Anaconda pass : hasan242Mh2

What is DATA?

DDDM - Data-driven decision-making (DDDM) is defined as **making decisions based on hard data as opposed to intuition, observation, or guesswork**. The value of data-driven decisions is dependent on the quality of the data and its analysis and interpretation.

Automation & Data science – Automation a intelligence nai, command only but Data science a intelligence ace.

Machine learning – 1 statistical 2 nural network

Two different types of machine learning model:

* Supervised
* Unsupervised

Simple Linear Regression:

linear mean amra Jodi kono problem k akta line r mote akta model dea solve korte chai then seta hobe linear model.

Regression mean real value, output hisabe predict korte hobe

So, linear regression mean line r model dea real value predict kora bujhay

* Model ta baka hole linear regression thakto na seta polynomial regression hoy jeto
* So linear regression holo dependent and independent variable r moddhe relation make korar process matro

In csv file ‘area’ column hocce feature & ‘price’ column hocce level

# linear regression hocce superviced machine learning processs

# feature hoy 2 dimension as input

28/03/22

We are importing 4 libraries

* Pandas
* Matplotlib.pyplot
* Numpy
* Sklearn

Scikit-learn (sklearn): <https://scikit-learn.org/stable/getting_started.html>

# What is statistical modeling?

Sklearn r moddhe linear regression ace

#task – car driving risk analysis - <https://www.kaggle.com/datasets/studymart/car-driving-risk-analysis>

For training purpose 60% data and for test purpose 40% data

Speed r basis a risk analysis korbo

Need to find –

* best fitted line
* Need to predict risk for different speed
* Score
* Accuracy
* M – coefficient
* C – intercept
* Prove y

<https://stackoverflow.com/questions/45627784/unable-to-obtain-accuracy-score-for-my-linear>

29/03/22

# how to plot graph using matplotlib.pyplot library

Plt.scatter(df[‘speed’], df[‘risk’], marker = ‘\*’, color = ‘blue’)

Plt.title(‘Graph title’)

Plt.xlebel(‘speed in km’)

Plt.ylebel(‘risk in percentage %’)

Doing the task – car driving risk analysis

Breaking the task into sequence.

Importing 3 libraries

(pandas, matplotlib, numpy)

Read the csv file

Store features and level in x & y variable

from sklearn.model\_selection import train\_test\_split

xtrain, xtest, ytrain, ytest = train\_test\_split(x,y, test\_size = .40, random\_state = 1)

from sklearn.linear\_model import LinearRegression

store LinearRegression in a variable ‘reg’

reg.fit(xtrain, ytrain)

plt.plot(df.speed, reg.predict(df[['speed']]))

30/03/22

* Data preprocessing korte hobe mean nan value fill up korte hobe. To fill up the nan value we need to find the .mean() or .median() value first. After that we need to store the .median() value in a variable then we need to do df.column\_name = df.column\_name.fillna(variable). That’s how we can fill up the nan data.
* We need to use ‘from sklearn import linear\_model’ this library to use linear\_regression model.
* Linear regression for multiple variables.
* User warning: x does not have valid feature names . problem solved.

31/03/2022

Performance Metrics

* Regression
* R-Squared Value
* Classification
  + - Confusion Matrix
    - Accuracy
    - F1 Measure
    - Recall
    - ROC Curve
* Clustering
  + Elbow Method

Performance Metrics mean R-Squared Value find korbo

Merge to data sets using pandas.

‘pd.merge(data\_frame\_name1, data\_frame\_name2, on =’CustomerID’)’

If we had difference in data sets then by default it’ll create intersection datatsets(common data) so if we need all data we need to decleare how = ‘outer’

‘pd.merge(data\_frame\_name1, data\_frame\_name2, on =’CustomerID’, how = ‘outer’)’

We can also use how = ‘left’ or ‘right’

* ‘pd.merge(data\_frame\_name1, data\_frame\_name2, on =’CustomerID’, how = ‘outer’, indicator = True)’
* If we use indicator = True it’ll show which data are common