V Semester Diploma Examination, February/ March 2023

Artificial Intelligence and Machine Learning (20CS51I)

SCHEME OF VALUATION

Section-I

1a. Summarize the challenges associated with Machine Learning.

Any 5 relevant challenges can be considered 1*5=5M

1b. How AI Software Development life cycle differs from traditional software development?

Any five differences 1*5=5M

1c. Perform the following operations on Car manufacturing company dataset auto-mpg.csv given below using pandas. 10M

a) Read data from an existing file 2.5M
b) statistical details of dataset 2.5M
c) Get all cars with 8 cylinders 2.5M
d) Get the number of cars manufactured in each year. 2.5M

2a. Differentiate between supervised machine learning and Unsupervised machine learning.

Any five differences 1*5=5M

2b. Explore different sources of big data in machine learning . 5M

Writing any 5 sources 1*5=5M

2c. Ramesh 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 10 days. 10M

Code to add1000 steps to all the observations. 5M

Code to find out the days on which Ramesh walked more than 7000 steps. 5M

Section-II

3a. How to handle missing values in the dataset? Explain. 10M

List approaches to handle the missing values - 2M + Explanation of any four methods.(2*4=8M)

3b. A dataset is given to you for creating machine learning model. What are the steps followed before using the data for training the model? Elaborate each step. 10M

Listing of steps 2M+Explaination of any four steps (2*4=8M)

4a. 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) 10M .

Check for missing values 5M + Replace missing values with mean value 5M

4b. Statistical summary of Credit card dataset is as follows. Analyze and explain statistical metrics from above summary. 10M

Analysis of given summary 2M + Explanation of all statistical metrics 8M

Section-III

5a. A Machine learning model was built to classify patient as COVID +ve(1) or – ve(0). The confusion matrix for the model is as shown below. Evaluate accuracy, precision, recall, specificity and F1-Score. 10M

Accuracy – writing formula 1M+calculation 1M=2M
Precision- writing formula 1M+calculation 1M =2M
Recall -writing formula 1M+calculation 1M =2M
Specificity- writing formula 1M+calculation 1M =2M
F1-Score -writing formula 1M+calculation 1M=2M

5b. Compare overfitting with underfitting. **(1*5=5M)**

5c. How to Choose the Right Number of Clusters in k-means clustering? Explain any one method

(Listing methods 1M+Explanation 4M)

6a. 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. 10M

Calculating the distance from each point to all centres -5M
Finding which point belongs to which cluster and form new clusters 3M
Calculate the new cluster centre 2M

6b. Compare classification algorithms with clustering algorithm. 5M

Any five comparisons 1*5=5M

6c. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why? 5M **True(1M)+Explanation 4M**

Section IV

7a. N-grams are defined as the combination of N keywords together. Consider the given sentence:
"The greatest glory in living lies not in never falling, but in rising every time we fall"
a. Generate bi grams for the above sentence
b. Generate tri-grams for the above sentence

1. Generate bi grams (1M import library +4M for generating and printing output)
2. Generate tri-grams (1M import library +4M for generating and printing output)

7b. Discuss importance of dimensionality reduction in machine learning. 5M

Any five valid points1*5=5M

7c. Summarize different strategies of production deployment. 5M

Listing of deployment strategies 1M+Explaination 4M

8a. Demonstrate Stemming and Lemmatization concepts with suitable examples. 10M

Stemming - Explanation 2.5M +Example 2.5M
Lemmatization - Explanation 2.5M +Example 2.5M

8b. Discuss different techniques of cross validation. 5M

Listing of techniques 1M+ explanation of any two 4M

8c. What are MLOps? brief different stages that are involved in the MLOps lifecycles 5M

(Definition 1M+ Explanation of MLOps lifecycle, any four stages 4M)

Section V

9a. With a neat diagram explain components of Docker . 10M

Diagram 2M+ Docker component explanation 8M

9b. Demonstrate Simple Linear Regression considering a dataset that has two variables: salary (dependent variable) and experience (Independent variable) 10M

Importing lib 1M +Reading Dataset 1M + pre-processing 2M+Split, build model 5M +finding score 1M

10a. Summarize any two cloud deployment models . 10M

Any two cloud deployment models 5*2=10M

10b. Discuss any five ethical challenges in AI. 10M

Any five ethical challenges 5*2=10M

V Semester Diploma Examination, February/ March 2023

Artificial Intelligence and Machine Learning (20CS51I)

MODEL ANSWERS

Section-I

1a Summarize the challenges associated with Machine Learning 5M

1. Inadequate Training Data and Poor quality of data

The major issue that comes while using machine learning algorithms is the lack of quality as well as quantity of data. Although data plays a vital role in the processing of machine learning algorithms, many data scientists claim that inadequate data, inaccurate data, noisy data, and unclean data are extremely exhausting the machine learning algorithms. This leads to less accuracy in classification and low-quality results. Hence, data quality can also be considered as a major common problem while processing machine learning algorithms.

2. Non-representative training data

To make sure our training model is generalized well or not, we have to ensure that sample training data must be representative of new cases that we need to generalize. The training data must cover all cases that are already occurred as well as occurring.

3. Overfitting

Overfitting is one of the most common issues faced by Machine Learning engineers and data scientists. Whenever a machine learning model is trained with a huge amount of data, it starts capturing noise and inaccurate data into the training data set. It negatively affects the performance of the model. The main reason behind overfitting is using non-linear methods used in machine learning algorithms as they build non-realistic data models.

4. Underfitting:

Underfitting is just the opposite of overfitting. Whenever a machine learning model is trained with fewer amounts of data, and as a result, it provides incomplete and inaccurate data and destroys the accuracy of the machine learning model.

5. Monitoring and maintenance

As we know that generalized output data is mandatory for any machine learning model; hence, regular monitoring and maintenance become compulsory for the same. Different results for different actions require data change; hence editing of codes as well as resources for monitoring them also become necessary.

6. Lack of skilled resources

Although Machine Learning and Artificial Intelligence are continuously growing in the market, still these industries are fresher in comparison to others. The absence of skilled resources in the form of manpower is also an issue. Hence, we need manpower having in-depth knowledge of mathematics, science, and technologies for developing and managing scientific substances for machine learning.

7. Deployment: Deploying ML models in production environments can be difficult, as it requires expertise in both ML and software engineering.

8. Model Selection: There are many different ML algorithms to choose from, and selecting the right one for a particular problem is challenging.

9. Feature Engineering: Extracting useful features from raw data is a critical step in ML, but it can be difficult to identify the most relevant features.

1b. How AI Software Development life cycle differs from traditional software development?5M

1. Data-Driven: AI software development is heavily dependent on data and requires large amounts of high-quality data to train and test models. In traditional software development, data is often an afterthought.

2. Experimentation and Iteration: AI software development often involves a lot of experimentation and iteration, as different algorithms and approaches are tried and tested to see which ones work best. Traditional software development is typically more linear and follows a specific plan or design.

3. Model Selection: In AI software development, selecting the right model for a particular problem is critical and can be a time-consuming process. In traditional software development, the choice of algorithms and techniques is often predetermined.

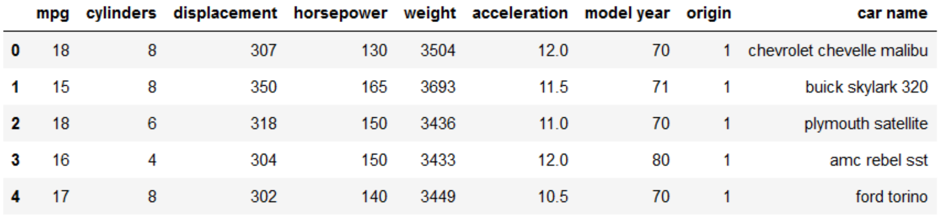
4. Model evaluation and performance: In AI software development, model performance is evaluated using different metrics and techniques, such as accuracy, precision, recall, and F1 score. In traditional software development, model evaluation is often based on functional requirements.

