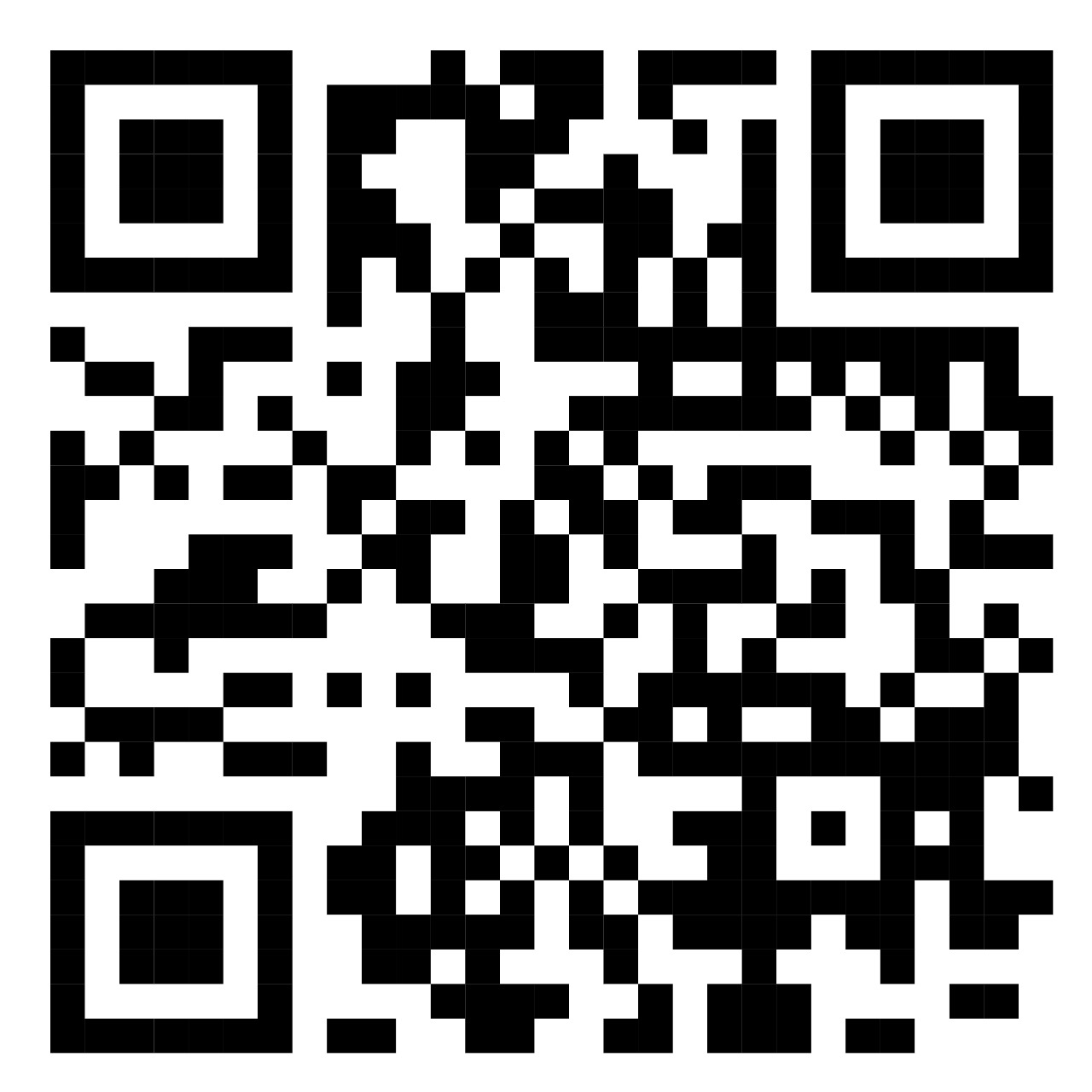
****

**DAY 1:12-02-2024**

IDE - INTEGRATED DEVELOPMENT ENVIRONMENT

Numpy-python library for numerical computing it supports arrays,matrices,scientific computing data analysis machine learning

GOOGLE COLAB

**Array creation-**

array(),zeros(),ones(),arange()

**Array manipulation-**

Reshaping of an Array-reshape(),

Slicing the Array-slice(),

Stacking the array in vertical using “vstack” and horizontal using “Stack”

split()

