A green circle with black text

Description automatically generated

Name: Amer Samamh



The DML program

An easier way to visualize analyses and apply machine learning Data technics for your data.

The project has started with the Data mining course that was held in the first semester of 2024. Since the we have enhanced the model and added much more functionality that we will discussed in this documentation.

**THE GUI**

We have used the (tkinter) library witch provide an easy way to make a simple yet effective GUI .

**Tkinter** is the standard Python interface to the Tcl/Tk GUI toolkit12. It is the most commonly used library for developing GUI (Graphical User Interface) in Python2. As Tk and Tkinter are available on most Unix platforms, including macOS, as well as on Windows systems, developing GUI applications with Tkinter becomes fast and easy2.

The tkinter package is not a thin wrapper but adds a fair amount of its own logic to make the experience more pythonic1. It provides a set of wrappers that implement the Tk widgets as Python classes34.

Running python -m tkinter from the command line should open a window demonstrating a simple Tk interface, letting you know that tkinter is properly installed on your system1.

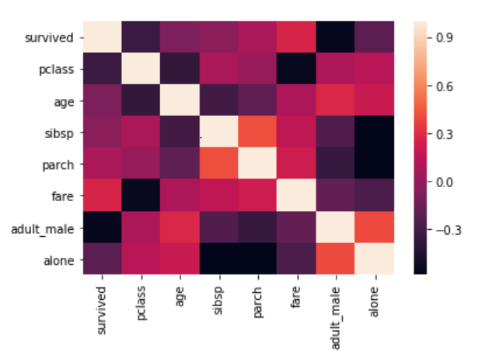
It supports a range of Tcl/Tk versions, built either with or without thread support1. The official Python binary release bundles Tcl/Tk 8.6 threaded1.

Tkinter allows you to create a variety of widgets, such as labels, buttons, and text boxes, in your application. It also provides methods for organizing these widgets in a variety of ways to create your desired layout2.

**data visualization**

Data visualization where integrated in various parts in our program but it also has its own dedicated page, why is that? well the data visualization is very important a step in the process of understanding the data that is provided to do the correct data processing for it later on

SEABORN: we have used the Seaborn since it is a powerful tool for data visualization in Python, providing an intuitive and high-level interface for creating beautiful and informative statistical graphics.

**1.A heat map** : heat maps are a powerful tool in data visualization,

providing an intuitive means to understand complex datasets at a glance.

They are widely used in various fields, it will find the correlation between

Every variable in the data . and the highst correlation will be 1 (the data with

Itself and if the correlation is 0 there are no correlations between the data

A graph showing a red line and blue dots

Description automatically generated**2.ScatterPlot:**

A scatter plot in Seaborn is a type of data visualization that uses

dots to represent the values obtained for two different variables

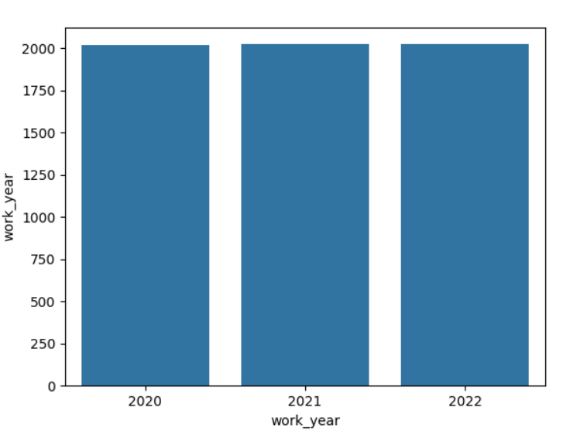
one plotted along the x-axis and every other variable in data well

be considered the Y-axis.

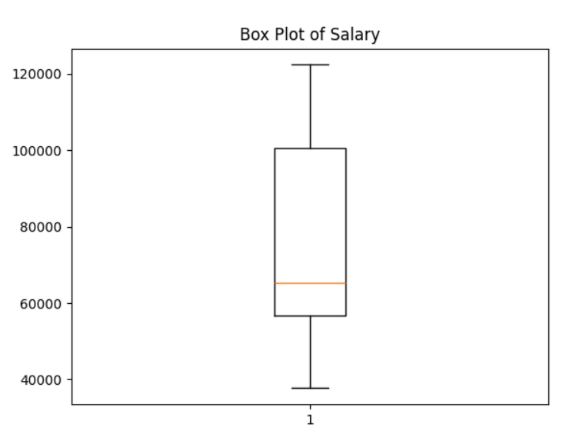
Scatter plots are particularly useful for visualizing

