CODE FIRST GIRLS

CFGDEGREE

DATA EXAM MATERIAL RELEASE

THEORY QUESTIONS ANSWERS

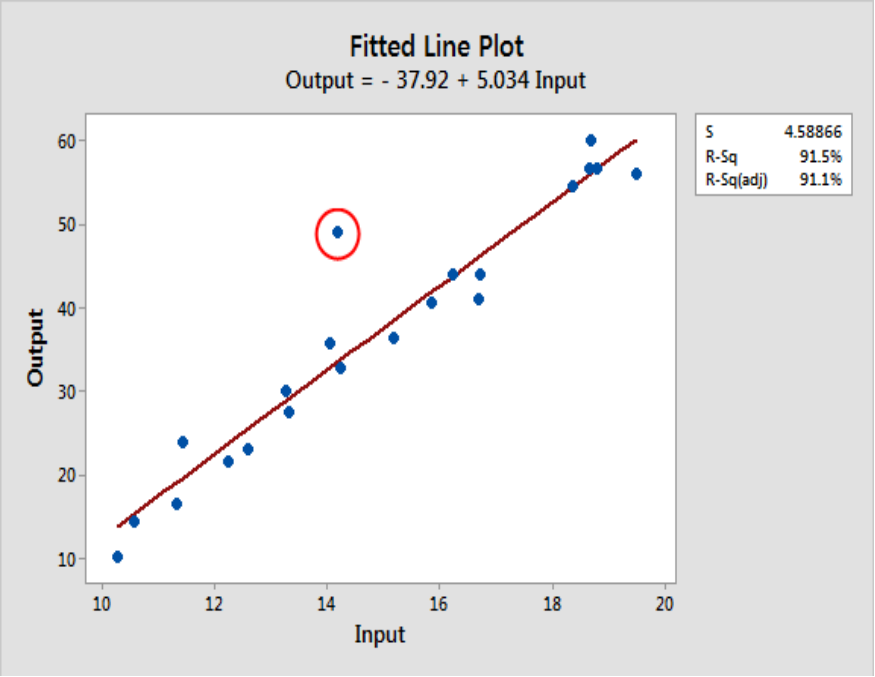
Section 1: Theory Questions [25 points] and Answers

1.1 In your own words, what does the role of a data scientist involve? A data scientist's role involves analyzing data to uncover insights and help make informed decisions. They gather and clean the data, create models to spot patterns or predict future trends, and explain their findings clearly. Their goal is to use data to solve problems and guide decisions.	**2 points**

1.2 What is an outlier? Here we expect to see the following: a. Definition Outliers are data points that are very different from most other points. They are values that don't match the general trend of the data. b. Examples Here are some examples of the outliers: 1. Medical Data: Most people's blood pressure readings are around 120/80, but one	**4 points**

reading is 300/200.
2. Sales Data: Most days, a store's sales are between £80 and £400, but one day it's £4,000 because of a big sale.
3. Temperature Data: Most days, the temperature is between 10°C and 32°C, but one day it's -10°C.

In the graph below there is (the source: https://statisticsbyjim.com/basics/outliers/) the example of outlier, the circled point that doesn't fit the line for the model.



Both values, the input and output are not unusual for this dataset, but don't fit the model that could cause problems in regression analysis.

c. Should outliers always be removed? Why?

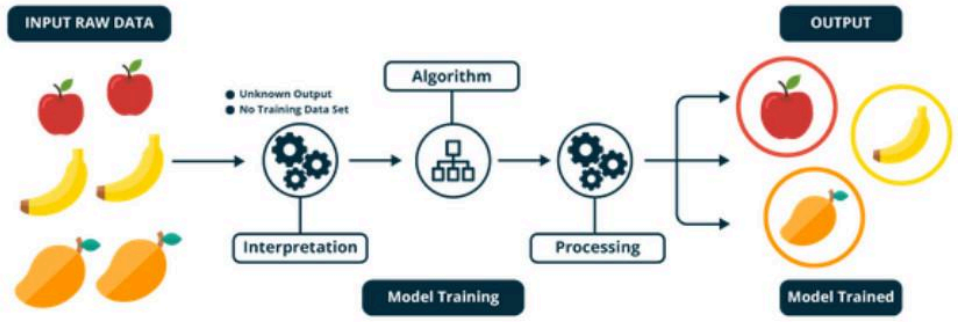
Outliers should not always be removed because they might provide important information, such as uncovering new insights or pointing out unusual behavior. It is better to remove them if they are due to errors (such as a data entry mistake) and you can't fix them, or if they are not relevant to your study. However, outliers should be kept if they are part of the natural variation, as they can reveal significant trends or phenomena that are crucial for understanding the full scope of the data.