Transpose of a matrix

**Mathematical operations-**

1)Addition of a Arrays

Element wise addition

Broad cast addition

2)Subtraction of Arrays

3)Multiplication of Arrays

4)Transpose of a matrix

**Statistical operations**

Mean

Median

Variance

Standard deviation

**Dataset-**

Creating a dataset,importing a dataset into google colab and using in a program

Functions-

Loadtxt

savetxt

**Binary classification-**

Binary classification-it has only 2 classes either 0 or 1(T or F)

Multi classification- +ve(1),-ve(0),neutral(0.5)

**Confused matrix-**4 classes(11,00,10,01)

i)completely true

ii)not completely true

iii)not completely false

iv)completely false

Data preprocessing-

Data cleaning

Data training

**DAY-2 13-02-2024**

**Module-2-Pandas**-Library used for working with datasets.It has functions for analysing, cleaning,exploring,and manipulating data.”pandas” has a reference to both “Panel Data” and “Python Data Analysis”.

**Applications of Pandas:**

1. Data cleaning and preparation
2. Data analysis
3. Data visualisation
4. Machine learning

**Importing Pandas Library:**

Import pandas as pd(we shorten the imported name to pd for better readability of code using Pandas)

Assigning indices(values) to the list elements using Series method

var\_nam=pd.Series(list\_name,index=[values\_to\_be\_assigned])

**Loading.csv(comma separated value file) to our program**

df=pd.read\_csv(“file\_name.csv”)

To print the file data we use print(df) # df stands for dataframes\_variable

* To print top n rows in the dataset we use df.head(n) #n = no. of rows
* To print bottom n rows in the dataset we use df.tail(n) #n = no. of rows

**Functions in .csv files:**

min(),max(),mean(),concat()

**Loading.txt(text file) to our program**

df=pd.read\_csv(“file\_name.txt”,sep=” “)

We can print a particular index location in the dataset by using .loc() function

print(df.loc[1]) #index 1 portion

print(df.loc[2:4]) #index 2 to 3 portion by slicing operator

**Loading.xlsx(Excel spreadsheet file) to our program**

df=pd.read\_excel(“file\_name.xlsx”)

To print particular sheet in the .xlsx file we specify the sheet number as:

df=pd.read\_excel(“file\_name.xlsx”,sheet\_name=n)

To delete the duplicate lines in the .xlsx file

df=df.drop\_duplicates()

**Matplotlib Library Functions:**

1. .scatter() - a visual representation of how 2 variables relate to each other.
2. .plot() - draws a line from point to point.
3. .subplot() - provides a way to plot multiple plots on a single figure.
4. .title() - to specify the title of the visualisation depicted.
5. .show() - provides the figure,which contains all the plot elements.
6. .legend() - used to place a legend on the axes
7. .pie() - draws a pie-shaped wedge bounded by the intersection of an ellipse and 2 radials.

plt.plot(x=’X’,y=’Y’,color='b',linestyle='-',label='Line Plot',marker='o')

- x='X' and y='Y' specify the columns to be plotted on the x and y-axes.

- marker='o' adds circular markers to data points.

- linestyle='-' specifies a solid line connecting the points.

- color='b' sets the color to blue.

- label='Line Plot' adds a label for the legend.

**DAY-3:14-02-2024**

**Seaborn**

python data visualisation library built on top of matplot.it provides a high-level interface for drawing attractive and informative statistical graphics.

**Applications of Seaborn:**

Making statistical graphics in python.

**Importing seaborn Library:**

Import seaborn as sns(we shorten the imported name to sns for better readability of code using seaborn)

**list of all built-in datasets in Seaborn:**

* **anagrams:** A dataset containing lists of anagrams.
* **anscombe:** A classic dataset containing four sets of bivariate data.
* **attention:** A dataset containing reaction times and error rates for a visual attention task.
* **brain\_networks:** A dataset containing functional connectivity matrices for different brain networks.
* **car\_crashes:** A dataset containing information about car crashes in the US.
* **diamonds:** A dataset containing information about diamonds.
* **dots:** A dataset containing the locations of dots on a grid.
* **exercise:** A dataset containing information about exercise and physical activity.
* **flights:** A dataset containing information about flights.
* **fmri:** A dataset containing fMRI data from a study of working memory.
* **gammas:** A dataset containing gamma ray burst data.
* **iris:** A classic dataset containing information about iris flowers.
* **mpg:** A dataset containing information about fuel economy in cars.
* **planets:** A dataset containing information about planets.
* **tips:** A dataset containing information about tips at a restaurant.
* **titanic:** A dataset containing information about passengers on the Titanic.