the relationship between two numerical variables where

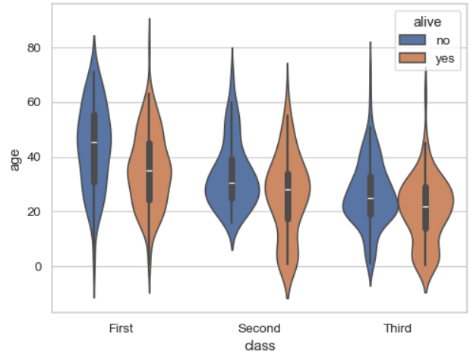
each dot represents an observation in the dataset.

**3. BarPlot:** is used for the variable to find the count of

each attribute it has .

**4.boxPlot:**used to know the Q1,Q2,Q3, can see

the outlier values and the min max values it has

****

**5.volion plot:** depicts distributions of numeric

data for one or more groups using density curves

A graph with a line and a red line

Description automatically generated**6.** **Linear regression:** Linear regression is a statistical

method that allows us to study the relationship between two

continuous variables, for example we can find the outliers

between two variables and manually remove them.

**Data Preprocessing**

**In the data preprocessing section we have a few functionality that we will discuss with you**

**1.show data button : it will allow you to see the data in a table as you see it the excel file any new data you would like to add or remove will be there to see and comeback for**

**2.Show statics :provides a**  **Statistics Description Mean,Median,Mode,Standard Deviation and Variance for the data we have**

**3.smoothing : we provide the functionality to smooth the data by mean ,median or bounders**

**Choosing for your self the ben size**

**4.normalztion: provides the min max ,z score , or decimal number .**

**5.droping a column: made a new column that you don’t have use for ? no problem u can just drop it**

**Supervised learning**

Supervised learning is a machine learning technique that uses labeled data to train algorithms. Let me break it down for you:

Labeled Data: In supervised learning, we have a dataset where each example (or observation) is associated with a known output (also called a label). For instance, if we’re building a spam filter, the dataset would contain emails labeled as either “spam” or “not spam.”

Training Process: During training, the algorithm learns from this labeled data. It adjusts its internal parameters (often represented as weights) to map input features (such as words in an email) to the correct output (spam or not spam). The goal is to minimize the error between the predicted output and the actual label.

Types of Problems:

Classification: In classification problems, the algorithm assigns input data to specific categories or classes. For example, classifying emails as spam or not spam.

Regression: Regression deals with predicting a continuous value (e.g., predicting house prices based on features like square footage and location).

Algorithms in our programs will be discussed further here is a short explanation :

Linear Classifiers: These algorithms create a linear boundary to separate different classes.

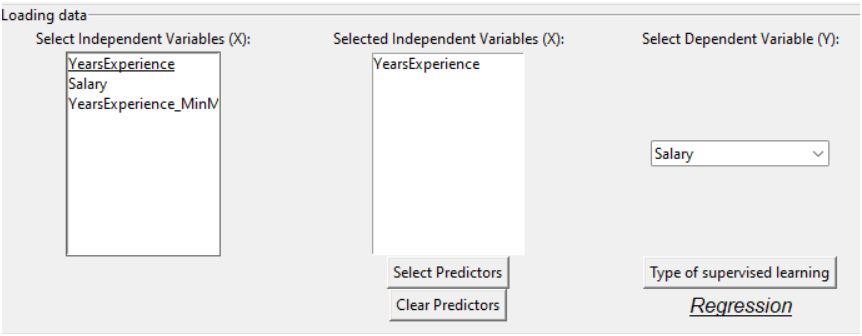
Support Vector Machines (SVM): SVMs find the best hyperplane to separate data points.

k-Nearest Neighbors (k-NN): k-NN assigns a label based on the majority class among its k nearest neighbors.

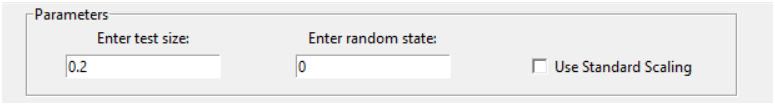
------------------------------------------------------------------------------------------------------------------

**firstly,** the use has to specify the X values (the independent values) and the Y values

(the dependent values) we have created a simple selection for the user to prompt.

