Fr. Conceicao Rodrigues College of Engineering, Bandra

Department of Computer Engineering

SEM-VI

DATA WAREHOUSING & MINING

PROJECT REPORT

Topic
REACTIVE MINER

By

8312	**Abhishek Ahirrao**	**8320**	**Carol Sebastian**
8316	**Princeton Baretto**	**8322**	**Pratik Chowdhury**
8317	**Amurto Basu**	**8323**	**Rahim Chunara**
8318	**Shubham Bhate**	**8324**	**Ariane Correa**
8319	**Simran Bindra**	**8330**	**Ria Dmello**

Teacher-in-Charge
Prof. Dr. Sujatha Deshmukh

TABLE OF CONTENTS

PROBLEM STATEMENT:

Design of a graphical interface to visualize the working of various Data Mining Algorithms.

Color and geometry representations are easily recognized and interpreted by the human brain. Hence, data visualization technologies provide data mining results with natural and intuitive operation interfaces. Currently, data miners always plot 2-dimensional and 3-dimensional diagrams using the traditional data visualization tools. However, it is difficult to visualize big data in real time using these tools. To enhance the visualization quality, this project proposes a realtime data mining and visualization system for big data.

FUNCTIONALITIES OFFERED:

1. **Regression :** A set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.
   a. **Linear Regression :** A linear approach to modeling the relationship between a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

2. **Classification :** Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.
   a. **Decision Tree :** The goal of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from prior data (trained data). For predicting a class label for a record we start from the root of the tree, compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.
   b. **SVM :** A Support Vector Machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.
   c. **PCA :** Principal Component Analysis is a feature extraction method that uses orthogonal linear projections to capture the underlying variance of the data. It is a method used to reduce the number of variables in the data by extracting an important one from a large pool. It maps each instance of the given dataset present in a 'd' dimensional space to a 'k' dimensional subspace such that k < d.
   d. **KNN Classifier :** K Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

3. **Clustering :** The task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
    a. **Fuzzy C-Means:** Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster. Clustering or cluster analysis involves assigning data points to clusters such that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Different similarity measures may be chosen based on the data or the application.
    b. **KMeans :** A method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.
    c. **Hierarchical Clustering :** A method of cluster analysis which seeks to build a hierarchy of clusters. The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

TOOLS & TECHNOLOGY USED:

1. **Hardware Requirements:**

A machine or server capable of hosting the web app

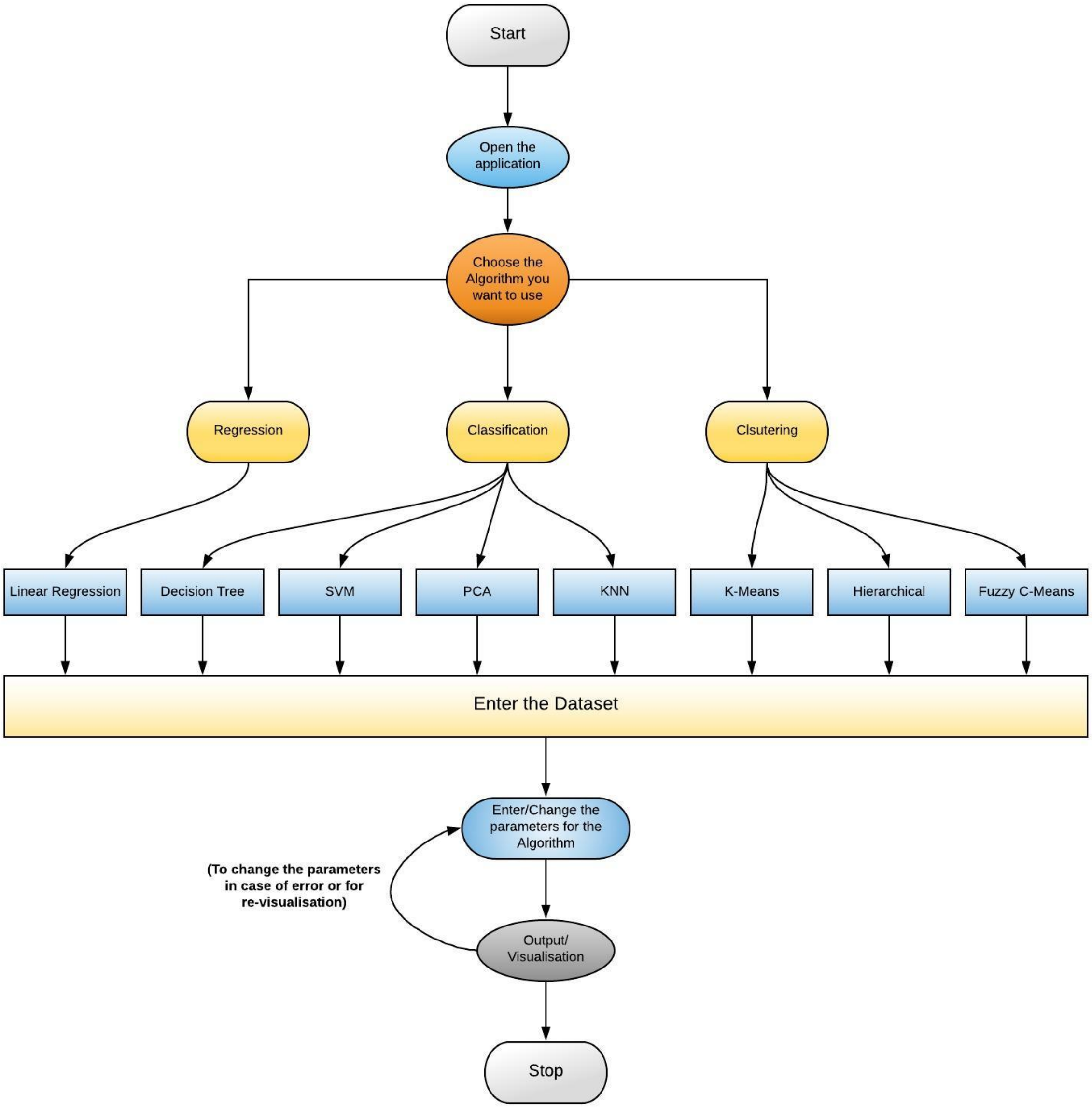
2. **Software Requirements (for Deployment)**

A HTML5 Web Browser capable of running Progressive Web Applications

3. **Development Dependencies:**

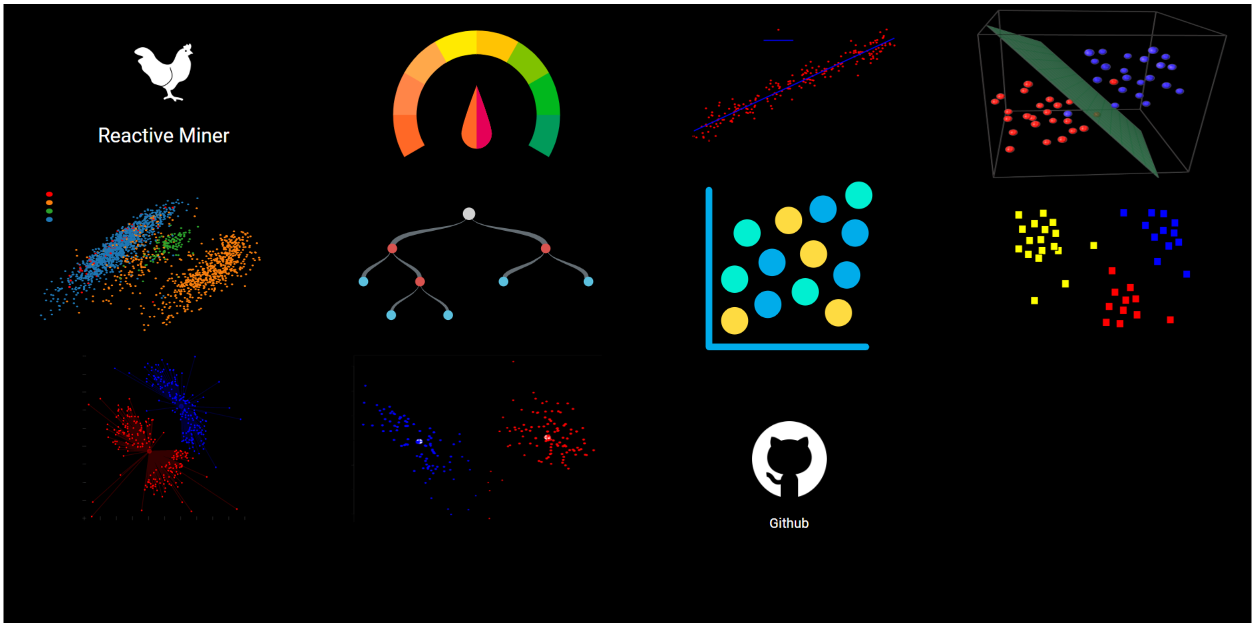
- ML.JS
- LIBSVM
- TENSORFLOW.JS
- MATERIAL-UI
- REACT.JS
- D3.JS

