BERLIN SCHOOL OF BUSINESS & INNOVATION

Essay Title: Introduction to basic data analytics techniques

Name: Abijith M A

Date: 27/02/2023

Statement of compliance with academic ethics and the avoidance of plagiarism

I honestly declare that this essay is entirely my own work and none of its part has been copied from printed or electronic sources, translated from foreign sources and reproduced from essays of other researchers or students. Wherever I have been based on ideas or other people texts I clearly declare it through the good use of references following academic ethics.

(In the case that is proved that part of the essay does not constitute an original work, but a copy of an already published essay or from another source, the student will be expelled permanently from the postgraduate program).

Name and Surname (Capital letters): Abijith Mullancherry Asokan

Date: .27/02/2023

TABLE OF CONTENTS

INTRODUCTION

Machine learning is a field of study that focuses on developing algorithms and models that enable computer systems to automatically learn patterns and make predictions from data. There are primarily three types of machine learning, namely supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning involves training a model on a labeled dataset where the target variable is known. The goal of supervised learning is to develop a model that can make accurate predictions on new, unseen data. Examples of supervised learning include regression, classification, and decision tree algorithms.

Unsupervised learning involves training a model on an unlabeled dataset where the target variable is not known. The goal of unsupervised learning is to discover patterns and relationships in the data, such as clustering similar data points together or identifying anomalies in the data. The three methods of clustering, dimensionality reduction, and anomaly detection are examples of unsupervised learning.

Reinforcement learning involves training a model to learn through trial and error. The model interacts with an environment and learns by receiving feedback in the form of rewards or punishments for certain actions. The goal of reinforcement learning is to develop a model that can make optimal decisions in complex environments. Robotics and game playing are two examples of reinforcement learning.

CHAPTER 1: Supervised vs Unsupervised Learning

Supervised Learning:

Supervised learning involves training a model using labeled data in order to make predictions on brand-new, untainted data. In other words, using samples from a dataset that contains both the input characteristics and the matching output or target variable, the algorithm learns to map inputs to outputs.

How Supervised Learning Work?

The outcome of the supervised machine learning algorithm is already known. A mapping exists between the input and the output. Thus, the machine is provided with a lot of training input data in order to construct a model (having input and corresponding output known).

These are a few examples of supervised learning algorithms:

- Decision Trees
- K-Nearest Neighbor
- Support Vector Machine
- Linear Regression

Examples of Supervised Learning:

Classification: Algorithms are used in classification issues to divide the data into groups, such as true-false or more specific groups like apples and oranges. An example is whether to label an email as spam or not. Other categorization methods include Support Vector Machine and Decision Tree, among others.

Regression: Algorithms are used in regression issues to predict numerical values or establish a connection between input and output variables. One illustration of regression is weather forecasting. A regression algorithm is linear regression.

Unsupervised Learning:

Finding patterns and correlations in unlabeled data without any predetermined target variables is the goal of unsupervised learning, a form of machine learning technique.

The target values in this model are unidentified or unlabeled since there is no output that is linked to the input. The system must independently discover hidden patterns from the data that is sent to it.

How does Unsupervised Learning Work?

Certain strategies are used to mine data rules, patterns, and groupings of data with similar kinds since there are no known output values that can be utilized to establish a logical model between the input and output. These groupings assist the end users in finding a relevant result and in better comprehending the data.

Like training data, the inputs are not provided in an appropriate format (in supervised learning). It can include anomalies, noisy data, etc. The system receives all of these inputs at once. The inputs are clustered during the model training process.

The model automatically modifies its parameters as it goes through the process of identifying patterns in the data; for this reason, it is also known as self-organizing. By determining the commonalities between the inputs, the clusters will be created.

Examples of Unsupervised Learning:

Clustering: Data is clustered when it is grouped according to similarities or differences. One illustration of clustering is market segmentation. K-Means One of the clustering algorithms is clustering.

Association: The group of objects that appear together in the dataset is determined by association. Market Basket Analysis is a real-world illustration.