****

* Is the data suit for Binary Classification ,Multi-class Classification or Regression? Not sure? No problem we can tell what it is just use the button (Type of supervised learning and we will tell what is the best model for your data



Why Split Data into Train and Test Sets?

**Why Split Data into Train and Test Sets?**

1.Training the Model: The training set is used to train the machine learning algorithm. The model learns from this data, identifying patterns and building its predictive capabilities.

2.Testing the Model: The test set is used to evaluate the performance of the trained model. It provides a measure of how well the model generalizes to new, unseen data.

Importance of Splitting Data:

Avoid Overfitting: Without a proper split, the model might overfit, meaning it performs exceptionally well on the training data but fails to generalize to new data. Overfitting occurs when the model learns the noise and details in the training data to an extent that it negatively impacts its performance on new data.

Measure Model Performance: By evaluating the model on the test set, we get an estimate of its real-world performance. This helps in understanding how well the model will perform on new, unseen data.

Consequences of Not Splitting Data:

Overfitting: The model will likely memorize the training data, resulting in high accuracy on training data but poor performance on new data.

Inaccurate Performance Metrics: Without a test set, we cannot reliably measure the model's performance, making it difficult to assess its effectiveness.

By properly splitting the data into training and testing sets, we ensure that our model is robust, generalizes well to new data, and performs reliably in real-world scenarios.

**What is random state?**

General Definition: random state is a parameter used to control the randomness in various computational processes. It can be found in numerous libraries and functions where random number generation is involved, such as data splitting, random sampling, model training, and others.

Why is random state Important?

Reproducibility: In any computational process involving randomness, setting a random state ensures that you can reproduce the same results every time you run the code. This is crucial for:

Scientific Research: Ensuring that experiments can be replicated and verified by others.

Debugging: Allowing developers to consistently reproduce and diagnose issues.

Comparing Models: Ensuring fair and consistent comparisons between different models or configurations.

Determinism: By fixing the random seed, processes that involve randomness become deterministic. This means that the same sequence of random numbers will be generated each time, leading to consistent behavior.

How random state Works

Random Number Generators (RNGs): Many computational processes use RNGs to perform tasks that require randomization. The random state acts as a seed for these RNGs. When the same seed is used, the RNG produces the same sequence of numbers, leading to reproducible results.

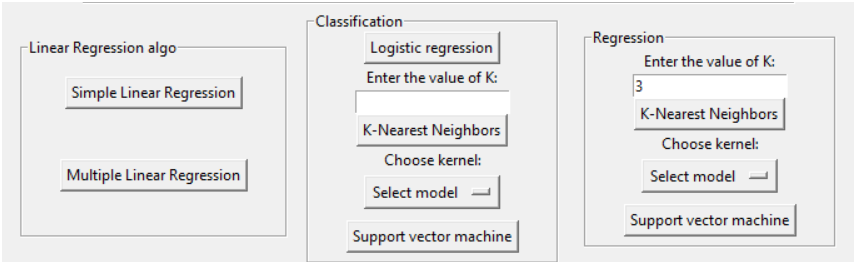
Standard scaling, often referred to as standardization or Z-score normalization, is a data preprocessing technique used to transform the features of a dataset so that they have the properties of a standard normal distribution with a mean of 0 and a standard deviation of 1.

**Why Standard Scaling**

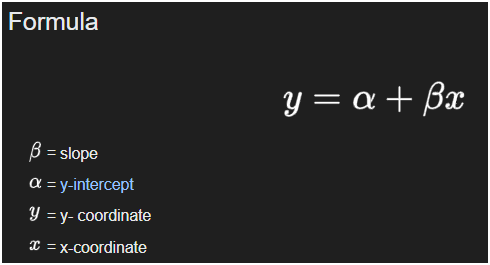
1.Consistency: Different features might be measured on different scales (e.g., age in years, income in dollars). Standard scaling brings all features onto a common scale without distorting differences in the ranges of values.

2.Performance: Many machine learning algorithms, especially those involving gradient descent (e.g., logistic regression, neural networks) and distance-based algorithms (e.g., k-nearest neighbors, SVMs), perform better when features are on a similar scale. This helps the algorithms converge faster and achieve better performance.

3.Comparability: Standard scaling makes it easier to compare the importance of different features.