5. Deployment and Maintenance: AI software deployment and maintenance requires additional considerations, such as retraining models over time and deploying them in production environments. In traditional software development, deployment and maintenance are often simpler and more straightforward.

6. In traditional programs, a developer designs logic or algorithms to solve a problem. The program applies this logic to input and computes the output.But in Machine Learning, a model is built from the data, and that model is the logic.

1c.Perform the following operations on Carmanufacturing company dataset auto-mpg.csv given below using pandas. 10M

a) 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.



	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

a) Reading data from an existing file:

```
import pandas as pd
data = pd.read_csv("auto-mpg.csv")
```

b) Statistical details of dataset:

```
stats = data.describe()
print(stats)
```

c) Get all cars with 8 cylinders:

```
eight_cylinder_cars = data[data['cylinders'] == 8]
print(eight_cylinder_cars)
```

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

```
cars_by_year = data.groupby('model year')['model year'].count()
print(cars_by_year)
```

Any equivalent code can be considered

2a. Differentiate between supervised machine learning and Unsupervised machine learning 5M

Supervised Learning	**Unsupervised Learning**

Supervised learning algorithms are trained using labelled data.	Unsupervised learning algorithms are trained using unlabelled data.
Supervised learning model takes direct feedback to check if it is predicting correct output or not.	Unsupervised learning model does not take any feedback.
Supervised learning model predicts the output.	Unsupervised learning model finds the hidden patterns in data.
In supervised learning, input data is provided to the model along with the output.	In unsupervised learning, only input data is provided to the model.
The goal of supervised learning is to train the model so that it can predict the output when it is given new data.	The goal of unsupervised learning is to find the hidden patterns and useful insights from the unknown dataset.
Supervised learning needs supervision to train the model.	Unsupervised learning does not need any supervision to train the model.
Supervised learning can be categorized in Classification and Regression problems.	Unsupervised Learning can be classified in Clustering and Associations problems.
Supervised learning can be used for those cases where we know the input as well as corresponding outputs.	Unsupervised learning can be used for those cases where we have only input data and no corresponding output data.
It includes various algorithms such as Linear Regression, Logistic Regression, Support Vector Machine, Multi-class Classification, Decision tree, etc.	It includes various algorithms such as Clustering, KNN, and Apriori algorithm.
Supervised learning model produces an accurate result.	Unsupervised learning model may give less accurate result as compared to supervised learning.

(Any 5 can be considered)

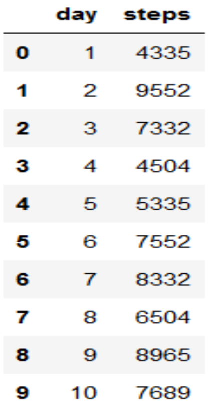
2b. Explore different sources of big data in machine learning 5M

Although Big data originates from multiple sources, below are the most common ones.

- Data collected from social Media sites like Facebook, WhatsApp, Twitter, YouTube, Instagram, etc
- Sensor placed in various places of the city that gathers data on temperature, humidity, etc. A camera placed beside the road gather information about traffic condition and creates data. Security cameras placed in sensitive areas like airports, railway stations, and shopping malls create a lot of data.
- IoT Appliance: Electronic devices that are connected to the internet create data for their smart functionality, examples are a smart TV, smart washing machine, smart coffee machine, smart AC, etc. It is machine-generated data that are created by sensors kept in various devices
- Customer feedback on the product or service of the various company on their website creates data. For Example, retail commercial sites like Amazon, Walmart, Flipkart, and Myntra gather customer feedback on the quality of their product and delivery time.
- E-commerce: In e-commerce transactions, business transactions, banking, and the stock market, lots of records stored are considered one of the sources of big data. Payments through credit cards, debit cards, or other electronic ways, all are kept recorded as data.
- Global Positioning System (GPS): GPS in the vehicle helps in monitoring the movement of the vehicle to shorten the path to a destination to cut fuel, and time consumption. This system creates huge data on vehicle position and movement.
- Transactional Data: Transactional data, as the name implies, is information obtained through online and offline transactions at various points of sale. The data contains important information about transactions, such as the date and time of the transaction, the location where it took place etc

2c. Ramesh 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 10 days

10M

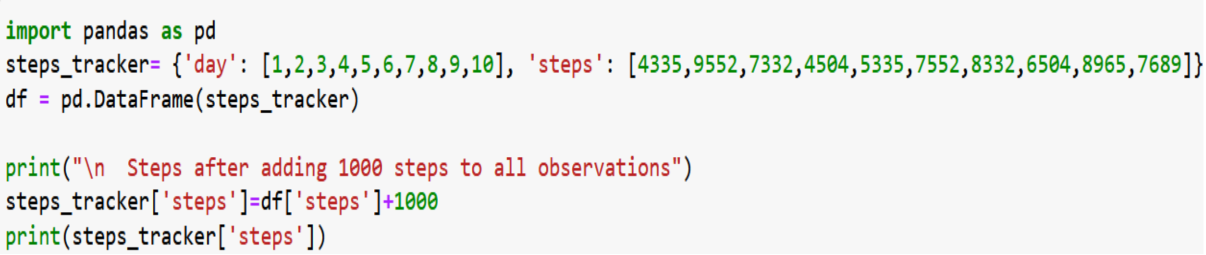


	day	steps
0	1	4335
1	2	9552
2	3	7332
3	4	4504
4	5	5335
5	6	7552
6	7	8332
7	8	6504
8	9	8965
9	10	7689

a) Perform an appropriate operation to add 1000 steps to all the observations

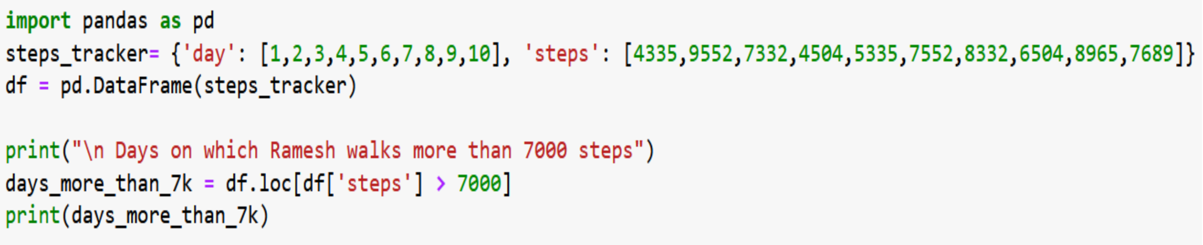
b) Find out the days on which he walked more than 7000 steps

1a)



```
import pandas as pd
steps_tracker= {'day': [1,2,3,4,5,6,7,8,9,10], 'steps': [4335,9552,7332,4504,5335,7552,8332,6504,8965,7689]}
df = pd.DataFrame(steps_tracker)

print("\n  Steps after adding 1000 steps to all observations")
steps_tracker['steps']=df['steps']+1000
print(steps_tracker['steps'])
```

1b)



```
import pandas as pd
steps_tracker= {'day': [1,2,3,4,5,6,7,8,9,10], 'steps': [4335,9552,7332,4504,5335,7552,8332,6504,8965,7689]}
df = pd.DataFrame(steps_tracker)

print("\n Days on which Ramesh walks more than 7000 steps")
days_more_than_7k = df.loc[df['steps'] > 7000]
print(days_more_than_7k)
```

