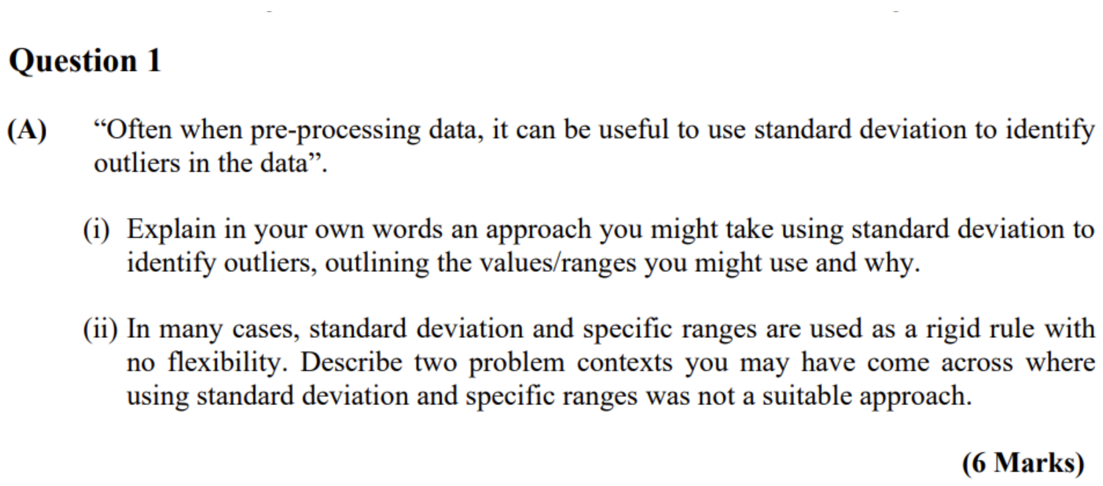


Question 1

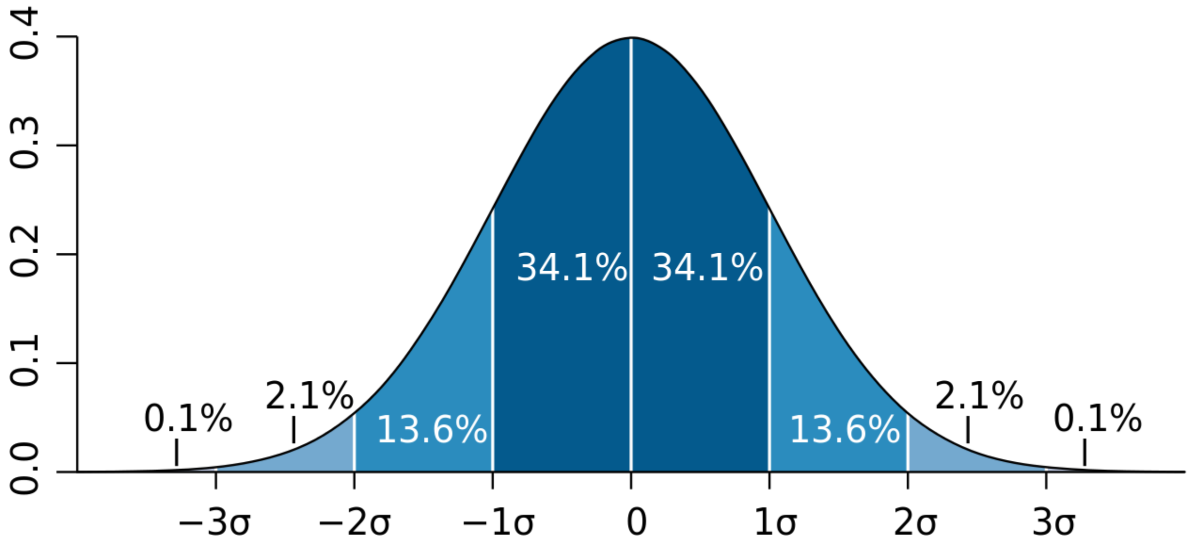
(A) "Often when pre-processing data, it can be useful to use standard deviation to identify outliers in the data".

(i) Explain in your own words an approach you might take using standard deviation to identify outliers, outlining the values/ranges you might use and why.

(ii) In many cases, standard deviation and specific ranges are used as a rigid rule with no flexibility. Describe two problem contexts you may have come across where using standard deviation and specific ranges was not a suitable approach.

(6 Marks)

a.i.



Mean – Expected average value of the distribution.