* **Linear regression algorithm**
* **Simple Linear regression :** **Models the relationship between one independent variable and one dependent variable using a straight line.**

****

* **Multiple linear regression:** **Models the relationship between two or more independent variables and one dependent variable using a plane or hyperplane.**

**Formula Y = a + b1X1 + b2X2**

* **Classification section**

**-logistic regression: this button varies of thing**

**1.Prepares the data by extracting the specified features and target, encoding if necessary, and optionally scaling the features.**

**2.Splits the data into training and testing sets.**

**3.Trains a logistic regression model on the training set.**

**4.Predicts the target variable for the test set.**

**5.Calculates and displays the accuracy and confusion matrix(explained later).**

**6.Visualizes the confusion matrix using a heatmap.**

**confusion matrix** is a performance measurement tool for machine learning classification models. It provides a detailed breakdown of the predictions made by the model compared to the actual outcomes, allowing you to understand the types of errors the model is making. The confusion matrix is a powerful tool for understanding the performance of a classification model, providing insight into not just the overall accuracy but also the types and frequencies of classification errors. It helps in diagnosing the strengths and weaknesses of the model, enabling more informed decisions for model improvement.

* **K-Nearest Neighbors:** is a straightforward and versatile algorithm used for classification. It is based on the idea that points close to each other in the feature space are likely to have similar labels. Despite its simplicity, KNN can perform well for a variety of tasks, particularly when the number of dimensions is relatively low, and the dataset size is manageable.
* **How it works?**

Parameter Selection (K):

- Choose the number of neighbors ( K ). This is the number of closest training samples that will be used to make predictions.

2. Distance Calculation:

- Calculate the distance between the new input (the query point) and all the points in the training

- Euclidean Distance

3. Find Neighbors:

- Identify the ( K ) closest neighbors to the query point based on the calculated distances.

4. Voting (for classification) :

- Classification: The class of the query point is determined by majority voting among the ( K ) nearest neighbors. The class that appears most frequently among the neighbors is assigned to the query point.

5. Output the Result:

- The class label for classification is returned as confustion matrix.

- **Support Vector Machines (SVM)** are a set of supervised learning methods used for classification, regression, and outlier detection. The primary goal of an SVM is to find a hyperplane that best separates the classes in the feature space. SVMs are particularly effective in high-dimensional spaces and are versatile due to the use of different kernel functions

**How SVM for classification Works ?**

1. Linear SVM:

- SVM aims to find the optimal hyperplane that separates the data points of different classes. This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the closest data points from each class (called support vectors).

2. Non-linear SVM:

- When the data is not linearly separable, SVM can use a kernel function to map the data into a higher-dimensional space where a hyperplane can be used to separate the classes.

Kernels in SVM

- Kernels are functions that enable SVM to perform in higher-dimensional spaces without explicitly transforming the data points. They do this by computing the inner products between the images of all pairs of data in the feature space. This allows SVM to fit complex boundaries.

models we chose for the kernel

1-liner :- Used when the data is linearly separable.

2-poly : - Represents the similarity of vectors in a polynomial feature space.

3-rbf :A small gamma value will result in a wide kernel, leading to a simpler model with low variance and high bias

Imports:

Import necessary libraries for SVM classification, including SVC for the SVM classifier, confusion\_matrix and accuracy\_score for evaluation metrics, and MinMaxScaler for feature scaling.

Data Preparation:

x\_vars, y\_var, test\_var, and rand\_var are obtained from user inputs or GUI elements. They likely represent the indices or names of selected features, the target variable, the test size proportion, and the random state for reproducibility.

X and y are extracted from the DataFrame df, presumably containing the feature variables and target variable, respectively.

Data encoding and scaling are performed using custom functions (encodingc and MinMaxScaler).

The data is split into training and testing sets using train\_test\_split().

Feature Scaling:

If use\_scaling is True, the features are standardized using StandardScaler to have zero mean and unit variance.

Model Training:

An SVM classifier (SVC) is initialized with parameters such as the specified kernel (e.g., linear, polynomial, or radial basis function) and random\_state.

The classifier is trained on the training data (X\_train, y\_train) using the fit() method.

Prediction and Evaluation:

Predictions are made on the test data using the predict() method.

Accuracy is calculated using accuracy\_score.

A confusion matrix (cm) is computed using confusion\_matrix.

A heatmap visualization of the confusion matrix is created using sns.heatmap().