Any other equivalent code can be considered

Section-II

3a.How to handle missing values in the dataset? Explain. **10M** Different approachesto handle the missing valuesare as follows

1. Keep the missing value as is
2. Remove data objects with missing values (Deleting the entire column)
3. Remove the attributes with missing values (Deleting the entire row)
4. Estimate and impute missing values

1. Keep the missing value as is
   Sometimes missing data is very less number of rows (say less than 3%) then we can simply ignore themissingdata.Thereisnohardruleto keepthemissingdata itdepends onus
2. Remove data objects with missing values (Deleting the entire column)

   If a certain column has many missing values, then you can choose to drop the entire column.Codetodroptheentirecolumnis asfollows:
   df=train_df.drop(['Dependents'],axis=1)
   df.isnull().sum()
3. Remove the attributes with missing values (Deleting the entire row)
   If a row has many missing values, then you can choose to drop the entire row.
   Codeto droptheentirerowisasfollows:

```
df =train_df.dropna(axis=0)
df.isnull().sum()
```

4. Estimateandimputemissingvalues

- Replacing with Arbitrary Value
  If you can make an educated guess about the missing value, then youcan replace it with some arbitrary value using the following code.
  Ex: In the following code,we are replacing the missing values of the ‘Dependents’column with‘0’
  train_df['Dependents']=train_df['Dependents'].fillna(0)

- Replacing with Mean
  This is the most common method of imputing missing values of numeric columns.One can use the ‘fillna’ method for imputing the columns ‘Loan Amount’ with theme an oft he respective column values as below
  train_df['LoanAmount'].fillna(train_df['LoanAmount'].mean())

- Replacing with Mode
  Mode is the most frequently occurring value. It is used in the case of categorical features.
  You can use the ‘fillna’ method for imputing the categorical columns ‘Gender’, ‘Married’, and‘Self_Employed’.
  train_df['Gender'].fillna(train_df['Gender'].mode()[0])

- Replacing with Median
  Median is the middlemost value. It’s better to use the median value for imputation in the case ofoutliers. You can use ‘fillna’ method for imputing the column ‘Loan_Amt’ with the median value.
  train_df['Loan_Amt']=train_df['Loan_Amt'].fillna(train_df['Loan_Amt'].median()[0])

Note: Code is not mandatory for any approach

3b. A dataset is given to you for creating machine learning model. What are the steps followed before using the data for training the model?Elaborate each step. 10M

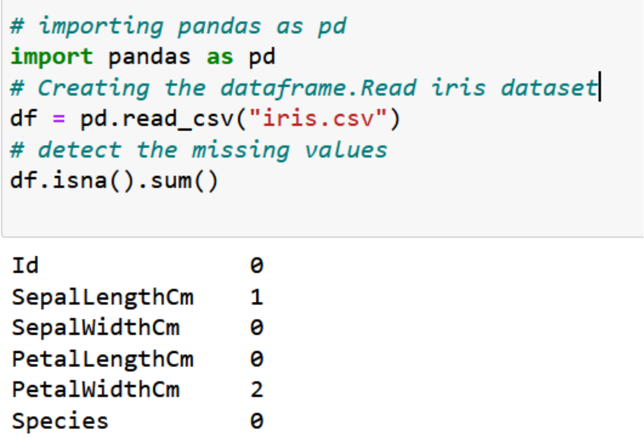
1. Data Exploration: The first step is to explore the data and understand the characteristics of the dataset. This includes understanding the number of observations and variables, the data types of each variable, and the distribution of the data. This can be done by using summary statistics and visualizations such as histograms, box plots, and scatter plots.
2. Data Cleaning: The next step is to clean the data. This includes handling missing or corrupted data, removing outliers, and addressing any other data quality issues. This step is important because dirty data can lead to inaccurate or unreliable models.
3. Data Transformation: After cleaning the data, it may be necessary to transform the data to make it suitable for the machine learning model. This can include normalizing the data, scaling the data, or creating new variables.
4. Feature Selection: Once the data is cleaned and transformed, it is important to select the relevant features that will be used to train the model. This step can be done by using techniques such as correlation analysis, principal component analysis, or mutual information.
5. Data Splitting: The next step is to split the data into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune the model's parameters, and the test set is used to evaluate the model's performance.
6. Feature Engineering: This step is to create new features that will be useful in the model. This can include creating interaction terms, polynomial terms, or binning variables.
7. Evaluation Metric: Selecting the right evaluation metric will help to evaluate the model's performance. Common evaluation metrics include accuracy, precision, recall, F1 score, and area under the ROC curve.
8. Model Selection: After the data is prepared, the next step is to select the appropriate machine learning model. This can be done by comparing the performance of different models using the evaluation metric.

.

4a.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)
10M

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

Using isna() function to detect the missing values in a dataframe

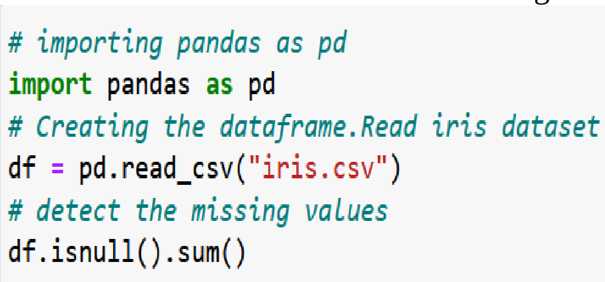


```
# importing pandas as pd
import pandas as pd
# Creating the dataframe.Read iris dataset
df = pd.read_csv("iris.csv")
# detect the missing values
df.isna().sum()
```

```
Id               0
SepalLengthCm    1
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     2
Species          0
```

Or

one can use isnull method to check missing values as below

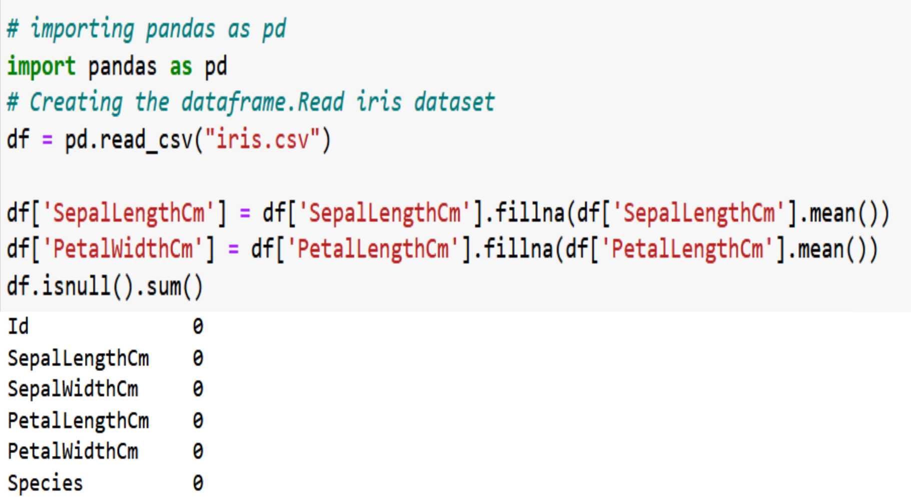


```
# importing pandas as pd
import pandas as pd
# Creating the dataframe.Read iris dataset
df = pd.read_csv("iris.csv")
# detect the missing values
df.isnull().sum()
```

Replacing missing values with mean value

This is the most common method of imputing missing values of numeric columns. You can use the ‘fillna’ method for imputing the numerical columns.