Key Differences between Supervised and Unsupervised Learning:

Goal: As contrast to unsupervised learning, which processes enormous amounts of data to uncover intriguing insights, patterns, and correlations existing in the data, supervised learning involves training the model using labeled data so that it predicts the proper output when given test data.

Output Feedback: Whereas unsupervised learning lacks a feedback mechanism since the model is oblivious of output, supervised learning provides a direct feedback mechanism because the machine is trained on labeled data.

Complexity: In comparison to supervised models, unsupervised models require more data to create outputs and insights, making them more complicated.

Applications: For tasks like spam detection, weather forecasting, and price prediction, supervised algorithms are excellent. Nonetheless, unsupervised algorithms are effective for medical imaging, recommendation engines, and anomaly detection.

CHAPTER 2: K Means Clustering

K-Means clustering is an unsupervised learning technique. This clustering does not employ labeled data, in contrast to supervised learning. K-Means groups objects into clusters that are distinct from one another while sharing characteristics.

K is a number, not a word. The number of clusters you need to build must be specified to the system. K = 2, for instance, designates two clusters. There is a method for determining the best or optimum value of K given a set of data.

Let's look at the procedures for making these clusters.

Distance Measures:

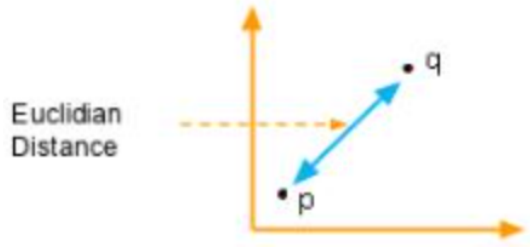
The distance between two pieces defines how similar they are and affects how clusters are constructed.

Several types of distance metrics are supported by K-Means clustering, including:

- **Euclidian Distance Measure:**

  The euclidean distance is a regular straight line if P and Q are two points. It is the separation in Euclidean space between the two points.

$$d=\sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}$$



- **Squared Euclidian Distance Measure:**

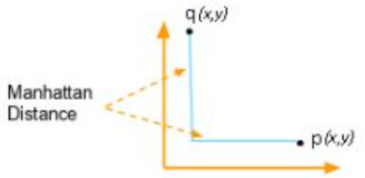
The only difference between this and the Euclidean distance measurement is that it does not include the square root.

$$d = \sum_{i=1}^{n} (q_i - p_i)^2$$

- **Manhattan Distance Measure:**

The Manhattan distance, or the distance between two places measured along axes at right angles, is the simple sum of the horizontal and vertical components.

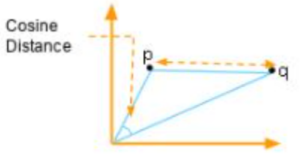
$$d = \sum_{i=1}^{n} |q_x - p_x| + |q_x - p_y|$$