Visualization:

The confusion matrix heatmap is displayed using plt.show().

Overall, this function encapsulates the entire process of SVM classification, from data preparation to model training, evaluation, and visualization of results. It provides a convenient way to perform SVM classification with adjustable parameters and visual feedback on model performance.

* Regression section
* - **K-Nearest Neighbors:**
* **How k-NN for regression works?**

Parameter Selection (K):

- Choose the number of neighbors ( K ). This is the number of closest training samples that will be used to make predictions.

2. Distance Calculation:

- Calculate the distance between the new input (the query point) and all the points in the training

- Euclidean Distance

3. Find Neighbors:

- Identify the ( K ) closest neighbors to the query point based on the calculated distances.

4. Voting (for classification) :

- Classification: The class of the query point is determined by majority voting among the ( K ) nearest neighbors. The class that appears most frequently among the neighbors is assigned to the query point.

5. Output the Result:

- The class label for classification is returned as liner graph .

**How SVM for classification Works ?**

1. Linear SVM:

- SVM aims to find the optimal hyperplane that separates the data points of different classes. This hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the closest data points from each class (called support vectors).

2. Non-linear SVM:

- When the data is not linearly separable, SVM can use a kernel function to map the data into a higher-dimensional space where a hyperplane can be used to separate the classes.

Kernels in SVM

- Kernels are functions that enable SVM to perform in higher-dimensional spaces without explicitly transforming the data points. They do this by computing the inner products between the images of all pairs of data in the feature space. This allows SVM to fit complex boundaries.

models we chose for the kernel

1-liner :- Used when the data is linearly separable.

2-poly : - Represents the similarity of vectors in a polynomial feature space.

3-rbf :A small gamma value will result in a wide kernel, leading to a simpler model with low variance and high bias

Data Preparation:

Obtain input variables (x\_vars) and target variable (y\_var) from user inputs or GUI elements.

Extract features (X) and target (y) from the DataFrame df.

Encode categorical variables and scale features using StandardScaler.

Splitting Data:

Split the data into training and testing sets using train\_test\_split().

Model Training:

Initialize SVR regressor with specified kernel.

Fit the regressor to the training data using fit() method.

Prediction and Evaluation:

Predict target values for the test data using predict() method.

Invert the scaling of predicted and actual target values.

Calculate mean squared error (MSE) between actual and predicted values.

Visualization:

Plot the actual vs. predicted values as a scatter plot.

Plot the decision boundary if the feature space is 2D using contour plot.

Display legend for different classes in the scatter plot.

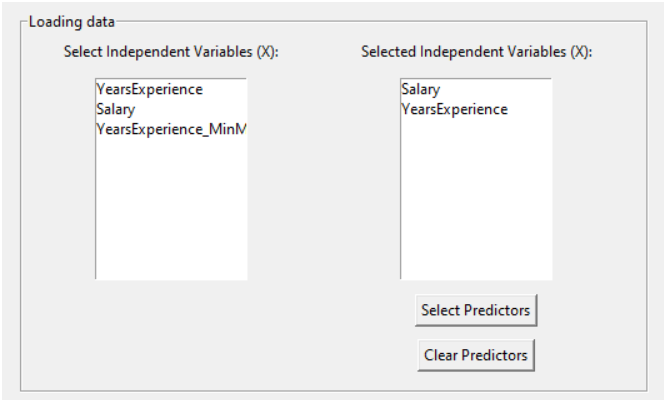
Error Handling:

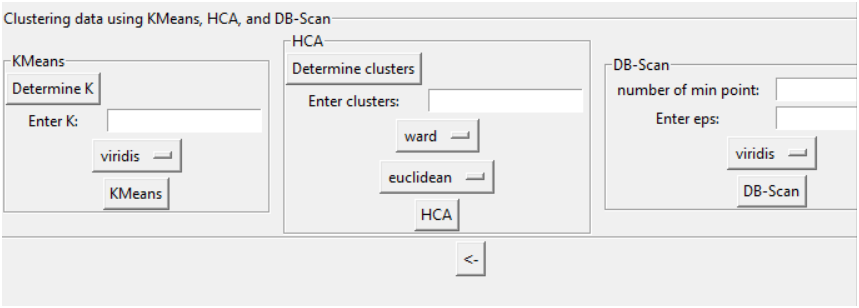
If the feature space is not 2D, display an error message indicating that the decision boundary visualization is only available for 2D feature spaces.

