WEEK 1

1. Artificial Intelligence (AI) is promising cutting-edge technology providing intelligent solutions in all sectors today. Define AI and describe applications of AI in different domains
2. Groups of developers want to work Collaboratively on big software project. Each developer in team is assigned one module. How git and GitHub help these developers to build project effectively?
3. Summarize the challenges associated with Machine Learning
4. How AI Software Development life cycle differs from traditional software development? Explain
5. Summarize any two cloud deployment models
6. Write steps to Create repository in GitHub and add file.
7. Is data which is collected by various applications ethical in nature? Justify your answer

WEEK 2

1. How Big data is different from the data stored in traditional databases? Elaborate
2. Differentiate between supervised machine learning and Unsupervised machine learning
3. Explain different machine learning types
4. Discuss various attributes of a high-quality data

WEEK 3

1. Here given the bating statistics of Indian great cricketer Anil Kumble. Perform the following operations-
   i) Aggregation
   ii) Grouping
   iii) Time series
   iv) Filter
   v) Vectorized

2. Create a dataframe with following data.

	First Name	Last Name	Type	Department	YoE	Salary
0	Aryan	Singh	Full-time Employee	Administration	2	20000
1	Rohan	Agarwal	Intern	Technical	3	5000
2	Riya	Shah	Full-time Employee	Administration	5	10000
3	Yash	Bhatia	Part-time Employee	Technical	7	10000
4	Siddhant	Khanna	Full-time Employee	Management	6	20000

   a) Make a pivot table which shows average salary of each type of employee for each department.
   b) Make a pivot table which shows the sum and mean of the salaries of each type of employee and the number of employees of each type.
   c) Make a pivot table which shows standard deviation for salary column.

3. Create two series as shown using pd.series() function.
Series_A = [10,20,30,40,50] Series_B = [40,50,60,70,80].
i. Get the items not common to both.
ii. Identify the smallest and largest element in the series A
iii. Find the sum of series B
iv. Calculate average in the series A
v. Find median in the given series B

4. Perform the following operations on Car manufacturing company dataset auto-mpg.csv given below using pandas

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
1	15	8	350	165	3693	11.5	71	1	buick skylark 320
2	18	6	318	150	3436	11.0	70	1	plymouth satellite
3	16	4	304	150	3433	12.0	80	1	amc rebel sst
4	17	8	302	140	3449	10.5	70	1	ford torino

Read data from an existing file

b) statistical details of dataset

c) Get all cars with 8 cylinders

d) Get the number of cars manufactured in each year.

5. Lee decides to walk 10000 steps every day to combat the effect that lockdown has had on his body's agility, mobility, flexibility and strength. Consider the following data from fitness tracker over a period of 10days

Day number	Steps walked
1	6012
2	4079
3	6386
4	5230
5	4598
6	5564
7	6971
8	7763
9	8032
10	8569

i) Represent the above data in a 10x2array. In each row, the first element should contain day number and second element should contain steps walked.

ii) Lee notices that the tracker's battery dies every day at 7 pm. Lee discovers that on an average, he walks 2000 steps every day after 7 pm. Perform an appropriate

iii)Write a program that returns the steps walked if the steps walked are more than 9000.

iv) Print an array containing steps walked in sorted order.

v) Perform an appropriate operation to add 1000 steps to all the observations using pandas

vi) Find out the days on which he walked more than 7000 steps using pandas

6. Write python code to explain map (), filter (), reduce (), lambda()

7. Assume Iris dataset and write the code

a. print first 5 record

b. print the size of the data for given data set

b. Use scatter plot to compare petal length and petal width

c. check for missing values

b. print summarizes of the dataset

e. Count plot for the spices

c. Visualize the distribution of any one column

d. Visualize the relationship between any two variable

e. Print the information of all column in the dataset

f. Visualize the spices column using bar graph

WEEK 4

1. The statistical summary of titanic dataset with respect to age and fare is as follows. Analysis and explain statistical matrices from below summary. And find the difference between Covariance and Correlation.

	Age	Fare
count	714.000000	891.000000
mean	29.699118	32.204208
std	14.526497	49.693429
min	0.420000	0.000000
25%	20.125000	7.910400
50%	28.000000	14.454200
75%	38.000000	31.000000
max	80.000000	512.329200

2. Difference between Covariance and Correlation.
3. Describe the univariate and multivariate analysis.
4. Justify the Significance of Exploratory Data Analysis?
5. A data set is given to you creating a machine learning model. What are the steps followed before using the data for training the model? Elaborate each step.
6. Explore different types of data in machine learning.