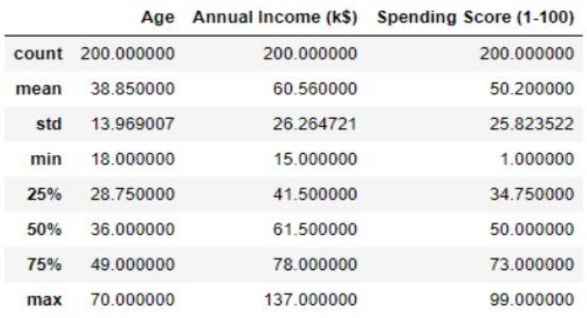
Suppose **SepalLengthCm** and **PetalWidthCm** columns are having missing values then below code can be written to replace null values with mean using fillna method



```
# importing pandas as pd
import pandas as pd
# Creating the dataframe.Read iris dataset
df = pd.read_csv("iris.csv")

df['SepalLengthCm'] = df['SepalLengthCm'].fillna(df['SepalLengthCm'].mean())
df['PetalWidthCm'] = df['PetalLengthCm'].fillna(df['PetalLengthCm'].mean())
df.isnull().sum()
```

```
Id               0
SepalLengthCm    0
SepalWidthCm     0
PetalLengthCm    0
PetalWidthCm     0
Species          0
```

4b. Statistical summary of Credit card dataset is as follows. Analyse and explain statistical metrics from above summary **10M**



	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000	137.000000	99.000000

Analysis

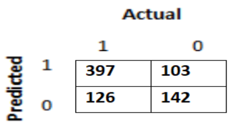
The statistical summary provided appears to be describing a dataset containing information about the credit card details of the customers, with columns “Age”, “Annual Income” and Spending Score”. The summary tells age of credit card holder,his annual income and spending score in the range 1-100

Explanation of statistical metrics

count - The number of not-empty valuesfor each variable
mean - The average (mean) value of each variable
std - The standard deviation column gives a measure of how much the values of each variable vary from the mean
min - the minimum value.
25% - The 25% percentile(is the same as the first quartile)
50% - The 50% percentile (is the same as the second quartile)
75% - The 75% percentile (is the same as the third quartile)
max - the maximum value of each variable

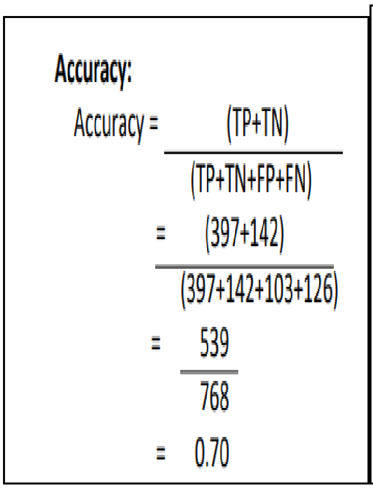
Section-III

5a. A Machine learning model was built to classify patient asCOVID +ve(1) or –ve(0). The confusionmatrix for the model is as shown below. Evaluate accuracy, precision, recall, specificity and F1-Score. 10M



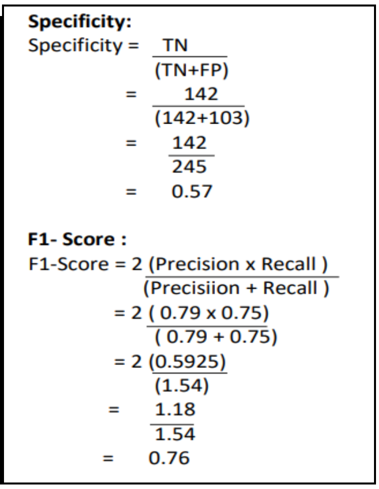
		Actual	
		1	0
Predicted	1	397	103
	0	126	142

TruePositive(TP):397
True Negative(TN) : 142
FalsePositive(FP):103
FalseNegative(FN): 126



Accuracy:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} = \frac{(397+142)}{(397+142+103+126)} = \frac{539}{768} = 0.70$$

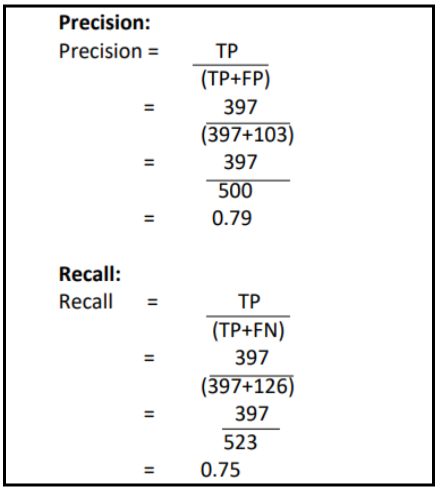


Specificity:

$$\text{Specificity} = \frac{TN}{(TN+FP)} = \frac{142}{(142+103)} = \frac{142}{245} = 0.57$$

F1- Score :

$$\text{F1-Score} = 2\,\frac{(\text{Precision x Recall})}{(\text{Precisiion + Recall})} = 2\,\frac{(0.79 \times 0.75)}{(0.79 + 0.75)} = 2\,\frac{(0.5925)}{(1.54)} = \frac{1.18}{1.54} = 0.76$$



Precision:

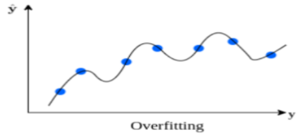
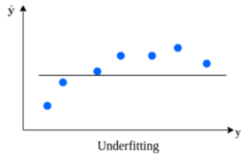
$$\text{Precision} = \frac{TP}{(TP+FP)} = \frac{397}{(397+103)} = \frac{397}{500} = 0.79$$

Recall:

$$\text{Recall} = \frac{TP}{(TP+FN)} = \frac{397}{(397+126)} = \frac{397}{523} = 0.75$$

5b. Compare overfitting with underfitting 5M

Overfitting	Underfitting
Overfitting happens when we train a	Underfitting occurs when the machine

machine learning model too much tuned to the training set. As a result, the model learns the training data too well, but it can't generate good predictions for unseen data. it produces low accuracy results for data points unseen in training, hence, leads to non-optimal decisions.Overfitting models produce good predictions for data points in the training set but perform poorly on new samples. It is accurate for training set It is not accurate for validation set  Simplest technique you can use to reduceOverfitting is Feature Selection	learning model is not well-tuned to the training set. The resulting model is not capturing the relationship between input and output well enough. Therefore, it doesn't produce accurate predictions, even for the training dataset. Resulting, an underfitting model generates poor results that lead to high-error decisions. It is not accurate for training set It is not accurate for validation set  Increase the number of features in the dataset to reduce underfitting

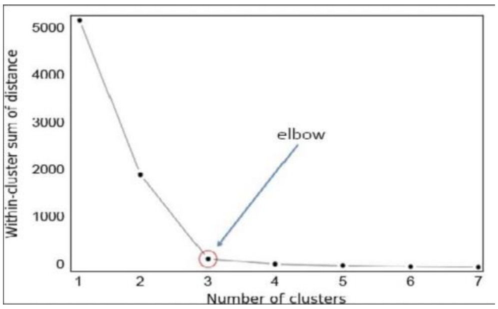
5c. How to Choose the Right Number of Clusters in k-means clustering? Explain any one method 5M

Three approaches to find the optimal number of clusters:

- The elbow method
- The optimization of the silhouette coefficient
- The gap statistic

Elbow method

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the total within-cluster sum of square (wss).
3. Plot the curve of wss according to the number of clusters k.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.



Average silhouette method

The algorithm is similar to the elbow method and can be computed as follow:

1. Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
2. For each k, calculate the average silhouette of observations (*avg.sil*).
3. Plot the curve of *avg.sil* according to the number of clusters k.
4. The location of the maximum is considered as the appropriate number of clusters.