- **Cosine Distance Measure:**

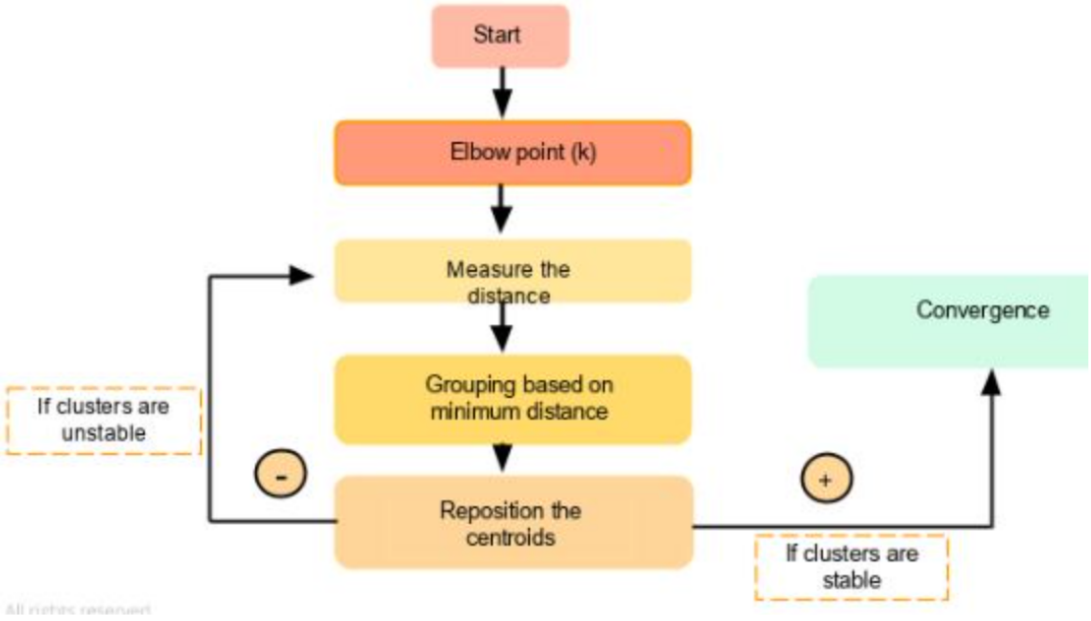
In this instance, the angle created by connecting the two vectors at the origin is used.

$$d = \frac{\sum_{i=0}^{n-1} q_i - p_x}{\sum_{i=0}^{n-1} (q_i)^2 \times \sum_{i=0}^{n-1} (p_i)^2}$$



Working of K-Means Algorithm:

The following flowchart illustrates how k-means clustering operates:



Step 1:

The Elbow approach is the most effective way to determine the cluster count. Running K-Means clustering on the dataset is what is referred to as the elbow approach.

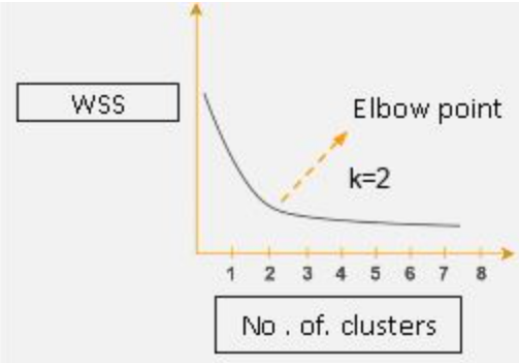
Finally, we determine the maximum number of clusters that may be generated for a particular data set using the within-sum-of-squares metric. the sum of the squared distances between each cluster member and its centroid is known as the within sum of squares (WSS).

$$WSS = \sum_{i=1}^{m} (x_i - c_i)^2$$

Where x_i = data point and c_i = closest point to centroid

For every K value, the WSS is calculated. The optimal value of K is chosen to have the least amount of WSS.

We now plot a curve between the number of clusters and WSS.



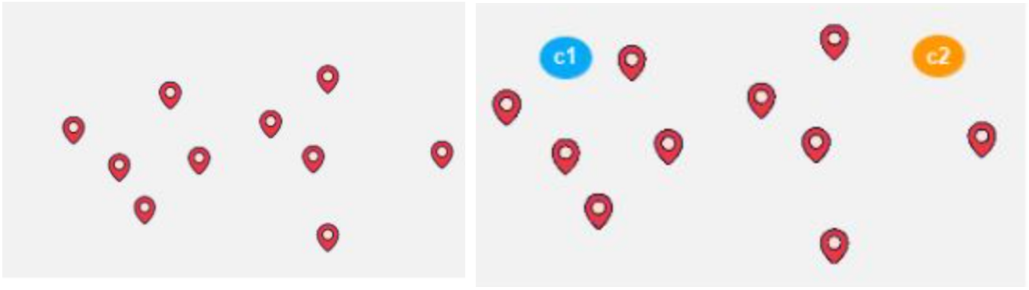
Here, the number of clusters is on the x-axis and WSS is on the y-axis.

When the K value rises from 2, you can observe that the value of WSS changes quite gradually.