WEEK 5

1. For the given data set Perform the following operations:

i) Check statistical info of the data set
ii) Plot a line plot showing total profit on y axis and number column on x axis
iii) Find the missing values
iv) Find the sum of total profit
v) Find the max value from Drawing sheets column

number	Pencil	textbooks	Drawing sheets	Total units	profit
1	300	250	100	700	80000
2	350	350	125	1075	9500
3	400	400	190	1320	10256
4	500	420	210	1510	12000
5	520	500	250		15000

2. How to handle the missing values in the dataset? Explain.
3. Explain different challenges involved in Data Integration.
4. How to handle the outliers in the dataset? Explain
5. Write python code to imputation the missing values in the dataset using, mean and median method
6. A company wants to study iris dataset to make predictions. However, the data gathered is not clean for analysis. The company requests you to write a python code to perform the following operations for data driven competitive

advantage (Assume dataset with missing values)

- Check for missing values
- Replace missing values with mean value

7. A company has collected customer comments on its products, rating them as safe or unsafe, using decision trees. The training dataset has the following features: id, date, full review, full review summary, and a binary safe/unsafe tag. During training, any data sample with missing features was dropped. In a few instances, the test set was found to be missing the full review text field. For this use case, which is the most effective course of action to address test data samples with missing features. Justify

WEEK 6

1. Assume a Boston housing dataset having two column built_up_area(independent variable) and rent(dependent variable). Build linear regression model
2. Assume a Boston housing dataset having two column built_up_area(independent variable) and rent(dependent variable).
    a. Import libraries
    b. Read data
    c. Write scatter plot to compare to show relationship between area and price
3. Assume that you are given a train data set having 1000 columns and 1 million rows the data set in based on a classification problem. your manager has asked you to reduce the dimension of this data so that model computations time can be reduced, your machine has memory constraints what would you do? (you are free to make practical assumptions).
4. For Breast cancer dataset build a machine learning model to predict or identify it and to perform the following operations
    i) import libraries
    ii) Perform pre-processing
    iii) Split the data set
    iv) Find the accuracy
    vi) Data prediction
5. Assume employee data with two column years of experience (independent variable) and salary (dependent variable).
    i) import libraries
    ii) Perform pre-processing
    iii) Split the data set
    iv) Find the accuracy
    vi) Data prediction
    vii) Build and test model
6. Perform comparative analysis and find model accuracy by splitting data for training and testing on iris dataset using K-means clustering and logistic regression.
7. Build a random forest model on diabetic's dataset and check accuracy.
8. Illustrate meaning of supervised learning and its types of classification
9. Compare overfitting with underfitting
10. Do comparative analysis by splitting data for training and testing on iris dataset
11. Discuss different techniques of cross validation
12. Demonstrate Simple Linear Regression considering a dataset that has two variables: salary (dependent variable) and experience (Independent variable).
13. Build a machine learning model and find the salary if years of experience is 3, test score and interview score is 18.

Exp	Test marks (20)	Interview marks (20)	Salary (Rs)
2	20	15	15,000
3	15	13	20,000
NAN	12	16	10,000
5	19	10	50,000

WEEK 7

1. Explain the evaluation matrix for classification as follows.
    1) confusion matrix
    2) Accuracy
    3) F1 – score
    4) AUC_ROC
    5) Precision and Recall

2. Build a data set and predict the heart disease based on BP, Sugar, Age, Gender and Cholesterol by using relevant operations

3. The confusion matrix for a model is as shown below. Evaluate accuracy, precision, recall, Specificity and F1-Score, AUC-ROC

		Actual: 1	Actual: 0
Predicted	1	397	103
	0	126	142

4. Given a dataset perform comparative analysis using decision tree and SVM algorithm and check accuracy

5. A machine learning model was built to classify people based on whether they speak English or Hindi. The confusion matrix for the model is as shown below. Compute accuracy, precision, recall, F1-Score and Specificity.