Gap statistic method

The algorithm works as follow:

1. Cluster the observed data, varying the number of clusters from k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_k.
2. Generate B reference data sets with a random uniform distribution. Cluster each of these reference data sets with varying number of clusters k = 1, …, k_{max}, and compute the corresponding total within intra-cluster variation W_{kb}.
3. Compute the estimated gap statistic as the deviation of the observed W_k value from its expected value W_{kb} under the null hypothesis: $Gap(k)=1B\sum b=1Blog(W*kb)-log(Wk)$
4. Compute also the standard deviation of the statistics.
5. Choose the number of clusters as the smallest value of k such that the gap statistic is within one standard deviation of the gap at k+1: $Gap(k)\geq Gap(k + 1)-s_{k + 1}$.

6a. 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 iteration10M

Iteration-01:

We calculate the distance of each point from each of the centre of the three clusters.
The distance is calculated by using the given distance function.
C1 : (2,10) C2 : (5,8) C3 : (1,2)

Calculating distance between A1(2,10) and C1,C2,C3
P(A1,C1) = |x2 – x1| + |y2 – y1|
= |2 -2 | + |10 - 10|
= 0
P(A1,C2) = |x2 – x1| + |y2 – y1|
= | 5-2 | + | 8 - 10|
= 5
P(A1, C3) = |x2 – x1| + |y2 – y1|
= | 1-2 | + | 2 - 10|
= 9

Calculating distance between A2(2,5) and C1,C2,C3
P(A2,C1) = |x2 – x1| + |y2 – y1|
= |2 -2 | + |10 - 5|
= 5
P(A2,C2) = |x2 – x1| + |y2 – y1|
= | 5-2 | + | 8 - 5|
= 6
P(A2,C3) = |x2 – x1| + |y2 – y1|
= | 1-2 | + | 2 – 5 |
= 4

Calculating distance between A3(8,4) and C1,C2,C3
P(A3,C1) = |x2 – x1| + |y2 – y1|
= | 2 -8 | + |10 - 4|
= 12
P(A3,C2) = |x2 – x1| + |y2 – y1|
= | 5-8 | + | 8 – 4 |
= 7
P(A3,C3) = |x2 – x1| + |y2 – y1|
= | 1-8 | + | 2 – 4 |
= 9

Calculating distance between A4(5,8) and C1,C2,C3
P(A4, C1) = |x2 – x1| + |y2 – y1|
= | 2 -5 | + |10 -8 |
= 5

P(A4, C2) = |x2 – x1| + |y2 – y1|
= | 5-5 | + | 8 -8 |
= 0
P(A4, C3) = |x2 – x1| + |y2 – y1|
= | 1-5 | + | 2 –8 |
= 10

Calculating distance between A5(7,5) and C1,C2,C3
P(A5,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 7 | + |10 - 5 |
= 10
P(A5,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 7 | + | 8 - 5 |
= 5
P(A5,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 7 | + | 2 - 5 |
= 9
Calculating distance between A6(6,4) and C1,C2,C3
P(A6,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 6 | + |10 - 4 |
= 10
P(A6,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 6 | + | 8 - 4 |
= 5
P(A6,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 6 | + | 2 - 4 |
= 7
Calculating distance between A7(1,2) and C1,C2,C3
P(A7,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 1 | + |10 - 2 |
= 9
P(A7,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 1 | + | 8 - 2 |
= 10
P(A7,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 1 | + | 2 - 2 | =0
Calculating distance between A8(4,9) and C1,C2,C3
P(A8,C1) = |x2 – x1| + |y2 – y1|
= | 2 - 4 | + |10 - 9 |
= 3
P(A8,C2) = |x2 – x1| + |y2 – y1|
= | 5 - 4 | + | 8 - 9 |
= 2
P(A8,C3) = |x2 – x1| + |y2 – y1|
= | 1 - 4 | + | 2 - 9 |
= 10

We draw a table showing all the results.Using the table, we decide which point belongs to which cluster.The given point belongs to that cluster whose centre is nearest to it.

Given Points	Distance from centre (2, 10) of Cluster-01	Distance from centre (5, 8) of Cluster-02	Distance from centre (1, 2) of Cluster-03	Point belongs to Cluster
A1(2, 10)	0	5	9	C1

A2(2, 5)	5	6	4	C3
A3(8, 4)	12	7	9	C2
A4(5, 8)	5	0	10	C2
A5(7, 5)	10	5	9	C2
A6(6, 4)	10	5	7	C2
A7(1, 2)	9	10	0	C3
A8(4, 9)	3	2	10	C2

From here, New clusters are-

Cluster-01: First cluster contains points-A1(2, 10)
Cluster-02: Second cluster contains points-A3(8, 4),A4(5, 8),A5(7, 5),A6(6, 4),A8(4, 9)
Cluster-03: Third cluster contains points-A2(2, 5),A7(1, 2)

- We re-compute the new clusters.
- The new cluster centre is computed by taking mean of all the points contained in that cluster.

For Cluster-01:
We have only one point A1(2, 10) in Cluster-01, so, cluster centre remains the same.
For Cluster-02:
Canter of Cluster-02
= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5)
= (6, 6)
For Cluster-03:
Canter of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

This is completion of Iteration-01.

6b. Compare classification algorithms with clustering algorithm 5M

Classification	Clustering
Classification is a supervised learning technique	Clustering is an unsupervised learning technique
Classification is used to predict a class or label for a given data point	Clustering is used to group similar data points together
classification algorithms include Decision Trees, Random Forest, Naive Bayes	Clustering algorithms include K-means, Hierarchical, DBSCAN
classification requires labelled data for training	Clustering does not require labelled data
classification is mainly used for prediction	Clustering is used to find pattern in data
classification is used to make predictions based on known patterns	Clustering can be used to discover hidden patterns or relationships in the data

Any other relevant points can also be considered

6c. K-means clustering with Euclidean distance suffer from the curse of dimensionality. Is the statement true and why? 5M

Yes, the statement is true. The curse of dimensionality refers to the phenomenon that as the number of dimensions (features) in a dataset increases, the amount of data needed to accurately estimate the relationships between the points also increases exponentially. This can make it difficult to use traditional statistical techniques, such as k-means clustering, which rely on calculating distances between points.

In k-means clustering, the distance between points is typically measured using the Euclidean distance, which is the straight-line distance between two points. As the number of dimensions increases, the distance between points can become much larger, even if the points are relatively close to each other in lower-dimensional space. This can cause k-means to become less effective at accurately identifying clusters in the data.

One way to mitigate the curse of dimensionality in k-means clustering is to use dimensionality reduction techniques, such as principal component analysis (PCA), to reduce the number of dimensions in the dataset before applying k-means. Another option is to use a different clustering algorithm, such as DBSCAN, which does not rely on calculating distances between points and is therefore less affected by the curse of dimensionality.