Variance – Spread of observation from the mean

Standard Deviation – Normalized spread of observations

We can the standard deviation of a numerical column of a dataset to find potential outliers in the data. Assuming the data follows a **Gaussian** distribution, like the way we assumed our datasets in the labs followed a **Gaussian** distribution, we can use the Standard Deviation to remove any potential outliers. But how is this done?

Well, we can first calculate the Standard Deviation from the data, using the Mean and Variance of the column. Then, we must multiply the Standard Deviation to get three times to get Three Standard Deviations to observe 99.7% of the data (empirical rule). The Standard Deviation is very useful in this context, and why we use it is because it is excellent at showing us the range of values data can be spread out in.

We can then use the Mean and the Standard Deviation to observe potential outliers within the data.

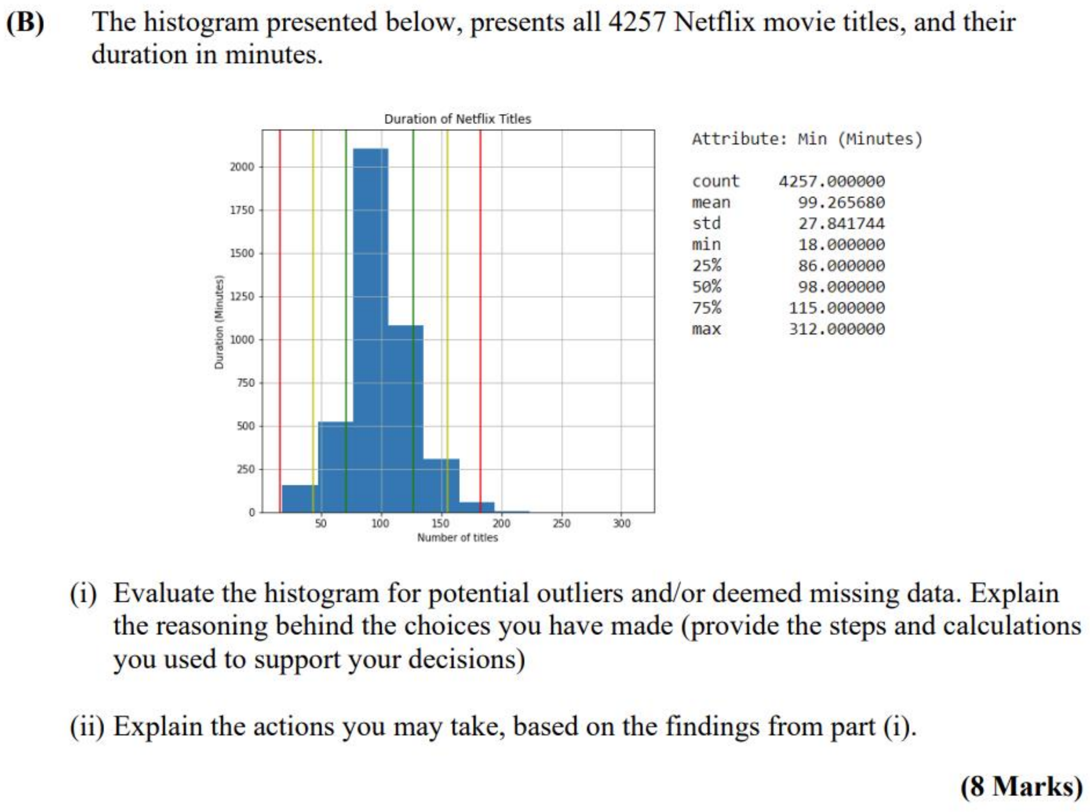
Mean – Three Standard Deviations = Anything below this value is **LIKELY** an outlier.

Mean + Three Standard Deviations = Anything above this value is **LIKELY** an outlier.

ii.

If our dataset column is not Standardized, then the Standard Deviation approach to removing outliers is not a valid one. We do not have the bell curve and it can't follow the Empirical Rule, so there is no use in using the Standard Deviation to remove any potential outliers. However, we COULD Normalize the data in order to fix this scenario. '

It may be the case that the data has a skewed distribution also, so it may be harder to find potential outliers.



(B) The histogram presented below, presents all 4257 Netflix movie titles, and their duration in minutes.

```
Attribute: Min (Minutes)

count    4257.000000
mean       99.265680
std        27.841744
min        18.000000
25%        86.000000
50%        98.000000
75%       115.000000
max       312.000000
```

(i) Evaluate the histogram for potential outliers and/or deemed missing data. Explain the reasoning behind the choices you have made (provide the steps and calculations you used to support your decisions)

(ii) Explain the actions you may take, based on the findings from part (i).