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

You can find in a dataset other possible issues such as missing or incomplete data, duplicates, bias, inconsistent data (e.g., dates in different styles), errors (incorrect data due to mistakes), outdated information (data that may no longer be relevant or accurate.

1.3 Describe the concepts of data cleaning and data quality. Here we expect to see the following: a. What is data cleaning? Data cleaning or data cleansing, scrubbing is the process of fixing mistakes and removing errors from data. It helps correct things like typos, missing information, and duplicate entries. b. Why is data cleaning important? Data cleaning is important because it ensures the data is accurate, consistent, and ready for use. This process makes the information reliable for analysis, helping to avoid misleading conclusions and ensuring better decision-making. c. What type of mistakes do we expect to commonly see in datasets? Common mistakes in datasets include missing values, typos, duplicate records, inconsistent formatting, and outdated information. These issues can affect the quality of the data and its reliability for analysis.	**4 points**

1.4 Discuss what is Unsupervised Learning - Clustering in Machine Learning using an example. Here we expect to see the following: a. Definition. Unsupervised learning is a type of machine learning where the algorithm groups similar data points together without needing labeled examples or predefined categories. The goal is to find hidden patterns or structures within the data. Below the scheme that shows how Unsupervised Learning works (source of the image is here)	**7.5 points**



b. When is it used?

Clustering in Machine learning is used when you have data without specific labels and you want to discover natural groupings or patterns within it. It's useful for exploring data, identifying patterns, or summarizing large datasets.

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

A real-world application of unsupervised learning is customer segmentation in marketing. For example, a company can use clustering to group customers based on their purchasing behavior. This helps in tailoring marketing strategies and personalized offers for different customer groups.

d. What are its main limitations?

Its main limitations are that results can be hard to predict or understand. It's tough to measure how accurate or effective the results are because there are no predefined answers to compare against. It can also be challenging to determine the right number of groups. Additionally, the results can be affected by irrelevant or noisy data.

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

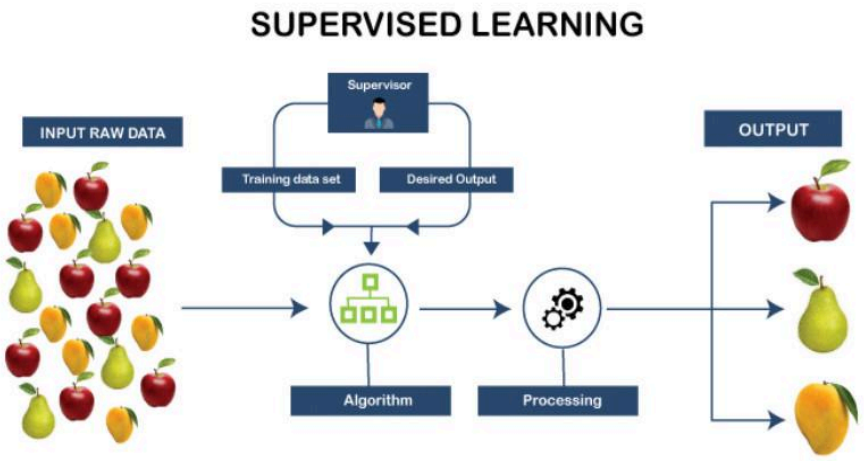
7.5 points

a. Definition.

Supervised Learning is a type of machine learning where the system is trained with examples that include both the input and the correct answer, so it means at least one part of the process requires human supervision. There are several types of Supervised Learning depending on the goal of the tasks, one of them is

Classification. Classification in Machine Learning is a type of Supervised Learning where the goal is to predict the category or class of an input based on its features, for example to classify emails as either "spam" or "not spam.

Below the scheme that illustrates Supervised Learning (the image source is here)



b. When is it used?

Supervised machine learning is used to sort new data into known categories and predict future trends based on past examples. It helps models learn patterns from labeled data so they can make accurate predictions, like forecasting house prices or customer buying habits. This method is useful for classifying files such as images or documents and for predicting future changes by understanding patterns in previous data.

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

A possible real-world application of supervised learning is email spam detection. The system is trained with examples of both spam and non-spam emails to learn the characteristics of each type. Once trained, the model can then automatically classify new emails as either spam or not spam based on what it has learned.

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

For supervised learning, we need a dataset with both input features and corresponding correct outputs (labels). The input features are the data points that the model uses to make predictions, while the labels are the known results that the model is trained to predict. For example, in spam detection, we need emails (inputs) labeled as "spam" or "not spam" (outputs).

Processing is necessary when building a supervised learning model. First, we need to gather and clean the data, transforming and simplifying it to make it suitable for the model. Next, we select a model based on the data type and

problem. Finally, we divide the data into training, validation, and testing sets to train, fine-tune, and evaluate the model's performance effectively.

Below the flowchart that illustrates Machine Learning workflow for Supervised Learning (the image source is here)