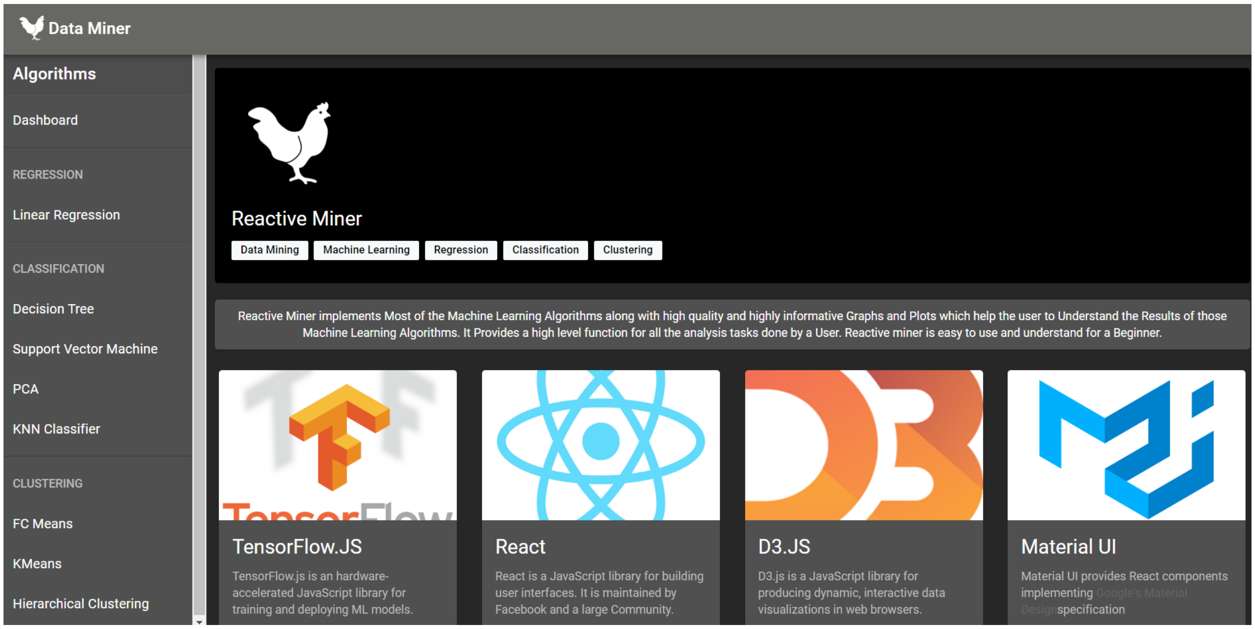
FLOW DIAGRAM:



SCREENSHOTS OF OUTPUT:

Overview Of Dashboard:





Reactive Miner implements Most of the Machine Learning Algorithms along with high quality and highly informative Graphs and Plots which help the user to Understand the Results of those Machine Learning Algorithms. It Provides a high level function for all the analysis tasks done by a User. Reactive miner is easy to use and understand for a Beginner.

TensorFlow.JS

TensorFlow.js is an hardware-accelerated JavaScript library for training and deploying ML models.

React

React is a JavaScript library for building user interfaces. It is maintained by Facebook and a large Community.