To load a dataset, you can use the load\_dataset() function from Seaborn.

var\_name = sns.load\_dataset("dataset\_name")

**Seaborn Library Functions:**

1. .scatterplot() - used to create a scatter plot.that displays individual data points on a 2D graph.
2. .violinplot() - create a violin plot. A violin plot is similar to a box plot but also includes a rotated kernel density plot on each side.
3. .corr() - used to calculate the correlation b/w numerical columns in a dataframe.
4. .heatmap() - used to visualise a matrix-like dataset such as correlation matrix.

Example:

sns.heatmap(corr\_matrix, annot=True, cmap="coolwarm")

- sns.heatmap() is used to create a heatmap of the correlation matrix.

- annot=True adds numerical annotations to each cell in the heatmap.

- cmap="coolwarm" sets the color map to cool to warm shades.

**MODULE-3 : MACHINE LEARNING-data analysis**

.

Machine learning -3 types

Imp are 2 steps

1.training data

2.Testing data

Types of Machine learning

1.Supervised machine learning

2.Unsupervised machine learning

3.Semi Supervised machine learning

Frequently used words in ML are Label,Features

Labelled data -Supervised learning

Unlabeled data-Unsupervised learning

Combination of Labelled & unlabeled–semi

Teachable Machine(website)-image project

->name classes(yes or no and more no.of class)

->upload images from kaggle dataset

->click on training Model

->for code click on export model if you to save then click on upload model it will be saved in teachable machine cloud and generate a click ,copy the link and paste in new tab you will see the project working

->to test the project without uploading on webcam or add file from your pc

->it will show the result according to class Yes or No

Machine learning for kids(website)->click on “try it now”(without registering you can use it)->free version available for 3 to 4 hrs

->give name

->select type

NEURAL NETWORK: inter-connection of the neurons

Algorithm

1.cnl-specific for image project

2.rnl-text project

CNN(Convolution Neural Network)

A convolutional neural network(CNN) is a type of deep learning.neural network architecture commonly used in computer vision.computer vision is a field of artificial intelligence that enables a computer to understand and interpret the image or visual data .

1.Input layer-to extract images

2.hidden layer

3.output layer-output depends on probability score

**Hidden layer**

1.Convolution layer-extract features from the input

2.Activation layer-applies activation function for the output by calculating weighted sum & bias

3.Pooling layer-reduces memory,size & fast computation

i)Max

ii)Avg

iii)Min

**Activation Function**

->It decides whether the neuron is activated or not by calculating the weighted sum and bias

->It’s purpose is element wise operation

1.Sigmoid function

2.Tanh

3.RELU

4.Softmax

ML

1.Classification-cnn

2.Regression -linear,logical

**DAY-4 15-02-2024**

**Linear Regression**

->Also a type of ML algorithm

->supervised ML algorithm

->learn from labelled dataset and maps the data points to the most optimised linear functions

->these points can be used for prediction on new datasets

Dependent & Independent variable

In Chemical reaction

-Mass of product increases with time

Mass-dependent-x

Time-Independent-y

Graph for above is Hyper plane in maths

Find the mean of both dependent and independent variable

X’-mean=5+7+12+16+20/5(ind)

Y’-mean=40+120+180+210+240/5(dep)

Find difference between each x point and x-mean(X-x’)

Find difference between each y point and y-mean(y-y’)

Find sum of squares of (x-x’)^2 sum(x-x’)^2

Find product of (x-x’) and (y-y’)

a=y’-bx’

b=Sxy/Sxx

=sigma(xi-x’)(yi-y’)/sigma(xi-x’)^2

**Logistic Regression**

->Used for Binary Classification

->Supervised learning

P>0.5,its a goal-classA(y=1)

p=<0.5,its a miss-cclassB(y=0)

PCA-Principle Component Analysis

Converts Sensible data into hideable data

Define data generator for train,validation and test sets

Imagedatagenerator-preprocessing

To convert image into rescale image,RGB=1./255

Train\_generator

Val\_generator

Test\_datagen

model.compilation(optimizer=’adam’,loss=’binary\_crossentropy’,matrics=[‘accuracy’])

history=model.fit(train\_generator,valdation\_data=val\_generator,epoch=5)

model.save(“Model.h5”,”label.txt”)