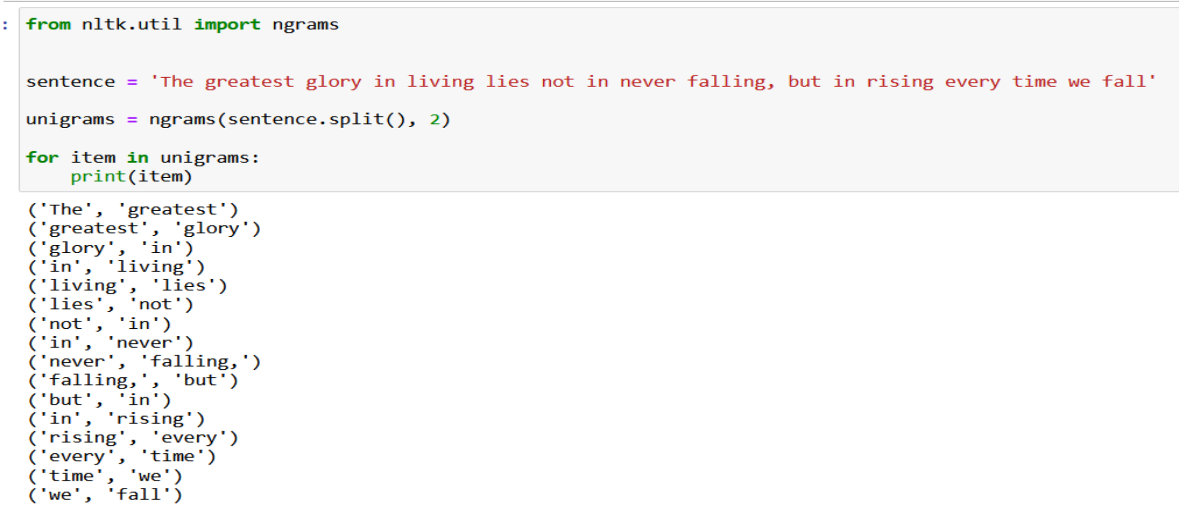
Section IV

7a. N-grams are defined as the combination of N keywords together. Consider the given sentence: "The greatest glory in living lies not in never falling, but in rising every time we fall"

a. Generate bi grams for the above sentence **10M**

b. Generate tri-grams for the above sentence

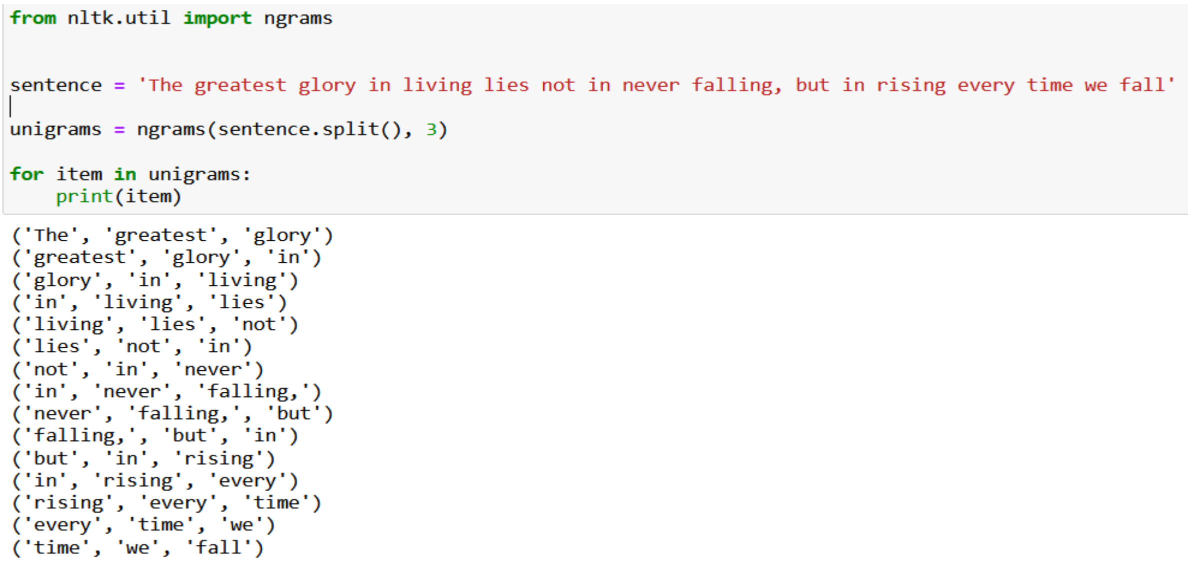
a. bigram



```
from nltk.util import ngrams

sentence = 'The greatest glory in living lies not in never falling, but in rising every time we fall'

unigrams = ngrams(sentence.split(), 2)

for item in unigrams:
    print(item)
```

```
('The', 'greatest')
('greatest', 'glory')
('glory', 'in')
('in', 'living')
('living', 'lies')
('lies', 'not')
('not', 'in')
('in', 'never')
('never', 'falling,')
('falling,', 'but')
('but', 'in')
('in', 'rising')
('rising', 'every')
('every', 'time')
('time', 'we')
('we', 'fall')
```

b.trigram



```
from nltk.util import ngrams

sentence = 'The greatest glory in living lies not in never falling, but in rising every time we fall'

unigrams = ngrams(sentence.split(), 3)

for item in unigrams:
    print(item)
```

```
('The', 'greatest', 'glory')
('greatest', 'glory', 'in')
('glory', 'in', 'living')
('in', 'living', 'lies')
('living', 'lies', 'not')
('lies', 'not', 'in')
('not', 'in', 'never')
('in', 'never', 'falling,')
('never', 'falling,', 'but')
('falling,', 'but', 'in')
('but', 'in', 'rising')
('in', 'rising', 'every')
('rising', 'every', 'time')
('every', 'time', 'we')
('time', 'we', 'fall')
```

7b.Discuss importance of dimensionality reduction in machine learning **5M**

The number of features or input variables that contribute to the prediction process to predict target or output in machine learning model is called dimensionality. However, it is observed that all feature variables do not have the same significance / contribution in the output.Therefore, before training a machine learning model, it is necessary to identify relevant features that contribute to output / target and this process is called dimensionality reduction

Dimensionality reduction is essential because of the following reasons:

1. To reduce complexity of model:More number of features / high dimensionality will lead to build a complex model specially when there is a high correlation exits in the feature variables. Thus, it is useful to select the right set of features to overcome this challenge.
2. To prevent overfitting:A dataset with high dimensionality sometimes may lead to overfitting of a machine learning model because the model captures key features and noise as well. Thus, the model performs well during training of the model, but performance degrades on testing on unknown data.
3. To achieve computational efficiency:A machine learning model with low dimensionality takes less time in training of the model because it takes less time in computations.
4. It is useful to lower model training time and lower the data storage requirement.
5. It prevents from curse of dimensionality.

7c. Summarise different strategies of production deployment 5M

Canary deployment releases the services incrementally in small phases, enabling the organizations to test their applications against real users and simultaneously analyze their different versions for improvement.

Shadow deployment contains a new version and its old version alongside to test the predictions. A more recent version is released after successful testing. A copy of the production environment traffic is sent to the older version to ensure successful working.

Blue/green deployment maintains two stages in parallel: Blue for staging and green for production. Quality assurance testing of new versions is carried out in the blue environment, and later, real traffic is sent from the green environment to the blue environment. Once the testing goes successful, the product is shifted to the green environment.

Rolling deployment: This involves releasing a new version of an application to a small subset of servers before releasing it to all servers. This allows for testing and monitoring before rolling out to the entire server fleet

A/B testing: This involves releasing different versions of an application to different subsets of users in order to test and compare the performance of each version.

8a. Demonstrate Stemming and Lemmatization concepts with suitable examples 10M

Lemmatization and Stemming are Text Normalization techniques. These techniques are used to prepare words, text, and documents for further processing.

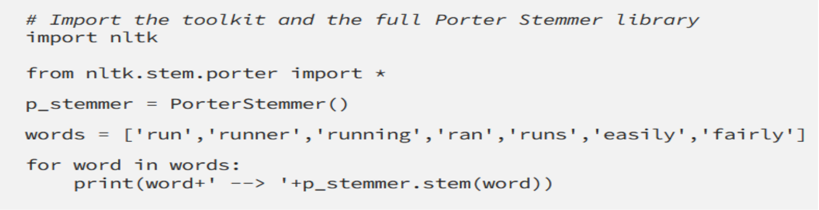
Stemming

It is the process of producing morphological variants of a root/base word. “boat” would be the stem for [boat, boater, boating, boats].

