the first two criteria are met then calculate mean+3 sd and mean – 3sd

X+3y and x-3y and then calculate area under the curve. This should be greater than equal to 99.7%. if all the 3 criteria are satisfied we can conclude that the variable read has a normal distribution. Similarly check the normal distribution for the variable write and math.

How to create a new variable in python.

To create a varible total which is the addition of marks read write nd math

Checl the normality for this new variable total

Create a new variable average which is the mean marks in read write and matths and then check the normality of this new variable average.

How to draw the distribution curve of the variable

Import the library matplotlib

Command:

from matplotlib import pyplot as plt

%matplotlib inline

Make the distribution cirve for the variable read.

Code:

norm\_data = pd.DataFrame(df[“read”])

norm\_data.plot(kind=”density”)

plt.show()

make the distributuin curve for the variables read, maths, wirte and total

How to import a dataset in python:

df = pd.read\_csv(C\hsb2.csv)

df.head()

df.info()

how to find the outliers in the data

outliers are the extreme values in any variable. For a good machine learning model we should remove any outliers if they are found.

To find the outliers in the variable read.

Step 1 ) find q1 (quartile one)

Q1= df[‘read’].quartile(0.25)

Find quartile 3 q3

Q3=df[‘read’].quartile(0.75)

Find inter quartile range

If any value in the read data is less than q1-1.5 times iqr then it is an outlier.

If any value is greater than q3+1.5iqr then it is also an outlier.

Ques 1) find the number of outliers in the variable math and sci

Create a variable total which is the addition of marks in read, write ,math ,sci and socsci and check for the presence of outliers.

Create a var, avg, which is the mean marks in the five subjects and check for the presence of outliers.

How to create a categorical variable for gender where 0 is a boy and 1 is a girl and then find the modal class

Code:

Bins = [-1, 0, 1]

Labels = [“Boy”, “Girl”]

Df[‘Gendar\_gr’] = pd.cut(df[‘Gender’], bins, labels = labels)

Df.head()

Create a categorical variable for the var prog whise name is prog\_gr ug pg phd and then find the modal class.

Create a categorical variable for school type and find the modal class where 1 is public and 2 is private.

Create a histogram for the variable read. , write, mth,sci ,socsci and average

Df[‘read’].hist(figsize=(20,8))

Plt.show()

Create a categorical variable for race and find the nodal class where 1 is north 2 is south stu 3 is east and 4 is western stu

Create a catego variable for ses soco eco bakcg where 1 is low 2 is miidddle and 3 is high and then find the modal class.

How to make bar plot using no of boys and girls

Import seaborn as sns

sns.countplot(x=df['Gender\_gr'])

plt.show()

Make a bar plot for gender,prog,ses,and sestyt