Overall, this function encapsulates the process of SVR regression, from data preparation to model training, evaluation, and visualization of results. It provides a comprehensive view of the regression performance and, if applicable, the decision boundary in the feature space.

**Unsupervised learning**

**firstly,** the use has to specify the X values (the independent values) since its unsupervised learning there is no Y .

****



Lets dive into the clustering data using Kmeans , Hca and DB-Scan .

* **K-means** clustering is a popular unsupervised machine learning algorithm used for partitioning a dataset into K distinct, non-overlapping clusters. It is widely used in various applications such as image segmentation, document clustering, and customer segmentation.

**-not sure whats the best number of K?**

We got you covered use the determen K button and it will result in a graph that allows you to to analysis the best number of k

How it works? Obtain input variables (`x\_vars`) from some source (possibly user input or GUI elements).

- Extract features (`X`) from the DataFrame `df`.

3. Elbow Method:

- Initialize an empty list `wcss` (Within-Cluster Sum of Squares).

- Iterate through a range of values for \( K \) from 1 to 10.

- For each value of \( K \), perform K-means clustering with `n\_clusters=i`.

- `init='k-means++'` initializes the centroids in a smart way to speed up convergence.

- `random\_state=42` ensures reproducibility of results.

- Compute the Within-Cluster Sum of Squares (WCSS) for each \( K \) and append it to the `wcss` list.

4. Plotting:

- Plot the number of clusters (\( K \)) on the x-axis and the corresponding WCSS on the y-axis.

- The plot shows the "elbow" point, which is the point of inflection where the rate of decrease in WCSS sharply changes.

- The optimal number of clusters can be chosen at this "elbow" point, as it represents the point where adding more clusters does not significantly decrease WCSS.

5. Plot Customization:

- Set plot title, x-axis label, and y-axis label.

- Show the plot.

Overall, this function provides a visual aid for selecting the optimal number of clusters in K-means clustering based on the Elbow Method, allowing for more informed decisions about the clustering process.

**The Elbow Method** : is a heuristic technique used to determine the optimal number of clusters (\( K \)) in a dataset for K-means clustering. It is based on the concept that as the number of clusters increases, the within-cluster sum of squares (WCSS) decreases. The optimal number of clusters is often identified at the "elbow point" in a plot of the WCSS versus the number of clusters.

**-What is cmap?** **camp is short for "colormap" in Matplotlib. It is a parameter used to specify the color map to use when plotting data that requires color mapping, such as scatter plots, contour plots, or heatmaps.**

**A colormap is a mapping from a range of data values to a range of colors. It provides a way to visually represent data by assigning colors to different data values. Colormaps can be sequential, diverging, or qualitative, depending on the nature of the data being visualized.**

**How it works ?**

1.Data Preparation:

Obtain input variables (x\_vars) from user inputs or GUI elements.

Extract features (X) from the DataFrame df.

Obtain the number of clusters (k) from user input.

2.K-means Clustering: Initialize KMeans object with the specified number of clusters (n\_clusters=k).

Fit KMeans to the data using the fit() method.

Assign cluster labels to the original dataset by adding a new column 'Cluster' containing the cluster labels.

3.Dimensionality Reduction:Perform dimensionality reduction using PCA to visualize the clusters in a 2D space.

Fit PCA to the original feature matrix X and transform it to a lower-dimensional space (reduced\_X).

4.Visualization:Scatter plot of the reduced data points (reduced\_X) colored by cluster labels (df['Cluster']) using a colormap (cmapk).

Scatter plot of cluster centroids transformed to the reduced space.

Set axis labels and plot title.

Display legend.

5.Display Visualization: Show the plot. Message Box:Display a message box indicating that KMeans clustering has been applied and clusters have been added to the original dataset.

* HCA**(Hierarchical clustering):** is a method of cluster analysis which seeks to build a hierarchy of clusters. It is an unsupervised learning technique that groups similar objects into clusters based on their pairwise distances. Hierarchical clustering methods can be broadly categorized into two types: Agglomerative and Divisive.

not sure what is the number of cluster?

use the determan cluster button it will provide the graph of the number of cluster you should use

how it works?

This code defines a function named deter\_hca() that performs hierarchical clustering using the Ward's method and visualizes the resulting dendrogram.

Here's a breakdown of each part of the code:

1.Data Preparation:

Obtains input variables (x\_vars) from some source (possibly user input or GUI elements).

