CFGDEGREE

DATA ASSESSMENT MATERIAL RELEASE

THEORY QUESTIONS

SECTION	MARK
1. Theory Questions	25
2. Pandas Questions	25
3. Matplotlib Challenge	25
4. Numpy Questions	25
TOTAL	100

Important notes:

- This document shares the first section of the Data Assessment which is composed of 5 Data Theory Questions
- The answers do not have to be long, but they have to answer each of the mention points for each question
- It is worth a quarter of your assessment mark
- You have 24 hours before the assessment to prepare.
- If any plagiarism is found in how you choose to answer a question you will receivea 0 and the instance will be recorded.
- Consequences will occur if this is a repeated offence. You can remind yourself of the plagiarism policy here.
- You are allowed to use any online images to support your answers.

Section 1: Theory Questions [25 points]

1.1 In your own words, what does the role of a data scientist involve?

2 points

A data scientist uses their statistical methods, programming, and machine learning expertise to analyze and interpret data sets, uncovering valuable insights and patterns that assist organizations in making informed decisions and solving complex problems. This role is instrumental in developing and implementing algorithms for predictive modeling and data-driven solutions.

They clean, process, analyse and visualize data, assisting businesses in understanding trends and making data-informed decisions. While data scientists often contribute to strategic decision-making and predictive modeling, data analysts play a key role in day-to-day operations by generating reports, identifying trends, and ensuring data quality.

1.2 What is an outlier? Here we expect to see the following:

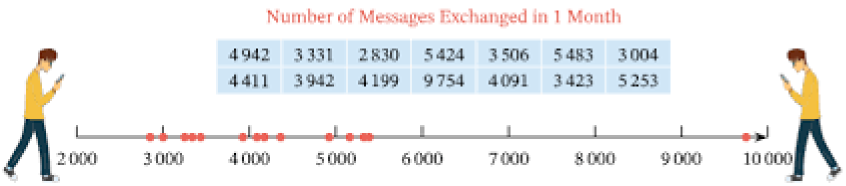
4 points

a. Definition

Outliers are data points that deviate significantly from the rest of the dataset.

b. Examples:

In this image I found online, the monthly volume of messages typically falls within the range of 2000 to 5500. However, there is a noticeable outlier where the number of messages exchanged in one month significantly exceeds this range, reaching 9754 messages. This value stands out as it is nearly double the upper limit of the range, so it's clearly an outlier.



Number of Messages Exchanged in 1 Month

4 942	3 331	2 830	5 424	3 506	5 483	3 004
4 411	3 942	4 199	9 754	4 091	3 423	5 253

(The image comes from this website https://www.nagwa.com/en/explainers/845148137695/)

c. Should outliers always be removed? Why?

Outliers don't always necessarily need to be removed, as they can offer crucial insights into the data distribution, discover interesting patterns, or spotlight significant anomalies within the dataset.

Before deciding to eliminate outliers, we need to evaluate the context and reasons behind the presence of outliers before deciding whether to exclude them.

Removing outliers without a careful prior assessment could result in information loss, biased outcomes, or an oversimplified portrayal of the true scenario.

d. What are other possible issues that you can find in a dataset?

Missing data: In some datasets, there could be missing values for some observations or variables.

Skewed distribution: When the data is not evenly distributed and shows a bias towards one end.

Data entry errors: Inaccuracies due to typing mistakes, incorrect measurements, or other human errors during data entry.

Inconsistent formatting: Differences in how data is formatted or recorded across different sources or entries.

Sampling bias: Occurs when the sample used in the dataset is not representative of the population it's supposed to represent.