D3.JS

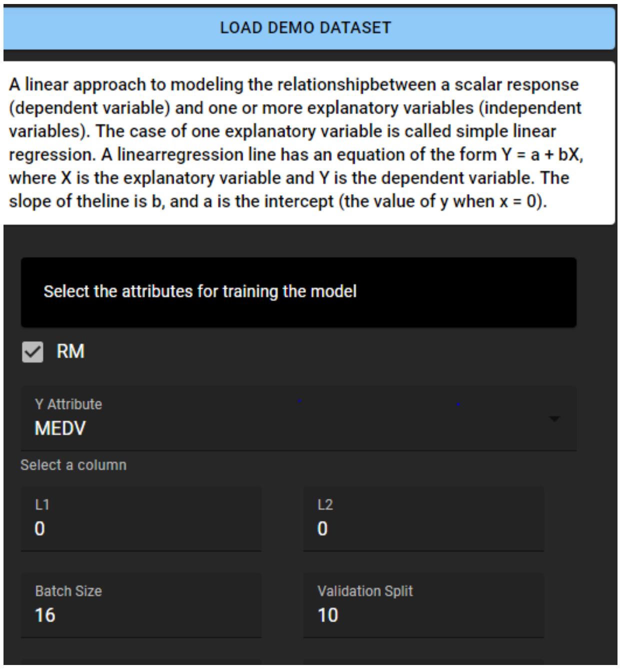
D3.js is a JavaScript library for producing dynamic, interactive data visualizations in web browsers.

Material UI

Material UI provides React components implementing Google's Material Design specification

1. Regression

a. Linear Regression



LOAD DEMO DATASET

A linear approach to modeling the relationshipbetween a scalar response (dependent variable) and one or more explanatory variables (independent variables). The case of one explanatory variable is called simple linear regression. A linearregression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of theline is b, and a is the intercept (the value of y when x = 0).

Select the attributes for training the model

☑ RM

Y Attribute
MEDV

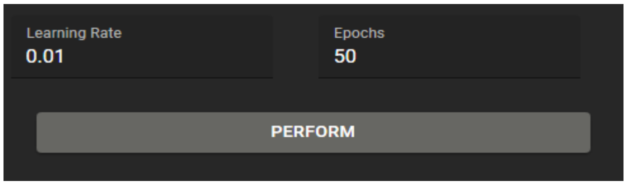
Select a column

L1
0

L2
0

Batch Size
16

Validation Split
10

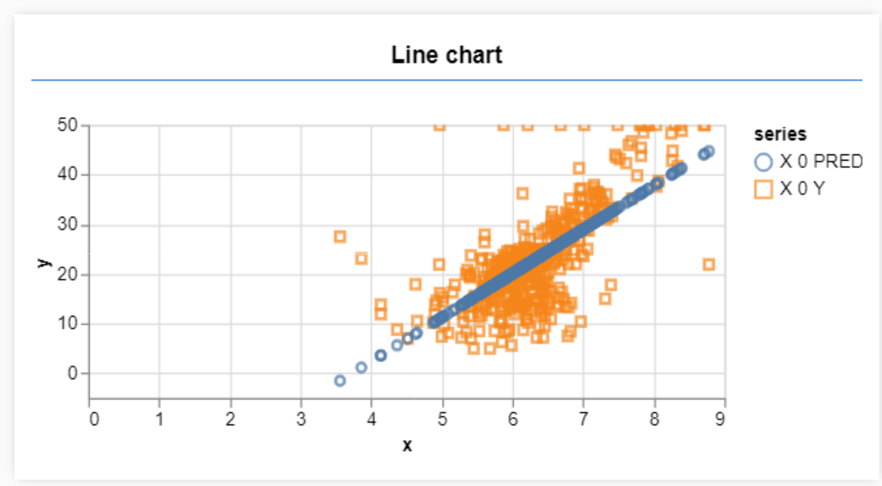


Learning Rate
0.01

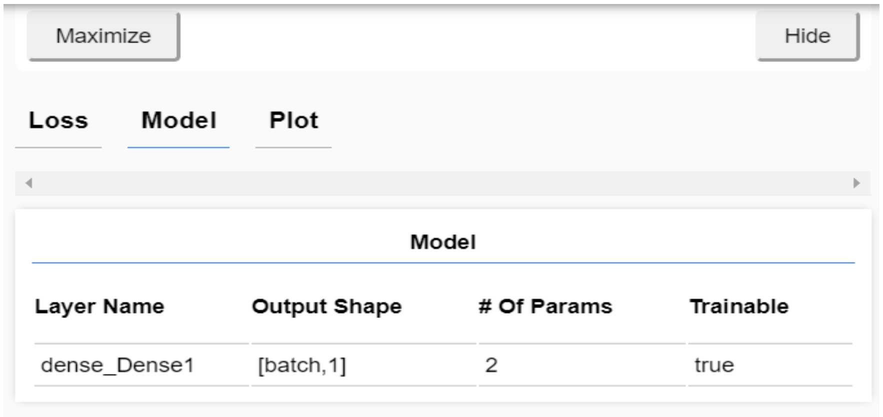
Epochs
50

PERFORM

Plot diagram:



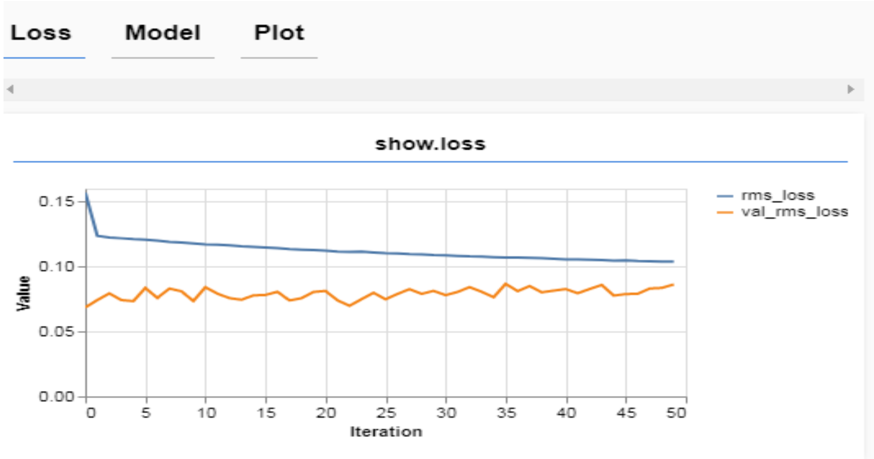
Model:



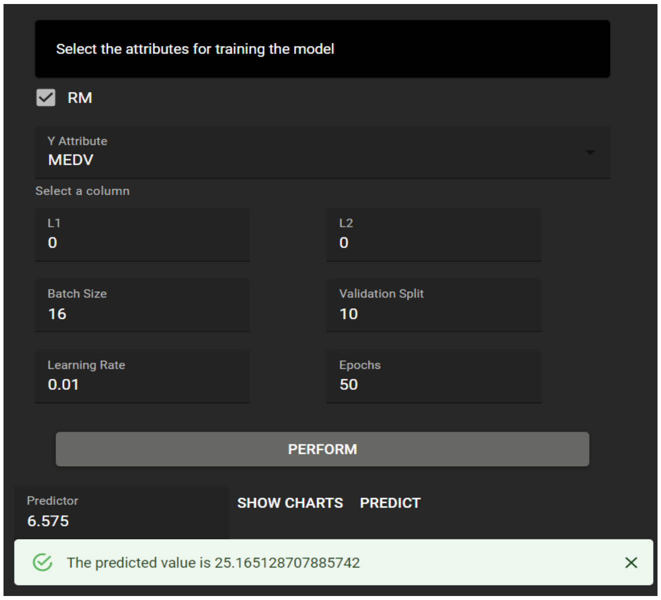
Model

Layer Name	Output Shape	# Of Params	Trainable
dense_Dense1	[batch,1]	2	true

Loss:



Predicted Value:



Select the attributes for training the model

☑ RM

Y Attribute
MEDV

Select a column

L1
0

L2
0

Batch Size
16

Validation Split
10

Learning Rate
0.01

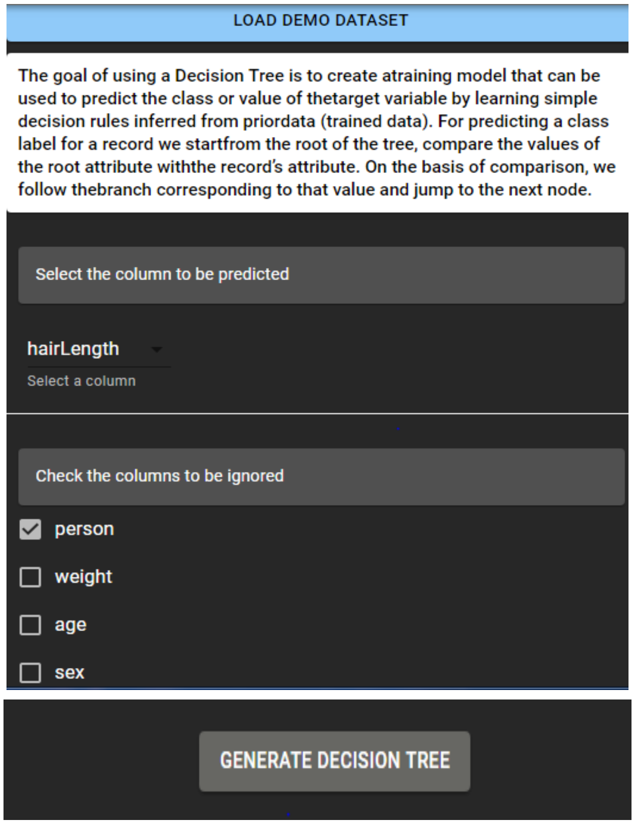
Epochs
50

Predictor
6.575

✓ The predicted value is 25.165128707885742

2. Classification

a. Decision Tree



The goal of using a Decision Tree is to create atraining model that can be used to predict the class or value of thetarget variable by learning simple decision rules inferred from priordata (trained data). For predicting a class label for a record we startfrom the root of the tree, compare the values of the root attribute withthe record's attribute. On the basis of comparison, we follow thebranch corresponding to that value and jump to the next node.

