Question 1

A retail company wants to analyze customer purchasing patterns but does not have labeled data. They aim to group customers based on their shopping behavior. Which type of machine learning approach should they use?

Options:

A) Predict outcomes for labeled data.

B) Train models with predefined answers.

C) Discover patterns and structures in unlabeled data.

D) Generate random outputs from datasets.

Answer: C

Explanation: The company should use unsupervised learning, which focuses on finding patterns and relationships within unlabeled datasets.

Question 2

A developer writes the following Python code to manage file operations, using google colab.
What will be the output?

```
f1 = open("file.txt", "w")
f1.write("ab\ncdefg")
f1.close()

f1 = open("file.txt", "r")
print(f1.readlines())
f1.close()
```

Options:

A) `ab\ncdefg`

B)

```
ab
cde
```

C) `('ab\n', 'cdefg')`

D) `['ab\n', 'cdefg']`

Answer: D

Explanation:

- The file is written with two lines: `"ab"` , `"cdefg"` , separated by a newline character (`\n`).
- `f1.readlines()` reads all lines from the file and returns them as a list.

Question 3

A student is working with text files and uses the following code to manipulate file pointers. He is using google colab.

```
with  open("kitab.txt", "w") as f1:
    f1.write("ab\ncd\nef\ngh")

with  open("kitab.txt", "r") as f2:
    f2.seek(3)
    print(f2.read())
```

What will be the output of the code?

Options:

A)

```
cd
ef
gh
```

B)

```
ncd\nef\ngh
```

C)

```
\ncd\nef\ngh
```

Answer: A

Explanation:

Here, `\n` represents a newline character.

2. The file is then opened in read mode (`"r"`) and `seek(3)` is used. The **seek()** function moves the file pointer to the specified position (character index `3`).
3. Let's break down the file content with indexes:

```
Index: 0 1 2 3 4 5 6 7 8 9 10

Content: a b \n c d \n e f \n g h
```

- `a` is at index `0`
- `b` is at index `1`
- `\n` (newline) is at index `2`
- `c` is at index `3`

  and so on

4. Since `seek(3)` moves the pointer to **index** `3` , reading from this position will output everything starting from `"c"` , including the subsequent newline and characters.

Question 4 (MCQ)

A supply chain analyst has two DataFrames:

python

```
df1 = pd.DataFrame({'item': ['X', 'Y', 'Z', 'W'], 'stock': [10, 20, 30, 40]})
df2 = pd.DataFrame({'item': ['Y', 'W', 'P', 'Q'], 'cost': [15, 25, 35, 45]})
```

The analyst performs an inner join on the column `'item'` . What will be the resulting DataFrame?

Options:

A)

```
   item  stock  cost
0   Y     20    15
1   W     40    25
```

B)

```
   item  stock  cost
0    X     10   NaN
1    Y     20   15.0
```

C)

```
   item  stock  cost
0   Y     20     15
1   W     40     25
2   P     NaN    35
```

D) Empty DataFrame

Answer: A

Explanation:

Inner join retains only matching 'item' values (Y/W). All other rows/columns are excluded.

Question 5 (MCQ)

A sales manager has two DataFrames containing sales and profit data:

```
df1 = pd.DataFrame({'ID': [101, 102, 103], 'sales': [200, 300, 400]})
df2 = pd.DataFrame({'ID': [102, 103, 104], 'profit': [50, 60, 70]})
```

The manager performs a right join on `'ID'`. What will be the resulting DataFrame?

Options:

A)

```
     ID   sales   profit
0   102  300.0     50
1   103  400.0     60
2   104    NaN     70
```

B)

```
     ID  sales   profit
0  101  200.0    NaN
1  102  300.0     50
```

C)

```
     ID   profit  sales
0   101    NaN    200.0
1   102    50     300.0
2   103    60     400.0
```

D) `KeyError`

Answer: A
Explanation:
Right join prioritizes df2's IDs (102/103/104) or, common values and missing values in df2's ID.

Question 6 (MCQ)

A weather analyst is working with two datasets: one containing temperature data for cities and another containing humidity data. The analyst performs a full outer join on the `'city'` column to combine the datasets.

```
df1 = pd.DataFrame({'city': ['Paris', 'London'], 'temp': [22, 18]})
df2 = pd.DataFrame({'city': ['London', 'Berlin'], 'humidity': [65, 70]})
```

How many missing values (NaN) will exist in the resulting DataFrame after this operation?

Options:
A) 0
B) 1

C) 2
D) 3

Answer: C
Explanation:

- Resultant DataFrame:

```
   city     temp    humidity
0  Berlin   NaN      70.0
1  London   18.0     65.0
2  Paris    22.0      NaN
```

NaN in `temp` (Berlin) and `humidity` (Paris) → 2 nulls.

Question 7 (MCQ)

A teacher is analyzing test scores of students stored in a DataFrame, where some scores are missing. The teacher wants to fill these missing values using the next valid score from below in the same column.

```
import pandas as pd
import numpy as np


data = {

'Student': ['A', 'B', 'C', 'D'],

'Test1': [90, np.nan, 85, 88],

'Test2': [np.nan, 78, 82, 80],

'Test3': [88, np.nan, 89, np.nan]

}

df = pd.DataFrame(data)
```

You apply the following operation:

```
df = df.bfill(axis=0)
```

What will be the effect of this operation?