1.3 Describe the concepts of data cleaning and data quality. Here we expect to see the following:

4 points

a. What is data cleaning? Data cleaning involves identifying, correcting, and eliminating inconsistencies or errors within the dataset. The purpose of cleaning data is to make sure its accurate, reliable, and suited for analysis.

b. **Why is data cleaning important?** The process of data cleaning holds immense significance for various reasons.
Firstly, it ensures that the analysis and decision-making based on the data are accurate, free from errors and inconsistencies that could otherwise lead to erroneous conclusions.
Moreover, it elevates the dependability of the results, fostering confidence in the insights drawn from the dataset.
Additionally, by upholding uniformity throughout the dataset, data cleaning facilitates meaningful comparisons and trend analysis.
This process also optimizes efficiency, streamlining the analysis and saving valuable time and resources by preemptively addressing errors and inconsistencies.
Ultimately, by displaying a dedicated commitment to quality and precision in data management, clean data cultivates trust among stakeholders and users.

c. What type of mistakes do we expect to commonly see in datasets?

Some of the most common mistakes are:

- **Missing data:** In some datasets, there could be missing values for some observations or variables.
- **Skewed distribution:** When the data is not evenly distributed and shows a bias towards one end.
- **Outliers:** Outliers are data points that deviate significantly from the rest of the dataset.
- **Inaccurate labeling:** Improper or inconsistent labels assigned to data points can mislead the model, leading to erroneous predictions.
- **Redundancy:** Excessively similar or overlapping features can cause overfitting, preventing the model from generalizing to new data.
- **Insufficient data size:** Small datasets can lead to overfitting, where the model memorizes specific training examples rather than learning generalizable patterns. Poor performance on unseen data results.
- **Data Corruption:** Errors or inconsistencies introduced during data collection, storage, or processing can severely impact the model's performance and lead to inaccurate predictions.
- **Lack of diversity:** Datasets lacking diversity in features, categories, or samples limit the model's ability to capture the complexities of real-world problems. Oversimplification impedes generalization and performance.
- **Dynamic data:** Constantly evolving datasets due to changing trends or new data points challenge static models that require continuous retraining and adaptation to maintain accuracy.

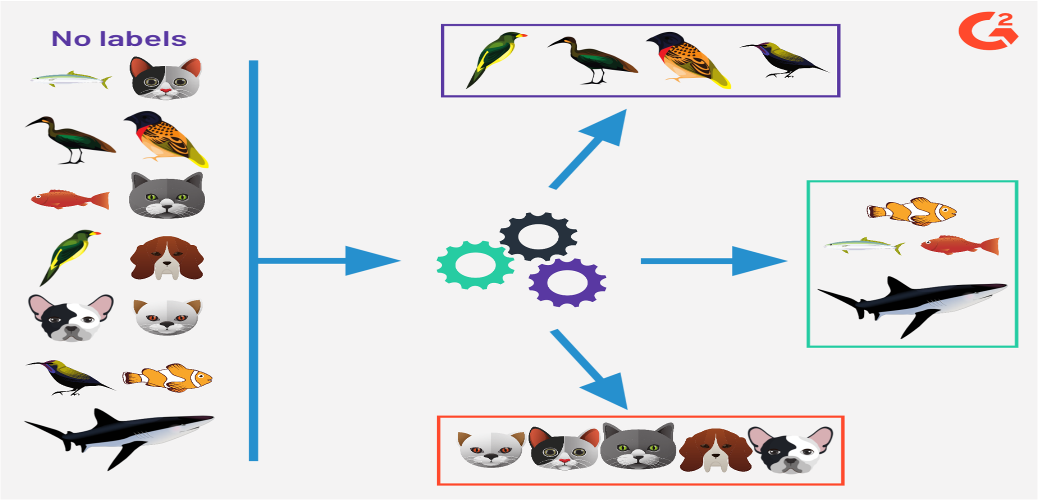
1.4 Discuss what is Unsupervised Learning - Clustering in Machine Learning using an example. Here we expect to see the following:

7.5 points

a. Definition.

It is a type of machine learning where there's no labeled data and it must learn to find patterns or structure within the data.

Clustering is a type of unsupervised learning that groups similar data in groups.



(The photo comes from this website https://www.g2.com/articles/supervised-vs-unsupervised-learning)
In the image we can see how there is a group of animals that are not labeled, but then groups animals based on similarity.

b. When is it used?

Clustering is used when there is no labeled output available for training, and the goal is to explore the inherent structure or relationships within the data. It is particularly useful when dealing with large datasets where manually labeling data points is impractical or costly.

c. What is a possible real-world application of unsupervised learning?