Select the column to be predicted

hairLength

Select a column

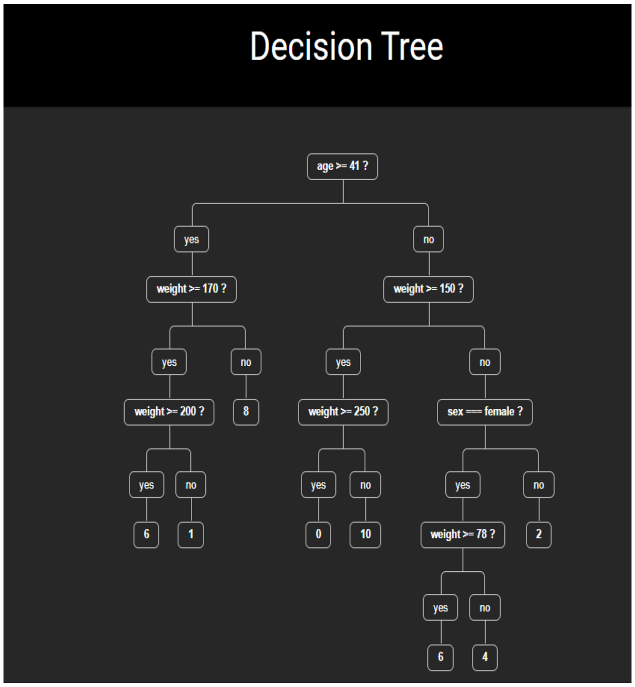
Check the columns to be ignored

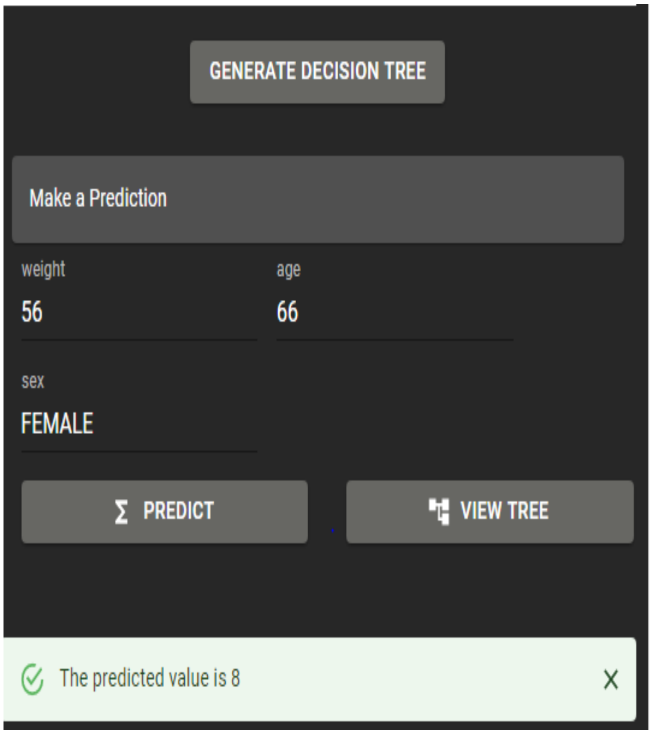
☑ person

□ weight

□ age

□ sex





Make a Prediction

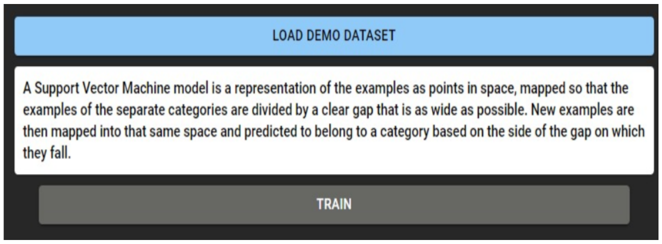
weight: 56

age: 66

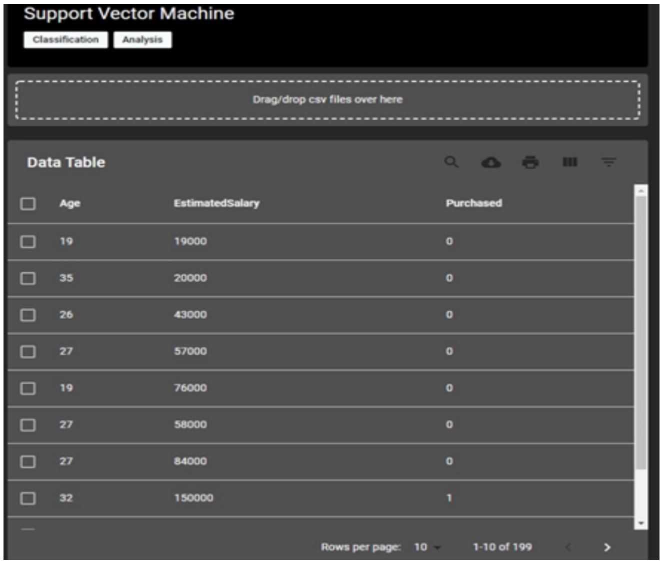
sex: FEMALE

✓ The predicted value is 8

b. SVM



A Support Vector Machine model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall.