Hence, the elbow point value may be used as the ideal K value. There ought to be no more than two, three, or four. After that, though, adding more clusters stabilizes the WSS value rather than drastically altering it.

Step 2:

Assume that these are where our deliveries will be made:



The cluster centroids are two positions that can be arbitrarily initialized.

Step 3:

Each location's separation from the centroid is now measured, and the data point closest to the centroid is given ownership of it.

Step 4:

Calculate the first group's actual centroid of data points.

Step 5:

Move the arbitrary centroid to the true centroid.

Step 6:

Calculate the second group's actual centroid of data points.

Step 7:

It is claimed that the k-means algorithm has converged after the cluster has reached stativity.

Centroids c1 and c2 make up the final cluster, as seen below:

CHAPTER 3: Decision Tree Classifier

Age	Sex	Education	Languages	Experience	Points	Accepted
25	M	3	2	3	4	No
22	F	4	1	2	3	No
21	F	3	2	5	1	No
29	F	4	3	4	5	Yes
24	M	5	4	7	4	Yes
26	M	2	2	8	4	Yes

As the name implies, a decision tree is a type of tree structure that follows the conditional logic. It is effective and contains powerful predictive analytic algorithms. Internal nodes, branches, and a terminal node are among its primary qualities.

With a collection of training feature vectors, we want to identify which characteristic is most effective in differentiating between the classes that need to be learnt.

- Knowledge gain reveals the significance of a specific feature vector characteristic.
- We'll use it to determine how to arrange the characteristics in a decision tree's nodes.

The Shannon Entropy is defined by the formula:

$$H(x) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Information Gain, or IG for short, divides a dataset according to a certain value of a random variable to calculate the decrease in entropy or surprise.

A higher information gain reflects a group or groups of samples with lower entropy, and so less surprise.

For our problem, we need to find the entropy of each column and the respective information gain.

Since all data are discrete values, we will convert them to continuous values to split.

Calculate the total entropy of the table.

$$H(x) = -P(yes)\log_2(P(yes)) - P(no)\log_2(P(no))$$

$$= -\frac{3}{6}\log_2(\frac{3}{6}) - \frac{3}{6}\log_2(\frac{3}{6})$$

$$= 1$$

Now, we need to find information gain from each feature of the table. To achieve the same, we would need to calculate the entropy of each column after splitting based on a splitting criterion since most columns are continuous values.

Age: We sort the table in ascending order first. After sortation, it can be seen that the target changes from "no" to "yes" when age changes from 22 to 24. We will take the splitting point to be the midpoint between them

Criteria	**Accepted**	**Not Accepted**
Age<23	0	2
Age>23	3	1

$$H_{Age}(x) = \frac{2}{6}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{4}{6}\left[-\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})\right]$$

$$= 0.541$$

Information Gain: $H(x) - H_{Age}(x) = 1 - 0.541 = 0.459$

Sex: Since Sex is a categorical value, the split can be done easily as M or F.

Criteria	**Accepted**	**Not Accepted**
Sex=M	2	1
Sex=F	1	2

$$H_{Sex}(x) = \frac{3}{6}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2(\frac{1}{3})\right] + \frac{3}{6}\left[-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})\right]$$

$$= 0.918$$

Information Gain: $H(x) - H_{Age}(x) = 1 - 0.918 = 0.082$

Education: The median split has been taken for calculating entropy for the feature.

Criteria	**Accepted**	**Not Accepted**
Education>=3.5	2	1
Education<3.5	1	2

$$H_{Education}(x) = \frac{3}{6}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2(\frac{1}{3})\right] + \frac{3}{6}\left[-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})\right]$$

$$= 0.918$$

Information Gain: $H(x) - H_{Education}(x) = 1 - 0.918 = 0.082$

Languages: Languages has been split at the median point.

Criteria	**Accepted**	**Not Accepted**
Languages>=2	3	2
Languages<2	0	1

$$H_{Education}(x) = \frac{5}{6}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2(\frac{2}{5})\right] + \frac{1}{6}\left[-\frac{1}{1}\log_2(\frac{1}{1})\right]$$

$$= 0.810$$

Information Gain: $H(x) - H_{Languages}(x) = 1 - 0.810 = 0.190$

Experience: Languages has been split at the median point.