Hierarchical Clustering:

Use the linkage() function from scipy.cluster.hierarchy to perform hierarchical clustering on the data.

The method used for linkage is set to 'ward', which minimizes the variance when merging clusters. Ward's method is known for producing balanced dendrograms.

Dendrogram Visualization:

Generate a dendrogram using the dendrogram() function from scipy.cluster.hierarchy.

The dendrogram() function takes the output of the linkage() function and plots the hierarchical clustering dendrogram.

The dendrogram visually represents the hierarchical clustering process, showing how clusters are merged as the algorithm progresses.

Plot Customization:

Set the title of the plot to 'Dendrogram'.

Label the x-axis as 'Customers', which might be specific to the dataset being analyzed.

Label the y-axis as 'Euclidean distances', indicating the distance metric used for clustering.

Display the Dendrogram:

Show the plot containing the dendrogram using plt.show() from Matplotlib.

Overall, this function provides a straightforward way to perform hierarchical clustering using Ward's method and visualize the resulting dendrogram, allowing for insights into the clustering structure of the dataset and the relationships between data points.

Lets dive into **HCA** algorithm

**linkage** :In hierarchical clustering, "linkage" refers to the method used to calculate the distance between clusters when merging or comparing them. The choice of linkage method significantly impacts the resulting clustering structure. Several linkage methods are commonly used in hierarchical clustering algorithms:

1.Single Linkage (Minimum Linkage):

In single linkage, the distance between two clusters is defined as the shortest distance between any two points in the two clusters.

It tends to produce clusters with elongated shapes and is sensitive to outliers and noise in the data.

Complete Linkage (Maximum Linkage):

2.In complete linkage, the distance between two clusters is defined as the longest distance between any two points in the two clusters.

It tends to produce compact, spherical clusters but can be sensitive to outliers and may create clusters of unequal sizes.

3.Average Linkage (UPGMA):

In average linkage, the distance between two clusters is defined as the average distance between all pairs of points in the two clusters.

It tends to produce clusters with similar sizes and shapes and is less sensitive to outliers compared to complete linkage.

4.Ward's Linkage:

Ward's method minimizes the variance when merging clusters. It computes the sum of squared differences within all clusters before and after merging and selects the merge that minimizes the increase in total variance.

It tends to produce clusters of relatively equal size and shapes, making it suitable for many datasets

**distance metric**: the choice of distance metric plays a crucial role in determining the similarity between data points or clusters. Different distance metrics capture different aspects of similarity or dissimilarity between data points. Here are explanations of some commonly used distance metrics:

1. Euclidean Distance:

- Euclidean distance is the straight-line distance between two points in a Euclidean space.

- It is the most commonly used distance metric and measures the length of the line segment connecting two points in a multidimensional space.

- Mathematically, the Euclidean distance between two points \( P = (p\_1, p\_2, ..., p\_n) \) and \( Q = (q\_1, q\_2, ..., q\_n) \) in an \( n \)-dimensional space is given by:

\[ \text{Euclidean distance} = \sqrt{\sum\_{i=1}^{n} (q\_i - p\_i)^2} \]

2. Manhattan Distance (City Block Distance):

- Manhattan distance is the sum of the absolute differences between the coordinates of corresponding points.

- It measures the distance between two points along the grid-like paths formed by the lines parallel to the axes.

- Mathematically, the Manhattan distance between two points \( P \) and \( Q \) is given by:

\[ \text{Manhattan distance} = \sum\_{i=1}^{n} |q\_i - p\_i| \]

3. Jaccard Distance:

- Jaccard distance measures the dissimilarity between two sets by comparing their intersection and union.

- It is commonly used for binary or categorical data where each data point represents a set of features.

- Mathematically, the Jaccard distance between two sets \( A \) and \( B \) is given by:

\[ \text{Jaccard distance} = 1 - \frac{|A \cap B|}{|A \cup B|} \]

4. Cosine Similarity:

- Cosine similarity measures the cosine of the angle between two vectors in a multidimensional space.

- It quantifies the similarity between two vectors regardless of their magnitudes.

- It is commonly used for text mining and high-dimensional data analysis.

- Mathematically, the cosine similarity between two vectors \( \mathbf{u} \) and \( \mathbf{v} \) is given by:

\[ \text{Cosine similarity} = \frac{\mathbf{u} \cdot \mathbf{v}}{\| \mathbf{u} \| \| \mathbf{v} \|} \]

5. Minkowski Distance:

- Minkowski distance is a generalization of both Euclidean and Manhattan distances.