- **A)** Missing values will be filled using the mean of the respective columns.
- **B)** Missing values will be filled using the next valid value(non-null value just above it) in the same column.
- **C)** Missing values will be replaced with the most recent non-null value(non-null value just below it) in the same column.
- **D)** Missing values will be dropped from the DataFrame.

Correct Answer: C

Explanation:

Axis=0 ensures that missing values are filled downwards within each column (column-wise backward fill)

- **Option C** is correct because `bfill` : missing values will be replaced with the most recent non-null value(non-null value just below it) in the same column.

Question 8

A data scientist is combining two DataFrames using following code :

```
import pandas as pd
df1 = pd.DataFrame({'A':  [1,2],  'B':  [3,4]}, index=[10,  20])
df2 = pd.DataFrame({'A':  [5,6],  'B':  [7,8]}, index=[30,  40])
result = pd.concat([df1, df2],ignore_index=True)
print(result)
```

What will be the output of this operation?

Options:

A)

```
    A  B
0   1  3
1  20  2
2   4  30
3   5   7
4  40  6
```

B)

```
    A  B
0   1  3
1   2  4
2   5  7
3   6  8
```

C)

```
    A  B
0   5  7
1   6  8
2   1  3
3   2  4
```

D) `ValueError`

Answer: B
Explanation:

- `ignore_index=True` creates a new integer index (0-3) instead of preserving original indexes
- Data is concatenated vertically while resetting index positions

Question 9

A data analyst is working with two DataFrames: one containing sales data and the other containing profit data, using pandas. The analyst wants to concatenate these DataFrames horizontally and executes following code.

```
df1 = pd.DataFrame({'A':  [1,2],  'B':  [3,4]})
df2 = pd.DataFrame({'C':  [5,6],  'D':  [7,8]})
result = pd.concat([df1, df2], axis=1, join='inner')
```

What will be the result variable store?

Options:

A) Merges columns with NaN values

B) Returns empty DataFrame

C) Combines only matching column labels

D) Results in a dataframe with all columns of df1 and df2

Answer: D

Explanation:

The `join='inner'` operation ensures that all rows are included in the concatenated DataFrame since there are no missing row indices. The resulting DataFrame will contain all columns from both `df1` and `df2` .

Question 10

A machine learning engineer is tasked with building a model to predict house prices based on features like square footage and number of bedrooms. The dataset includes labeled examples where the target variable is known.

Which of the following is correct statement for determining which type of learning ML Engineer should use ?

Options:

A) Supervised learning uses unlabeled data, while unsupervised learning uses labeled data.

B) Supervised learning uses labeled data, while unsupervised learning uses unlabeled data.

C) Both use labeled data.

D) Both use unlabeled data.

Answer: B

Explanation:

Supervised learning is appropriate because the dataset contains labeled examples (house prices), allowing the model to learn relationships between input features and output labels.

Question 11

A company wants to classify emails into "spam" or "not spam" based on their content. They have a dataset where each email is labeled as either spam or not spam. A company wants to classify emails into "spam" or "not spam" based on their content. They have a dataset where each email is labeled as either spam or not spam.

What type of data is used in this classification task?

Options:

A) Continuous Data
B) Categorical Data
C) Both Continuous and Categorical Data
D) None of the above

Answer: B
Explanation: Classification tasks involve categorical data, where the goal is to predict discrete categories such as "spam" or "not spam".

Question 12

An environmental scientist wants to predict daily rainfall amounts based on historical weather patterns like temperature and humidity levels.

What is the primary goal of this prediction task?

Options:

A) Predicting discrete categories.
B) Predicting continuous values.
C) Sorting unlabeled data into clusters.
D) Identifying anomalies in datasets.

Answer: B
Explanation: The scientist's goal is to predict continuous values (rainfall amounts), making this a regression task.

Question 13

An economist is modeling the relationship between advertising spend and revenue generation for a company using a linear regression approach.

Which equation represents this model?

Options:

A) y=mx+b

B) y=ax^2+bx+cy

C) y=1/x

D) y=e^x

Answer: A

Explanation: Linear regression models are represented by y=mx+b, where m is the slope and b is the y-intercept.

Question 14

A machine learning researcher wants to train a linear regression model. The goal is to find the best fit line that models the relationship between the input features (`X`) and target values (`Y`).

What is the primary purpose of training a linear regression model?

Options:

A) To plot a graph of the input vs output.

B) To establish a mathematical relationship between input and output.

C) To calculate statistical metrics for evaluation.

D) To generate random outputs.

Answer: B

Explanation: The primary goal of training a linear regression model is to establish a mathematical relationship (usually in the form `y = mx + b`) between input features and the target variable.

Question 15(NAT)

An engineer wants to build a model to predict the output values based on a simple linear relationship between input (`X`) and output (`Y`). The engineer finds that the relationship between `X` and `Y` is given by the equation:

$$y = 2x + 0$$

What is the slope in this linear equation?

Answer: 2

Explanation: In the linear equation `y = mx + b`, `m` represents the slope and `b` represents the intercept. Here, the slope is 2, and the intercept is 0.