Lemmatization is a method responsible for grouping different inflected forms of words into the root form, having the same meaning. It is similar to stemming, in turn, it gives the stripped word that has some dictionary meaning.

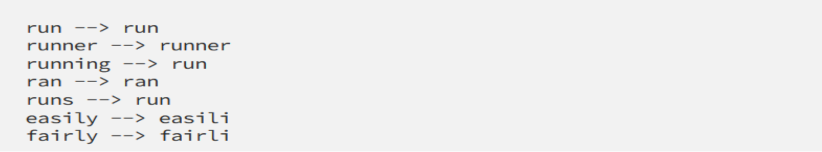
Lemmatization clearly identifies the base form of ‘troubled’ to ‘trouble’’ denoting some meaning whereas, Stemming will cut out ‘ed’ part and convert it into ‘troubl’ which has the wrong meaning and spelling errors.

‘troubled’ -> Lemmatization -> ‘troubled’, and error
‘troubled’ -> Stemming -> ‘troubl’
Ex for stemming



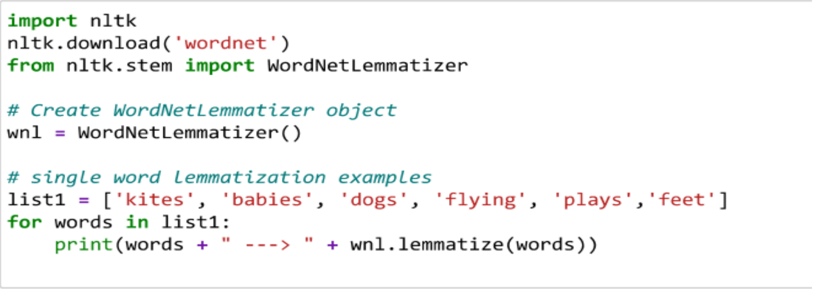
```
# Import the toolkit and the full Porter Stemmer library
import nltk

from nltk.stem.porter import *

p_stemmer = PorterStemmer()

words = ['run','runner','running','ran','runs','easily','fairly']

for word in words:
    print(word+' --> '+p_stemmer.stem(word))
```

output

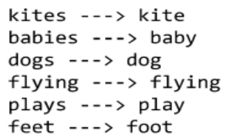


```
run --> run
runner --> runner
running --> run
ran --> ran
runs --> run
easily --> easili
fairly --> fairli
```

Ex for lemmatization



```
import nltk
nltk.download('wordnet')
from nltk.stem import WordNetLemmatizer

# Create WordNetLemmatizer object
wnl = WordNetLemmatizer()

# single word lemmatization examples
list1 = ['kites', 'babies', 'dogs', 'flying', 'plays','feet']
for words in list1:
    print(words + " ---> " + wnl.lemmatize(words))
```

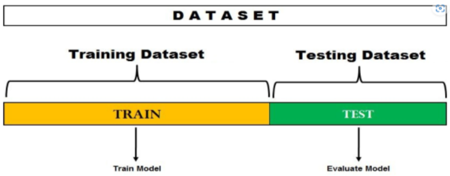


```
kites ---> kite
babies ---> baby
dogs ---> dog
flying ---> flying
plays ---> play
feet ---> foot
```

8b. Discuss different techniques of cross validation 5M

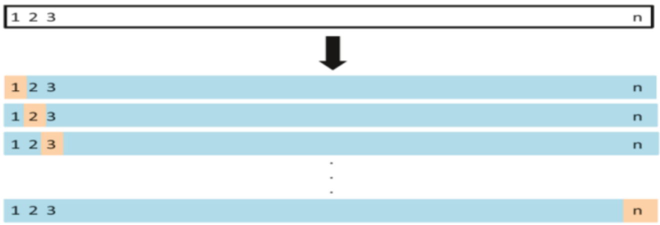
1. Hold Out method

This is the simplest evaluation method and is widely used in Machine Learning projects. Here the entire dataset(population) is divided into 2 sets – train set and test set. The data can be divided into 70-30 or 60-40, 75-25 or 80-20, or even 50-50 depending on the use case. As a rule, the proportion of training data has to be larger than the test data.



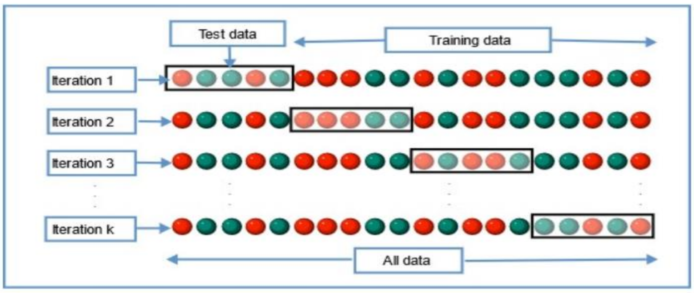
2. Leave One Out Cross-Validation

In this method, we divide the data into train and test sets – but with a twist. Instead of dividing the data into 2 subsets, we select a single observation as test data, and everything else is labelled as training data and the model is trained. Now the 2nd observation is selected as test data and the model is trained on the remaining data.



3. K-Fold Cross-Validation

In this resampling technique, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining k-1 sets. The test error rate is then calculated after fitting the model to the test data.In the second iteration, the 2nd set is selected as a test set and the remaining k-1 sets are used to train the data and the error is calculated. This process continues for all the k sets.



8c. What are MLOps? brief different stages that are involved in the MLOps lifecycles 5M

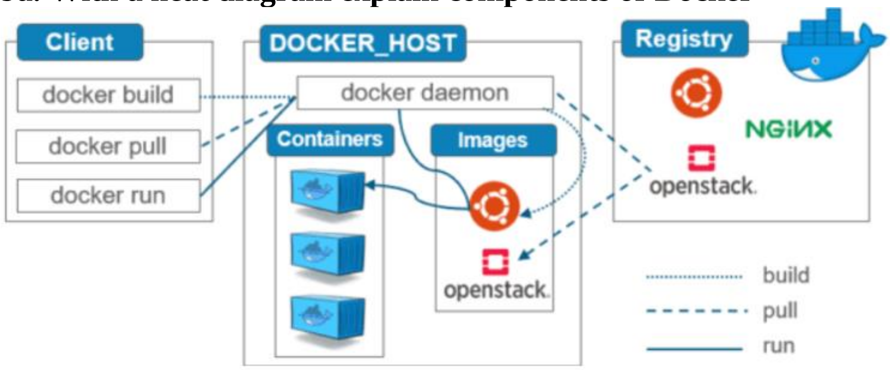
MLOps stands for Machine Learning Operations. MLOps is focused on streamlining the process of deploying machine learning models to production, and then maintaining and monitoring them. MLOps is a collaborative function, often consisting of data scientists, ML engineers, and DevOps engineers. The word MLOps is a compound of two different fields i.e. machine learning and DevOps from software engineering.

1. ML Development: This is the basic step that involves creating a complete pipeline beginning from data processing to model training and evaluation codes.
2. Model Training: Once the setup is ready, the next logical step is to train the model. Here, continuous training functionality is also needed to adapt to new data or address specific changes.
3. Model Evaluation: Performing inference over the trained model and checking the accuracy/correctness of the output results.
4. Model Deployment: When the proof of concept stage is accomplished, the other part is to deploy the model according to the industry requirements to face the real-life data.
5. Prediction Serving: After deployment, the model is now ready to serve predictions over the incoming data.
6. Model Monitoring: Over time, problems such as concept drift can make the results inaccurate hence continuous monitoring of the model is essential to ensure proper functioning.
7. Data and Model Management: It is a part of the central system that manages the data and models. It includes maintaining storage, keeping track of different versions, ease of accessibility, security, and configuration across various cross-functional teams.