Criteria	**Accepted**	**Not Accepted**
Experience>=4.5	2	1
Experience<4.5	1	2

$$H_{Experience}(x) = \frac{3}{6}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2(\frac{1}{3})\right] + \frac{3}{6}\left[-\frac{1}{3}\log_2(\frac{1}{3}) - \frac{2}{3}\log_2(\frac{2}{3})\right]$$

$$= 0.918$$

Information Gain: $H(x) - H_{Experience}(x) = 1 - 0.918 = 0.082$

Point: Points has been split at the median point.

Criteria	**Accepted**	**Not Accepted**
Points>=4	3	1
Points<4	0	2

$$H_{Points}(x) = \frac{4}{6}\left[-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2(\frac{1}{4})\right] + \frac{2}{6}\left[-\frac{2}{2}\log_2(\frac{2}{2})\right]$$

$$= 0.541$$

Information Gain: $H(x) - H_{Points}(x) = 1 - 0.541 = 0.459$

The information gain can be summarized as follows

Age	**Sex**	**Education**	**Languages**	**Experience**	**Points**
0.459	0.082	0.082	0.190	0.082	0.459

From the above table, it can be found that either Age or Points can be taken as the split criteria for root node splitting. We can take Age split as our root node. The result would contain two child nodes, where for one node the two values would be "no" for which no further split is required.

The table left with us for further split can be created as below:

Sex	Education	Languages	Experience	Points	Accepted
M	3	2	3	4	No
F	4	3	4	5	Yes
M	5	4	7	4	Yes
M	2	2	8	4	Yes

We need to calculate new entropy to perform further split.

$$H(x) = -P(yes)\log_2(P(yes)) - P(no)\log_2(P(no))$$

$$= -\frac{3}{4}\log_2(\frac{3}{4}) - \frac{1}{4}\log_2(\frac{1}{4})$$

$$= 0.811$$

Now repeat steps as done for the previous table.

Sex: Since Sex is a categorical value, the split can be done easily as M or F.

Criteria	**Accepted**	**Not Accepted**
Sex=M	2	1
Sex=F	1	0

$$H_{Sex}(x) = \frac{3}{4}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2(\frac{1}{3})\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2(\frac{1}{1})\right]$$

$$= 0.688$$

Information Gain: $H(x) - H_{Age}(x) = 0.811 - 0.688 = 0.123$

Education: The median split has been taken for calculating entropy for the feature.

Criteria	**Accepted**	**Not Accepted**
Education>=3.5	2	0
Education<3.5	1	1

$$H_{Education}(x) = \frac{2}{4}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right]$$

$$= 0.5$$

Information Gain: $H(x) - H_{Education}(x) = 0.811 - 0.5 = 0.311$

Languages: Languages has been split at the median point.

Criteria	**Accepted**	**Not Accepted**
Languages>=2.5	2	0
Languages<2,5	1	1

$$H_{Education}(x) = \frac{2}{4}\left[-\frac{2}{2}\log_2\left(\frac{2}{2}\right)\right] + \frac{2}{4}\left[-\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2})\right]$$

$$= 0.5$$

Information Gain: $H(x) - H_{Languages}(x) = 0.811 - 0.5 = 0.311$

Experience: It can be seen that when experience changes from 3 to 4, the outcome changes from "yes" to "no". Hence, we will split with the midpoint of 3 and 4.

Criteria	**Accepted**	**Not Accepted**
Experience>=3.5	3	0
Experience<3.5	0	1

$$H_{Experience}(x) = \frac{3}{4}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2(\frac{1}{1})\right]$$

$$= 0$$

Information Gain: $H(x) - H_{Experience}(x) = 0.811 - 0 = 0.811$

Point: Since there are only two values, we will split with midpoint of both.