- It can be tuned using a parameter \( p \), where \( p = 1 \) gives Manhattan distance, \( p = 2 \) gives Euclidean distance, and \( p \rightarrow \infty \) gives Chebyshev distance.

- Mathematically, the Minkowski distance between two points \( P \) and \( Q \) is given by:

\[ \text{Minkowski distance} = \left( \sum\_{i=1}^{n} |q\_i - p\_i|^p \right)^{\frac{1}{p}} \]

The choice of distance metric depends on the nature of the data and the specific requirements of the clustering task. Different distance metrics may lead to different clustering results.

**DB-Scan:** Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular clustering algorithm used in machine learning and data mining. It is particularly effective at identifying clusters of arbitrary shapes in spatial data while also being robust to noise.

**MinPts**:MinPts is a parameter used in DBSCAN to determine whether a data point is a core point.

A core point is a point that has at least MinPts neighboring points (including itself) within a distance of Eps.

If a point has fewer than MinPts neighboring points within Eps, it is labeled as noise or an outlier.

Increasing MinPts leads to fewer core points and potentially smaller but more cohesive clusters, whereas decreasing MinPts may result in more core points and larger but more fragmented clusters.

**Eps (epsilon):**Eps defines the maximum distance that specifies the neighborhood of a point.

Any two points within a distance of Eps are considered neighbors.

Eps influences the size and shape of clusters detected by DBSCAN.

Choosing an appropriate value for Eps is critical as it determines the scale at which clusters are identified. Too small an Eps may lead to sparse clusters, while too large an Eps may merge distinct clusters or classify too many points as noise.

Both MinPts and Eps are crucial parameters in DBSCAN, and their values need to be carefully selected based on the characteristics of the dataset and the desired clustering outcome. Adjusting these parameters can have a significant impact on the clustering results, affecting the number and shape of clusters identified, as well as the handling of noise and outliers.

Cmap is the same as we discussed in the K mean function

**Resources:**

Tkinter Resources:

[Python Tkinter Tutorial - GeeksforGeeks](https://www.geeksforgeeks.org/python-tkinter-tutorial/)

[tkinter — Python interface to Tcl/Tk — Python 3.12.3 documentation](https://docs.python.org/3/library/tkinter.html)

Seaborn Documentation and Tutorials:

[seaborn: statistical data visualization — seaborn 0.13.2 documentation (pydata.org)](https://seaborn.pydata.org/)

[Data Visualization using Seaborn. I am back with the seaborn tutorial… | by Mohit Sharma | Towards Data Science](https://towardsdatascience.com/data-visualization-using-seaborn-fc24db95a850)

Data Preprocessing:

[6.3. Preprocessing data — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/preprocessing.html)

[Data Preprocessing in Data Mining & Machine Learning | by Tarun Gupta | Towards Data Science](https://towardsdatascience.com/data-preprocessing-in-data-mining-machine-learning-79a9662e2eb)

Supervised Learning Algorithms:

1. [Supervised learning — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/supervised_learning.html)

[Machine Learning Mastery](https://machinelearningmastery.com/)

Unsupervised Learning Algorithms:

[Unsupervised learning — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/unsupervised_learning.html)

[A Brief Introduction to Unsupervised Learning | by Aidan Wilson | Towards Data Science](https://towardsdatascience.com/a-brief-introduction-to-unsupervised-learning-20db46445283)

K-means Clustering:

[KMeans — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) [K-means: A Complete Introduction. K-means is an unsupervised clustering… | by Alan Jeffares | Towards Data Science](https://towardsdatascience.com/k-means-a-complete-introduction-1702af9cd8c)

Hierarchical Clustering (HCA):

[2.3. Clustering — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/clustering.html)

[Understanding the concept of Hierarchical clustering Technique | by Chaitanya Reddy Patlolla | Towards Data Science](https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec)

DBSCAN:

[DBSCAN — scikit-learn 1.5.0 documentation](https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html)

[DBSCAN Clustering — Explained. Detailed theorotical explanation and… | by Soner Yıldırım | Towards Data Science](https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556#:~:text=e.g.%20anomaly%20detection.-,DBSCAN%20algorithm,many%20points%20from%20that%20cluster.)