   Any four stages explanation can be awarded marks

Section V

9a. With a neat diagram explain components of Docker 10M



Components of Docker

1. Docker Client

Docker client uses **commands** and **REST APIs** to communicate with the Docker Daemon (Server). When a client runs any Docker command on the Docker client terminal, the client terminal sends these Docker commands to the Docker daemon. Docker daemon receives these commands from the Docker client in the form of command and REST API's request.

Docker Client uses Command Line Interface (CLI) to run the following commands -

- Docker build
- Docker pull
- Docker run

2. Docker Registry

Docker Registry manages and stores the Docker images.There are two types of registries in the Docker Pubic Registry **-** Public Registry is also called as Docker hub**.**

Private Registry **-** It is used to share images within the enterprise.

3.Docker Daemon: This is the background process that runs on the host machine and manages the containers. It is responsible for creating, starting, stopping, and removing containers, as well as managing their network and storage resources.

4.Docker Images

Docker images are the read-only binary templates used to create Docker Containers. It uses a private container registry to share container images within the enterprise and also uses public container registry to share container images within the whole world. Metadata is also used by docket images to describe the container's abilities.

4.Docker Containers

Containers are the structural units of Docker, which is used to hold the entire package that is needed to run the application. The advantage of containers is that it requires very less resources.In other words, we can say that the image is a template, and the container is a copy of that template.

9b. Demonstrate Simple Linear Regression consideringa dataset that has two variables: salary (dependent variable) and experience (Independent variable) 10M

To implement the Simple Linear regression model in machine learning using Python, we need to follow the below steps:

Step 1: import libraries

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
```

step 2 :Load the dataset

```
df = pd.read_csv("salary_data.csv")
```

Step 3: pre-processing.Check for any missing values and handle it by any suitable method

```
df.isnull().sum()
mean_A = df['salary'].mean()
df['salary'] = df['salary'].fillna(mean_A)
df.isnull().sum()
```

Step 4: Split the data set. Extract the dependent and independent variables from the given dataset.

```
x = df['Experience']
y = df['salary']
x_train, x_test, y_train, y_test = train_test_split(x,y, train_size = 0.8, test_size=0.2, random_state = 21)
x_train=x_train.values.reshape(-1,1)
x_test =x_test.values.reshape(-1,1)
```

Step 5: Build model

```
model = LinearRegression()
model.fit(x_train, y_train)
```

Step 6: Find the accuracy

```
model.score(x_test,y_test)
```

10a.Summerize any two cloud deployment models 10M

1.Public Cloud

The public cloud makes it possible for anybody to access systems and services. The public cloud may be less secure as it is open to everyone. The public cloud is one in which cloud infrastructure services are provided over the internet to the general people or major industry groups. The infrastructure in this cloud model is owned by the entity that delivers the cloud services, not by the consumer. It is a type of cloud hosting that allows customers and users to easily access systems and services. This form of cloud computing is an excellent example of cloud hosting, in which service providers supply services to a variety of customers. In this arrangement, storage backup and retrieval services are given for free, as a subscription, or on a per-user basis. Example: Google App Engine etc.

Advantages of Public Cloud Model:

- Minimal Investment**:** Because it is a pay-per-use service, there is no substantial upfront fee, making it excellent for enterprises that require immediate access to resources.
- No setup cost: The entire infrastructure is fully subsidized by the cloud service providers, thus there is no need to set up any hardware.
- Infrastructure Management is not required: Using the public cloud does not necessitate infrastructure management.
- No maintenance: The maintenance work is done by the service provider (Not users).
- Dynamic Scalability: To fulfil company's needs, on-demand resources are accessible.

Disadvantages of Public Cloud Model:

- Less secure: Public cloud is less secure as resources are public so there is no guarantee of high-level security.
- Low customization: It is accessed by many public so it can't be customized according to personal requirements.

2.Private Cloud

The private cloud deployment model is the exact opposite of the public cloud deployment model. It's a one-on-one environment for a single user (customer). There is no need to share your hardware with anyone else. The distinction between private and public clouds is in how you handle all of the hardware. It is also called the "internal cloud" & it refers to the ability to access systems and services within a given border or organization. The cloud platform is implemented in a cloud-based secure environment that is protected by powerful firewalls and under the supervision of an organization's IT department. The private cloud gives greater flexibility of control over cloud resources.

Advantages of Private Cloud Model:

- Better Control: You are the sole owner of the property. You gain complete command over service integration, IT operations, policies, and user behaviour.
- Data Security and Privacy: It's suitable for storing corporate information to which only authorized staff have access. By segmenting resources within the same infrastructure, improved access and security can be achieved.
- Supports Legacy Systems: This approach is designed to work with legacy systems that are unable to access the public cloud.
- Customization: Unlike a public cloud deployment, a private cloud allows a company to tailor its solution to meet its specific needs.

Disadvantages of Private Cloud Model:

- Less scalable: Private clouds are scaled within a certain range as there is less number of clients.
- Costly: Private clouds are costlier as they provide personalized facilities.

Hybrid Cloud

By bridging the public and private worlds with a layer of proprietary software, hybrid cloud computing gives the best of both worlds. With a hybrid solution, you may host the app in a safe environment while taking advantage of the public cloud's cost savings. Organizations can move data and applications between different clouds using a combination of two or more cloud deployment methods, depending on their needs.

Advantages of Hybrid Cloud Model:

- Flexibility and control: Businesses with more flexibility can design personalized solutions that meet their particular needs.
- Cost: Because public clouds provide scalability, you’ll only be responsible for paying for the extra capacity if you require it.
- Security: Because data is properly separated, the chances of data theft by attackers are considerably reduced.

Disadvantages of Hybrid Cloud Model:

- Difficult to manage: Hybrid clouds are difficult to manage as it is a combination of both public and private cloud. So, it is complex.
- Slow data transmission: Data transmission in the hybrid cloud takes place through the public cloud so latency occurs.

10b.Discuss any five ethical challenges in AI **10M**

1. Bias and Discrimination: AI systems can perpetuate and even amplify biases and discrimination if they are not properly designed and tested. For example, an AI system that is used to make decisions about hiring or lending may discriminate against certain groups of people if it is trained on data that contains such biases.
2. Privacy and Security: AI systems can collect and process large amounts of personal data, which can raise concerns about privacy and security. For example, an AI system that is used to monitor people's behaviour or predict their behaviour may collect sensitive information about them, which could be used to discriminate against them or to cause harm.
3. Lack of Explain ability and Transparency: Many AI systems are based on complex algorithms that are difficult to understand or explain. This can make it difficult for people to understand how the AI system is making its decisions and to hold the system accountable for its actions.
4. Job Loss: AI systems can automate many tasks that are currently performed by humans, which can lead to job loss and other economic dislocation. It's important to consider the social and economic impacts of AI and to ensure that the benefits of AI are shared fairly.
5. Autonomy and Control: AI systems can be programmed to make decisions and take actions autonomously, which can raise concerns about who is in control of the system and who is responsible for its actions.
6. Ethical dilemmas: AI systems may be faced with ethical dilemmas, such as the trade-off between human lives and property damage in self-driving cars, and it may be difficult for the system to make the right decision.
7. Societal impact: The development and deployment of AI can have a significant impact on society, and it's important to consider the broader ethical, social, and political implications of AI and to ensure that the technology.

Certified that model answers prepared by me for code 20CS51I are from syllabus and scheme of valuation prepared by me is correct

K Savitha

SAVITHA KAMALAPURKAR
Lecturer Dept of Comp Science &Engg
S.J(Govt.) Polytechnic Bangalore

				Sujatha H.K.