.h5-converts a file into weighted file

Test your image

PIL-pillow(perform operations on image)

Load the model

Classes

Resizing the image to be 224\*224

Turn the image into a numpy array

Normalize

Load the image into array

Sequential

Simple rnn,dense,embedding

tokenizer

Pad sequences

to\_categorical

Predict the model

Tokenizing(word s are tokens)

.fit

len#total words

Creating input sequence and their corresponding next words

Padding sequence for consistent input size(entire sequence is considered analyzed)

Maximum sequence length categorical\_classify the text

Building a simple rnn model

Sequential,embedded,simple rnn,dense

Compilng the model

Training the model

Generating test using train model

Word

No.of words to print

**Day-5 :16-02-2024**

**Decision trees and Random forest**

**Decision trees-**

* Effective method for making decisions because they layout the problem and all the possible outcomes.
* Have nodes and leaves
* Node:Condition having the true and false branches
* Leaf:showing the dataset that is true and false to the condition
* Recurrently classify the data until pure node or pure leaf

**Random forest-**

* Collection of many decision trees
* keywords:Bootstrapping,aggregation
* Random forest create new dataset from original dataset
* This is called bootstrapping of data.
* Conditions for bootstrapping-

Select rows that are different and cover all the rows

Select features that are different and cover all features

Selecting same data(rows) and same features have no use

Prediction of new dataset-aggregation

Prediction is based on different decision tree

SVM-Support Vector Machine

**DAY 6:17-02-2024**

**Statistical analysis using SciPy and statsmodels for hypothesis testing,regression analysis and ANOVA**

Scientific computation-SciPy

SciPy is an open source scientific computing library for Python that builds on Numpy.It provides additional functionalities compared to Numpy,including optimization,integration,interpolation,eigen value problems,signal and image processing,statistical distribution and much more.

integrate.quad(lambda x: x\*\*2, 0, 1)

Integration of x^2 with limits 0 to 1

f(x)=ax^2+bx+c

Minimum function= -b/2a

=-3/2

x=independent,

y=dependent

Consider any 2 points

y1+[((x-x1)(y2-y1))/(x2-x1)]

X is called target variable

x1=2,x2=3,y1=1,y2=3,x=2.5

Statsmodels is a Python library that provides classes and functions for estimating and testing statistical models.It is built on top of Numpy,SciPy and Matplotlib and integrates with Pandas for data handling.Statsmodels includes a wide range of statistical models and tests making it powerful for testing

loc=mean,scale=standard deviation

**Hypothesis Testing**

1.This example code performing t-test to compare their mean value of two groups

2.Select scores from group\_A it filters the data ,dataframe to include only rows where the treatment column is A then select the score column.

**Performing the t-test**

stats.ttest\_ind() performs on independent samples t-test that means two sample tests to compare the mean value.

t\_stat →variable that stores t-statistical which measures the difference between mean of the two groups relative to the variation within the group.

p\_value →variable stores p value it consists of probability of observing the given result (less or more or extreme) under the null hypothesis(there is no difference between means of two groups)

**ANOVA(one-way ANOVA testing)**

1.Scores extraction for separate treatment category

stats.f\_oneway()-->perform one way procedure for ANOVA testing

x=mean value addition with standard deviation and finally standard deviation with multiple value of (target variable)

ANOVA(analysis of variance)-->it is the operation for performing two or more group to determine if the values are statistically significant or not

→this testing is extension on hypothesis testing t-test allowing comparison of mean value across multiple groups simultaneously.

**Step-1:1.**Null hypothesis

2.Alternative hypothesis→ at least one group mean is different from others

3.Assumption

i)independent observation→data points within each group or independent of each other

ii)Normality→each group data follows a normal distribution

iii)Homogeneity of variance-the variances of the data in each group are equal

**ANOVA testing statistical analysis**

f\_stat→It is calculated based on the ratio of the between group variability to the within group variability.If the f\_static is larger that means greater than expected null hypothesis it suggest that the group means are not equal and there are significant differences between the groups.

p\_value→probability of observing the given result (normal result,extreme result) under the assumption null hypothesis which is equal to true.a small p value (value less than 0.05) it suggested that the difference between group mean value are unlikely to have occurred by chance alone,leading to the rejection of null hypothesis.