(8 Marks)

B.

i. Context is important, so let's take into consideration that this is a Netflix dataset, and it is a histogram of **movies.**

The average movie judging from the histography seems to range from 80 minutes to 110 minutes. Initial observations show there are several movies outside of this range. Movies having abnormal length is nothing unusual. Lord of the Rings The Two Towers' theatrical release was 223 minutes. Seriously.

However, I am concerned about movies below the two standard deviations. A movie being under 50 minutes is very unusual, and it is worth considering that maybe some TV episodes have gotten into this dataset by mistake.

There doesn't seem to be any movies below 0 minutes, so there are no outliers there.

If we're looking for **theatrical releases**, it is probably best to remove or impute the values outside 2 standard deviations. If we want to be safe with our outliers, then it is probably best to remove or impute values outside 3 standard deviations.

ii. After marking the data we can do two of the following:

- Remove missing instances
- Impute mean values for the missing values

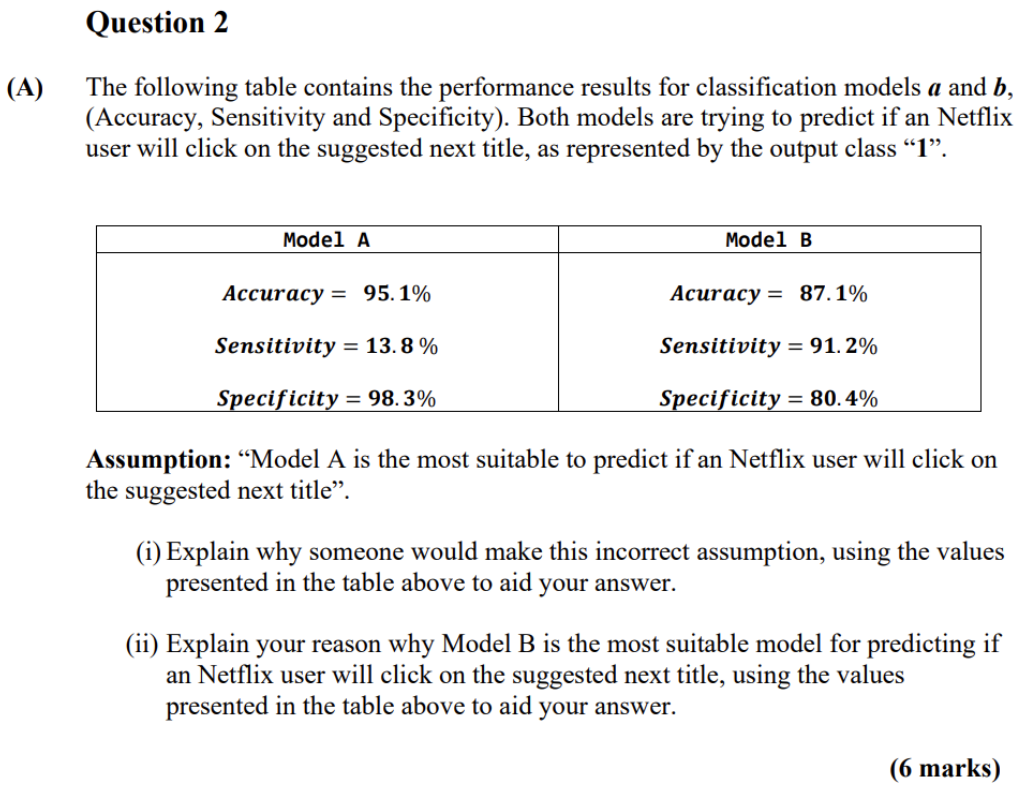
Removing the data might influence the dataset and model outcomes. If we were to remove data outside the two standard deviations like suggested previously, it might introduce a serious effect on how the model plays out by giving results more tailored towards theatrical releases. If we were to remove data outside the three standard deviations, there will be less of an effect due to the less data involved outside three standard deviations.

We could impute the marked outliers, with the average length of a movie. This would be using the mean to calculate the average length and would replace any outlier durations with that. This could also seriously affect the model and outcome.

Generally, if the missing data is below 5% of the total data, we impute the data with the average value. But if it's above 5%, we remove it.

C.

i. I would disagree. Pearson's correlation coefficient is great for feature selection when there are linear correlations. Spearmans might be better if there are little linear correlations.