		Actual: English Speaker	Actual: Hindi Speaker
Predicted	English Speaker	109	11
	Hindi Speaker	8	56

WEEK8

1. Define clustering and compare various clustering techniques.
2. Describe the advanced ensemble techniques
3. How to Choose the Right Number of Clusters in k-means clustering? Explain any one method
4. Cluster the following eight points (with (x, y) representing locations) into three clusters:

    A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

    Initial cluster centres are: A1(2, 10), A4(5, 8) and A7(1, 2).

    The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as

$P(a, b) = |x2 – x1| + |y2 – y1|$.

Use K-Means Algorithm to find the three cluster centres after the first iteration

5. Compare classification algorithms with clustering algorithm
6. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why?
7. The sinking of the Titanic is one of the most infamous shipwrecks in history. You are asked to build a machine learning model to predict whether a passenger survived or not. Describe each step you will follow to build this model. Description of dataset titanic.csv is as below

Name	Variable explanation
pclass	Passenger Class (1 = 1st;2 = 2nd;3 = 3rd)
Survived	Survival (0 = no, 1 = yes)
Name	Passenger name
Sex	Gender of passenger
Age	Age of passenger
Sibsp	(number of siblings/spouses aboard)
Parch	(number of parents/children aboard)
Ticket	Ticket number
Fare	Passenger fare (£)
Cabin	Cabin
Embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
Boat	Lifeboat
Body	Body Identification Number
Home.dest	Home/Destination

8. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why?
9. Compare Bagging and Boosting.

WEEK9

1. Discuss importance of dimensionality reduction in machine learning
2. Explain dimensionality reduction using PCA

WEEK 10

1. Discuss activation functions in Neural Network
2. Explain neural network architecture
3. Differentiate between forward propagation and Back propagation

WEEK 11

1. For the given sentence “Machine Learning is best platform. To group over selves”. Visualize the data for the better understanding and perform the following operations.
   - **i)** Word tokenization
   - **ii)** Sentence tokenization
   - **iii)** Dropping Stop word
   - **iv)** Stemming
   - **v)** Lemmatization
2. Companies monitor their call center agents’ live phone interactions or chat sessions with customers in real-time. Call duration with speech recognition automatically detects customer emotions. Companies can better

understand how customer satisfaction varies by product and call center services. Using this scenario Explain the applications and working of sentiment analysis in business.

3. You are working on an NLP model. So, you are dealing with words and sentences, not numbers. Your problem is to categorize these words and make sense of them. Your manager told you that you have to use embeddings. Explain Count Vector and TF-IDF Vector
4. List and explain tools used to deploy the AI model using google cloud platform
5. Demonstrate stemming and Lemmatization concepts with suitable examples
6. Write different algorithms used in sentimental analysis with neat diagram.

WEEK 12

1. What are ethics in AI and Why ethical practices should be followed while developing solutions using AI?
2. Write Ethical challenges in Artificial Intelligence
3. Create container and build docker image using docker file
4. Imagine a scenario where the developer's code is only working on his machine and he has to set it up in the machine of the testing team for every build. How docker containers solve this problem.
5. How will you deploy a trained machine learning model as a predictive service in a production environment? Explain.
6. Understand the difference between DevOps and MLOps and justify which is best suggested process for developing Machine Learning Models.
7. N-grams are defined as the combination of N keywords together. Consider the given sentence: "AI is simulation of human intelligence by machine. It includes expects system NLP speech recognition and machine vision"
    - Generate bi grams for the above sentence
    - Generate tri-grams for the above sentence
8. Summarize different strategies of production deployment
9. What are MLOps? brief different stages that are involved in the MLOps lifecycle
10. With a neat diagram explain components of docker.
11. For the following scenarios you are required to build a predictive model. Which machine learning technique/ algorithm can be applied / best suited for stated problems. Justify your recommendation.
    a. Predicting the food delivery time
    b. Predicting whether the transaction is fraudulent
    c. Predicting the credit limit of a credit card applicant
    d. To group similar customers of an online grocery store, based on their purchasing patterns, to offer discounts to its customers.
    e. Predict the probability of a mechanical system breakdown, based on its system vibration and operating temperature
12. For the below given scenarios, suggest best suited cloud deployment model and list the challenges with it.
    1. For ,
        a. Variable workload
        b. Test and Development
    2. For,
        a. Cloud bursting
        b. On demand access
        c. Sensitive data
13. How will you deploy a trained machine learning model as a predictive service in a production environment. Explain.