Criteria	**Accepted**	**Not Accepted**
Points>=4.5	3	0
Points<4.5	0	1

$$H_{Points}(x) = \frac{3}{4}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] + \frac{1}{4}\left[-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right]$$

$$= 0$$

Information Gain: $H(x) - H_{Points}(x) = 0.811 - 0 = 0.811$

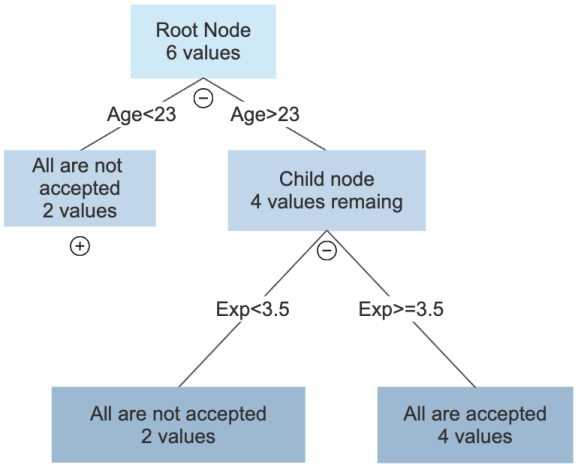
The information gain can be summarized as follows

Sex	**Education**	**Languages**	**Experience**	**Points**
0.123	0.311	0.311	0.811	0.811

If we use experience to split the node, it can be seen that the decision tree will have all terminal nodes as seen below.

It can be concluded that the tree has been split taking age as the root node with split at the value 23 and then experience has been taken at the next step at the split point of 3.5.

The decision tree can be visualized as follows.



CONCLUSIONS

In conclusion, machine learning is a powerful tool that can be used to make predictions, find patterns, and classify data. A model is trained on labeled data using supervised learning in order to generate predictions about fresh, unforeseen data. On the other hand, unsupervised learning involves finding patterns and relationships in unlabeled data without any specific target variables.

K-means clustering is a popular unsupervised learning algorithm used for clustering data points into groups based on their similarity. It works by iteratively assigning data points to the closest centroid, then updating the centroids based on the mean of the assigned data points. The process is repeated until convergence.

An example of a supervised learning method used for classification and regression problems are decision trees. They work by recursively splitting the data into smaller subsets based on the most significant feature until a stopping criterion is met. The result is a tree-like structure that can be used to make predictions on new data.

With the increasing availability of data and computational resources, machine learning is becoming more accessible and impactful in various industries, including healthcare, finance, and marketing.

BIBLIOGRAPHY

- Banoula, M. (2023) K-means clustering algorithm: Applications, types, and demos [updated]: Simplilearn, Simplilearn.com. Simplilearn. Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm (Accessed: February 27, 2023).
- Clustering algorithms - k-means algorithm (no date) Tutorials Point. Available at: https://www.tutorialspoint.com/machine_learning_with_python/clustering_algorithms_k_means_algorithm.htm (Accessed: February 27, 2023).
- Dwivedi, R. (no date) Introduction to decision tree algorithm in machine learning, Analytics Steps. Available at: https://www.analyticssteps.com/blogs/introduction-decision-tree-algorithm-machine-learning (Accessed: February 27, 2023).
- harshpreet0508, arvindpdmn (2022) Supervised vs unsupervised learning, Devopedia. Devopedia Foundation. Available at: https://devopedia.org/supervised-vs-unsupervised-learning (Accessed: February 27, 2023).
- Johnson, D. (2023) Supervised vs unsupervised learning: Difference between them, Guru99. Available at: https://www.guru99.com/supervised-vs-unsupervised-learning.html (Accessed: February 27, 2023).
- Types of machine learning: Supervised vs unsupervised learning (2023) Software Testing Help. Available at: https://www.softwaretestinghelp.com/types-of-machine-learning-supervised-unsupervised/ (Accessed: February 27, 2023).
- Understanding K-means clustering in machine learning (no date) Zilliz Vector database blog. Available at: https://zilliz.com/blog/k-means-clustering (Accessed: February 27, 2023).