Music genre classification:

Music streaming platforms such as Spotify collect a vast amount of data on users' listening habits. Unsupervised learning can be applied to analyze this data for insights into music genre preferences and to discover new or emerging genres.

Clustering algorithms can group music tracks based on musical characteristics like tempo, melody, rhythm, and harmony, enabling the platforms to offer more tailored music recommendations to users.

d. What are its main limitations?

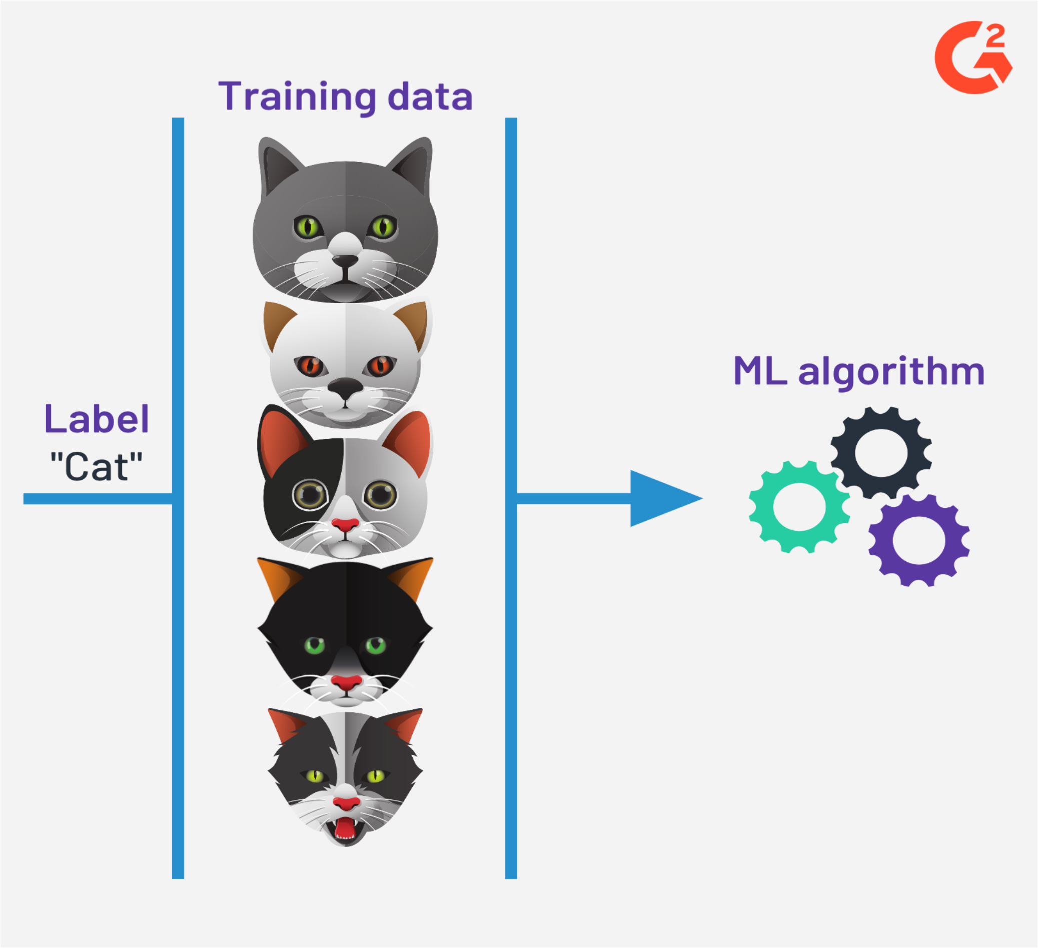
- Subjectivity in cluster interpretation: Interpreting the results and determining the best number of clusters can be unclear, resulting in subjective decision-making.
- Sensitivity to outliers: Clustering algorithms may be influenced by outliers, which can impact cluster formation and result in less accurate outcomes.
- Lack of ground truth: Because unsupervised learning does not have clearly labeled data, assessing the quality of clustering results can be difficult, as there is no definitive "correct" answer for the clusters generated.
- Dimensionality: Clustering becomes more challenging as the dimensionality of the data increases, making it difficult to visualize and interpret the clusters in high-dimensional spaces.