Question 2

(A) The following table contains the performance results for classification models ***a*** and ***b***, (Accuracy, Sensitivity and Specificity). Both models are trying to predict if an Netflix user will click on the suggested next title, as represented by the output class “**1**”.

Model A	Model B
$\textit{Accuracy} = \ 95.1\%$ $\textit{Sensitivity} = 13.8\ \%$ $\textit{Specificity} = 98.3\%$	$\textit{Acuracy} = \ 87.1\%$ $\textit{Sensitivity} = 91.2\%$ $\textit{Specificity} = 80.4\%$

Assumption: “Model A is the most suitable to predict if an Netflix user will click on the suggested next title”.

(i) Explain why someone would make this incorrect assumption, using the values presented in the table above to aid your answer.

(ii) Explain your reason why Model B is the most suitable model for predicting if an Netflix user will click on the suggested next title, using the values presented in the table above to aid your answer.

(6 marks)

Question 2:

Sensitivity = True Positive Rate

Specificity = True Negative Rate

(A).

i.

One may look at Accuracy, and get confused about the naming conventions of Accuracy, Sensitivity and Specificity. **While Accuracy is an important indicator of the success of a model, and Model A does indeed have a higher amount of Accuracy (95.1%) compared to Model B (87.1%), this does not mean it is a better model for predicting if someone will watch something on Netflix due to the low amount of Sensitivity**. If Model A had a higher amount, or similar amount of Sensitivity, we could pick Model A because it has a higher a decently significant more amount of Accuracy.

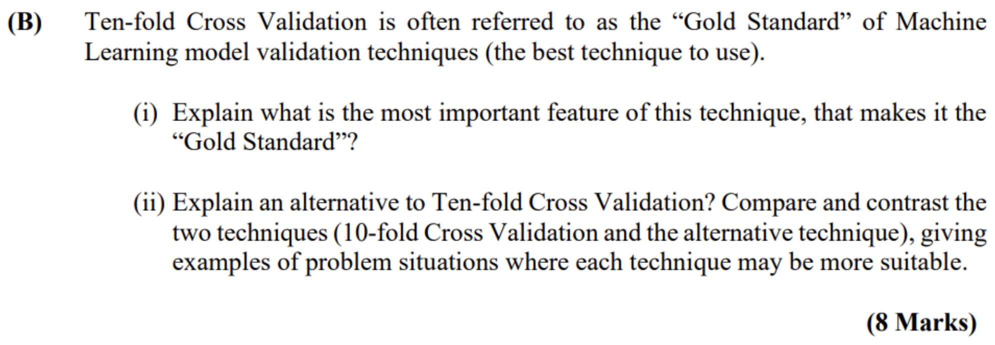
It is worth noting if the values of both Models were similar, picking one wouldn’t matter too much as they’d both produce similar results.

ii.

Output class 1 is positive, so we are looking for a positive outcome.

We are going to look at the True Positive Rate, or Sensitivity, which measures the proportion of positives correctly identified. In the context of the Netflix dataset, we are looking for how likely it is going to be that someone is going to watch something on Netflix. Model A has 13.8% Sensitivity. This is very poor, in terms of predicting a true positive rate. This means what we try to predict with a positive outcome is going to be terrible overall. Model B has 91.2% Sensitivity, so it is much better at predicting a positive outcome.

This means that it is a more suitable model overall, as well better at predicting whether someone is going to watch Netflix or not.



(B) Ten-fold Cross Validation is often referred to as the "Gold Standard" of Machine Learning model validation techniques (the best technique to use).

(i) Explain what is the most important feature of this technique, that makes it the "Gold Standard"?

(ii) Explain an alternative to Ten-fold Cross Validation? Compare and contrast the two techniques (10-fold Cross Validation and the alternative technique), giving examples of problem situations where each technique may be more suitable.

(8 Marks)

(B)

i. Ten-Fold Cross Validation is referred to as the Gold Standard of model validation techniques due to:

Being able to use all our data, as well as making sure test data is never present in training data at the same time. This provides that the model has been trained and tested using every possible instance. The dataset is partitioned into different subparts or 'k-folds', for each iteration selecting training data AND test data. This allows for the entire dataset to be utilized for both training and validation, while also allowing very possibility to be tested, repeating 10 times, or the number determined prior.