You will receive f\_statistic value and p\_value

→Based on the p value you can determine whether there are significant differences between the means of the three groups

→If the p\_value is less than the significance level, for example 0.05,you reject the null hypothesis value and conclude that there are significant differences between at least one pair of group means, otherwise you fail to reject the null hypothesis.

Tukey’s HSD(honestly significant difference method)

Confident interval=mean difference +or- 1.96 x standard error

1.96–critical value for 95% confidence interval

**Regression Analysis(simple linear regression)**

**Step-1:**Preparing data

Treatment variable is converted from categorical data to numerical format

sm.add\_constant()-->intercept-constant value it will allow the model to estimate the intercept parameter

**Step-2:**Creating the regression model

sm.OLS()→OLS-ordinary least square it will define OLS for regression model

data[‘Score’]-- is the dependent variable(response variable) it representing the test scores

X denotes independent variable (predictor variable) representing the treatment group

**Step-3:**Fitting the model

Regression model to the data it will fit coefficients(slope and intercept) by using OLS method.

**OLS**

1.OLS is the function is from stats model

**Step-1:**Sample data

x,y independent dependent variable

Adding a constant

Intercept-independent variable-array x

**Step-2:**It is necessary for estimating the intercept parameter in the linear regression model

Creating the OLS model

OLS regression it takes dependent variable y and independent variable x(input)

OLS model it will producing the coefficient including intercept

Time series analysis of Numpy or any library

**DAY 7:19-02-2024**

**WEB SCRAPING**

Web Page is written in HTML

Anything in webpage is element

Every element has special id and class

Scraping-stealing,copying,updating information data from a webpage

Beautiful soup library -helps to scrape data from webpage

selenium-automatically launch data

→loading data from websites

→unstructured in HTML(i.e..,output is HTML code)

→convertible into spreadsheets/DB

→major websites have their APIs for web scraping

→Can be saved in .csv file using (to\_csv)

Scrapper

→extract all the data on particular web site

→specific data that a user wants

Steps for web scraping

1.collect Url

2.you will get HTML code

3.elements(CSS/JS)(i.e.., take req elements)

4.Scrapes the required data

5.Saves it in required format(csv,xlsx,json)

Applications

1.email marketing

2.sentiment analysis

3.news monitoring

4.Market research

Libraries

1.beautiful soup-from bs4

2.request

3.pandas

3.webdriver

4.webdriver\_manager

5.selenium

BeautifulSoup

Used to scrape data from static website

Package bs4:subpackage BeautifulSoup

Function Purpose Attributes

beautifulSoup()--to extract html code from a webpage– .text

Html

find()--to find the first element of a kind–(‘element\_name’)

find\_all()-to find all elements of a kind–(‘element\_name’)

Requests

→used to send request to a web page

→get(‘url’)

Project 1

Extracting HTML code of any website

Program flow

→import libraries

→save the url

→using requests.get(),access the webpage

→using BeautifulSoup(),access the HTML code

Automatic web scraping using Selenium

→used to scrape data from dynamic website

→subpackage:webdriver

Function Purpose Attributes

Chromeoptions[]--create an instance of chrome

.get—access a webpage—‘url’

Find\_element--to find first element of a kind--By ID

ByxPATH

.click()--to click a button in webpage

Keys-push any string in text box

Id of the search box in amazon

id=”twotabsearchtextbox”

Search button=id=”nav-search-submit-button”

//tag(text()=”) #accessing text

//tag[@class=’’] #accessing a class

eg://span[@class=’a-size-medium a-color-base a-text-normal’]

Scikit learning

Tensorflow

**DAY8:-20-02-2024**

**APPLICATION PROGRAMMING INTERFACE(API)**

Using beautiful soup scrape the headlines from any news website(BBC,Times India) and save it in a dataframe

Scrape the prices of any 10 products from two different shopping websites(product name,prices).save both the results in two different dataframes.merge the data frame and compare the price of each product.In the code get the product name from the user,output should be flipkart has the lower price or amazon has the lower price.

Scrape IMDB website for horror movies on the parameters ratings,name of the movie(tv shows),small summary of that movie

**DAY 9:20-02-2024**

Project