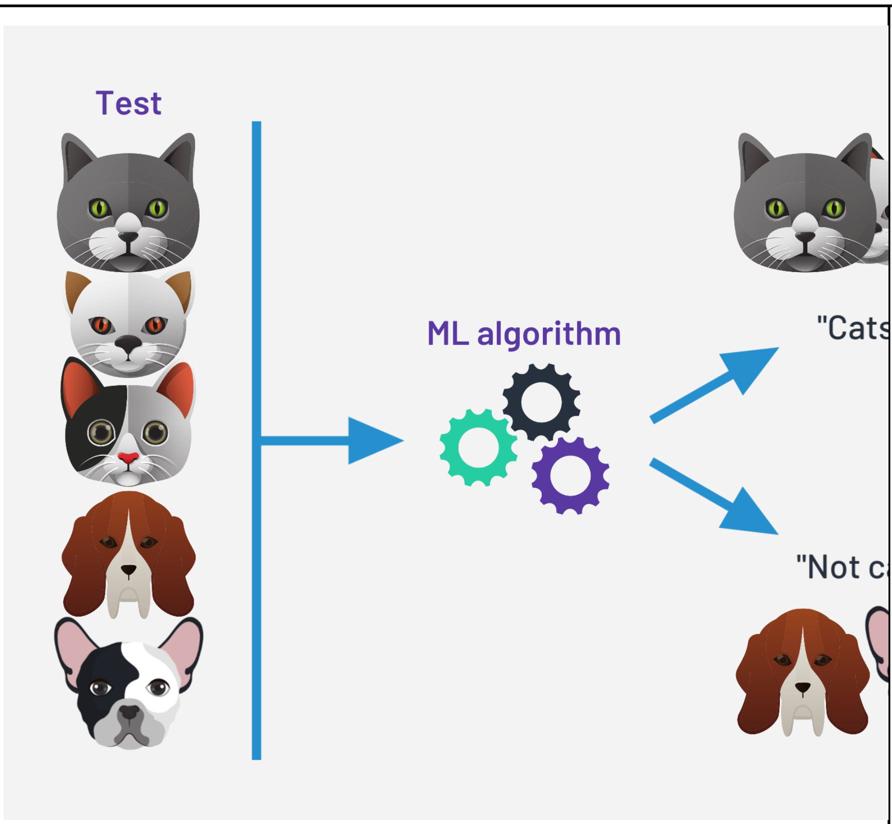
1.5 Discuss what is Supervised Learning - Classification in Machine Learning using an example. Here we expect to see the following:

7.5 points

a. Definition.

Supervised learning - classification is a type of machine learning where the algorithm is trained on a labeled dataset, meaning that the input data is paired with corresponding output labels. The goal of classification is to learn a mapping or relationship between the input features and predefined output categories.





b. When is it used?

Supervised learning, specifically classification, is utilized to predict or categorize new, unseen data into predefined classes or categories. This method is employed in diverse situations where labeled training data is accessible, aiming to generalize patterns and relationships between input features and output labels.

c. What is a possible real-world application of supervised learning?

The spam email filter algorithm is trained on a dataset of emails categorized as "spam" or "not spam." It identifies patterns in the features of these emails, such as words and sender information. When making predictions, it classifies new, unseen emails as either spam or not spam based on these learned patterns.

d. What data do we need for it? Is there any processing that needs to be done?

Supervised learning algorithms require labeled data, which means that the data has been tagged with the desired output or outcome. This labeled data is used to train the algorithm to learn the relationship between the input features and the output. The data can be numerical, categorical, or a combination of both.

This is the processing that needs to be done:

Missing value imputation: Filling in missing values in the data.

Data normalization: Scaling the data to a common range.

Feature engineering: Creating new features from the existing data. **Data cleaning**: Identifying and removing outliers and errors in the data. **Data transformation**: Transforming the data into a format that is suitable for the algorithm.	