It's excellent if we have limited data to test our model.

https://towardsdatascience.com/understanding-8-types-of-cross-validation-80c935a4976d

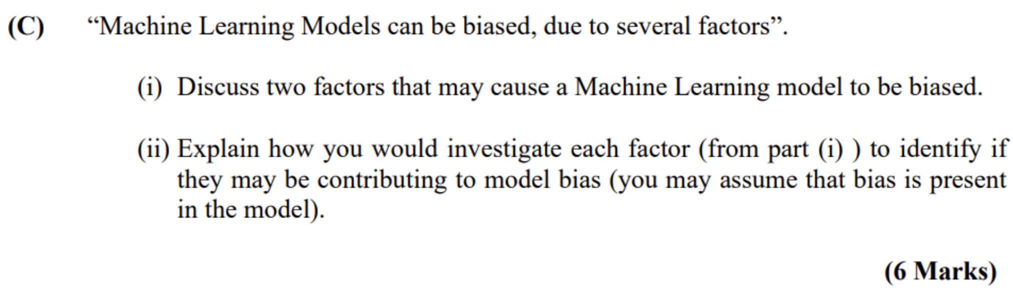
ii. An alternative to K-Fold Cross Validation is the hold out method. Or, Percentage Split.

This means we split the dataset into two, to help validate and test our model. We could split the dataset 60/40, using the first 60% for training and remaining 40% for testing. However, this could lead to some issues. The data in the training set could be too biased to the test set, leading to overfitting. While, the data in the training dataset could be too DIFFERENT to the test set, leading to underfitting.

K-Fold is a better technique because it tests both sets in 10 folds, while Percentage Split is less vigorous using only two folds.

Percentage Split is an easy technique to implement but it requires a high amount of computational time and it is not suited for an imbalanced dataset. There is also the fact a lot of data is isolated from the model, too. So, you will have to accommodate for that through rigorous testing. If this is an issue because you have a small dataset, K-Fold is a better technique to use as it uses both datasets to train and validate your model.

(C)



(C) "Machine Learning Models can be biased, due to several factors".

(i) Discuss two factors that may cause a Machine Learning model to be biased.

(ii) Explain how you would investigate each factor (from part (i)) to identify if they may be contributing to model bias (you may assume that bias is present in the model).

(6 Marks)

i. Missing data & it's removal:

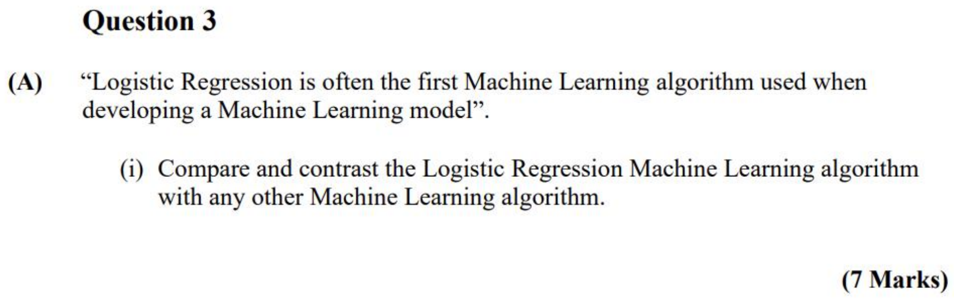
Removing missing data and outliers can significantly induce bias on the model. This may be in any direction, but at the same time, outliers and missing data in the dataset itself can induce bias also. This can happen either way.

Prejudice Bias & Diverse Data

In machine learning, ethics are very important. It is important to get diverse data within your dataset. There is a famous example of Amazon using an algorithm, but because their dataset was mostly male CVs, it became prejudiced towards women. It is really important to configure for this sort of thing as it may lead to machine learning models being racist or sexist, or even both.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Question 3:



Question 3

(A) “Logistic Regression is often the first Machine Learning algorithm used when developing a Machine Learning model”.

(i) Compare and contrast the Logistic Regression Machine Learning algorithm with any other Machine Learning algorithm.

(7 Marks)

i. Logistic Regression

- A supervised technique, training and test data.
- Used mostly for classification problems as well as regression problems
- Output must be a Categorical value
- No linear relationship required between dependent and independent variable.

Linear Regression:

- Supervised, like Logistic Regression. Training and test data.
- Used mostly for regression problems.
- Output must be a continuous value.
- Linear relationship required between dependent and independent variable.

(This is visible from Keith’s powerpoints. Straight line in linear regression, not a straight line in logistic.)

https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning

(B)

(i)

Logistic Regression is not applicable when you want the output to be a continuous variable. The output must be either 0 or 1. Or yes and no. It is binomial.

Logistic Regression requires the variables to have little to no multicollinearity amongst the independent variables.

(ii.)

No clue. My assumption is data scaling wouldn’t make a difference in performance.

(C)

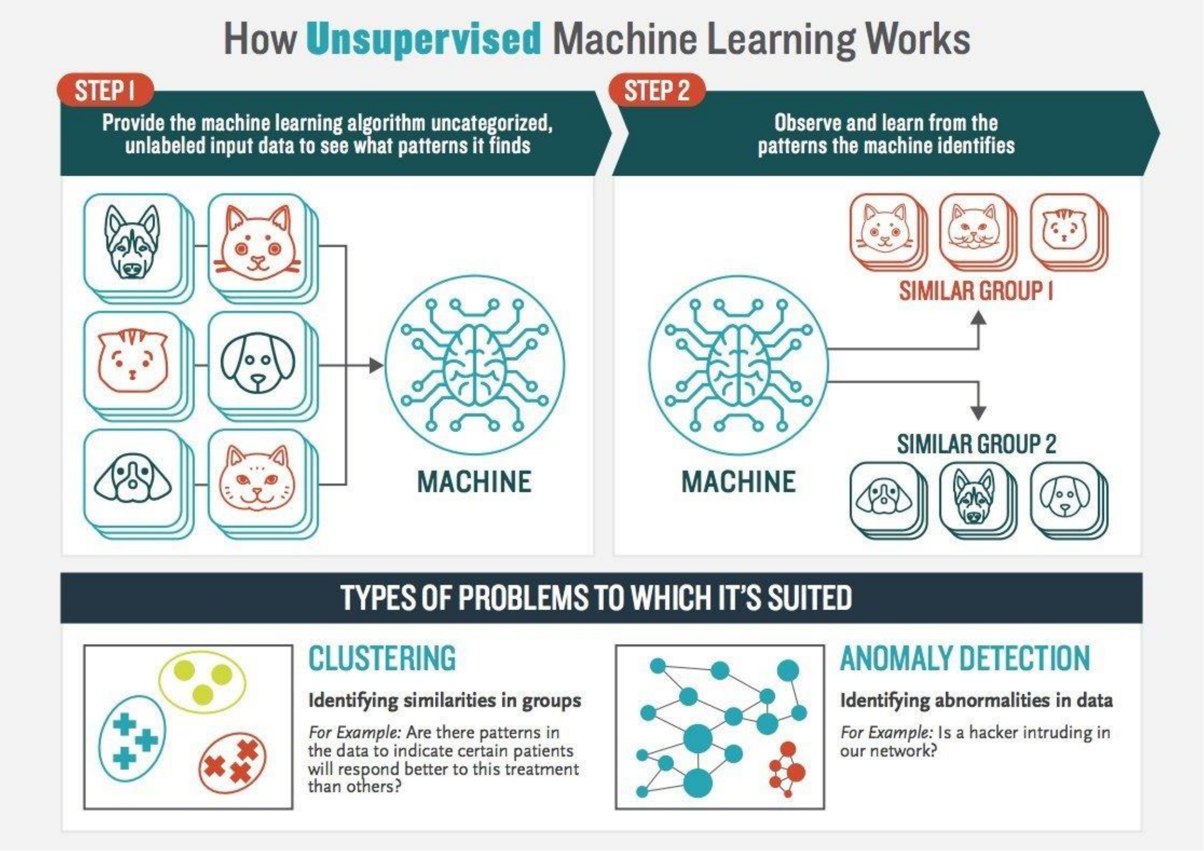
i.

Unsupervised machine learning is when you have only input data and no known output.

An example of this could be determining the difference between a cat and a dog or finding the differences between raw image data.

You could use an artificial neural network to help determine the difference, but I think you use a clustering algorithm first beforehand, to help identify the similarities between the two groups (in this case being a cat or a dog)

https://machinelearningmastery.com/how-to-develop-a-convolutional-neural-network-to-classify-photos-of-dogs-and-cats/


How Unsupervised Machine Learning Works
STEP 1
Provide the machine learning algorithm uncategorized, unlabeled input data to see what patterns it finds
STEP 2
Observe and learn from the patterns the machine identifies
MACHINE
MACHINE
SIMILAR GROUP 1
SIMILAR GROUP 2
TYPES OF PROBLEMS TO WHICH IT'S SUITED
CLUSTERING
Identifying similarities in groups
For Example: Are there patterns in the data to indicate certain patients will respond better to this treatment than others?
ANOMALY DETECTION
Identifying abnormalities in data
For Example: Is a hacker intruding in our network?