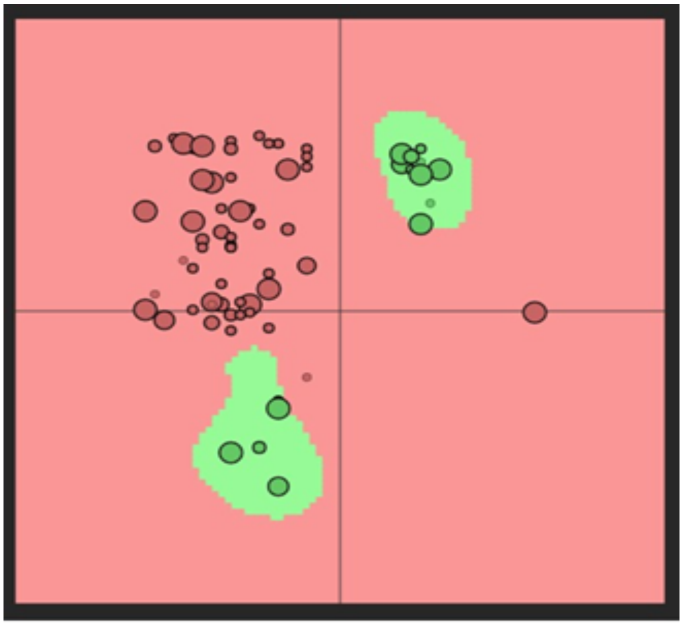


Support Vector Machine

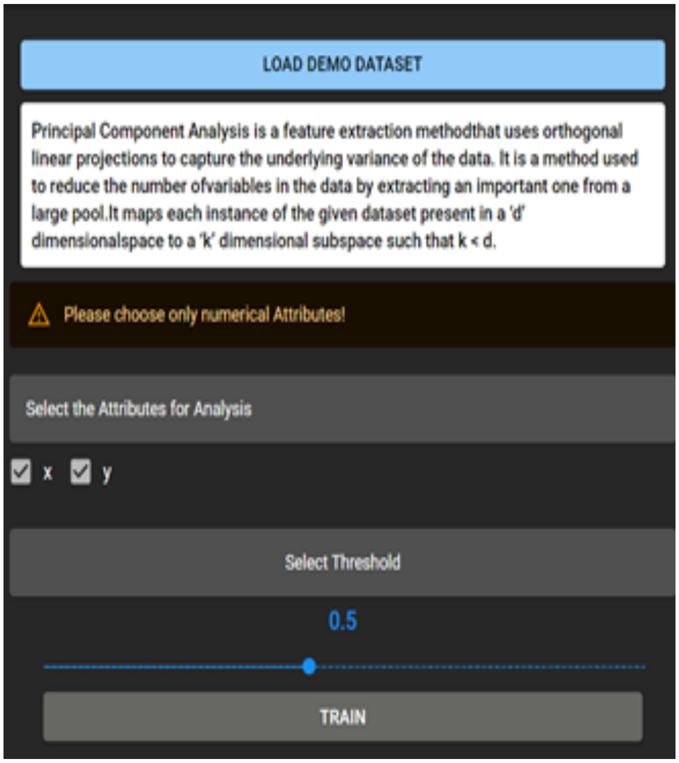
Data Table

□	Age	EstimatedSalary	Purchased
□	19	19000	0
□	35	20000	0
□	26	43000	0
□	27	57000	0
□	19	76000	0
□	27	58000	0
□	27	84000	0
□	32	150000	1

Rows per page: 10 1-10 of 199



c. PCA



Principal Component Analysis is a feature extraction methodthat uses orthogonal linear projections to capture the underlying variance of the data. It is a method used to reduce the number ofvariables in the data by extracting an important one from a large pool.It maps each instance of the given dataset present in a 'd' dimensionalspace to a 'k' dimensional subspace such that k < d.

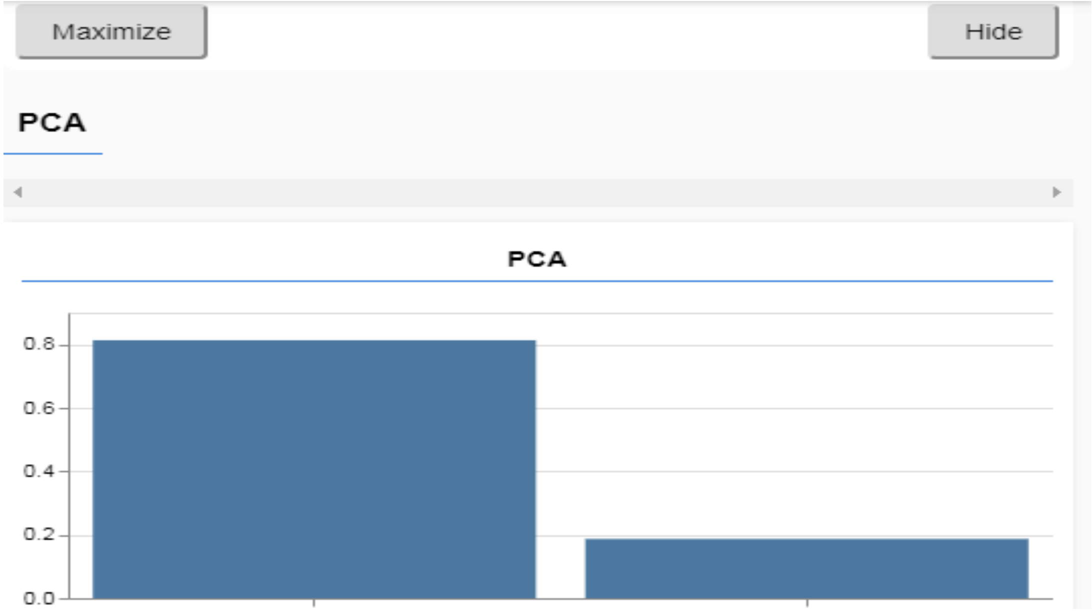
⚠ Please choose only numerical Attributes!

Select the Attributes for Analysis

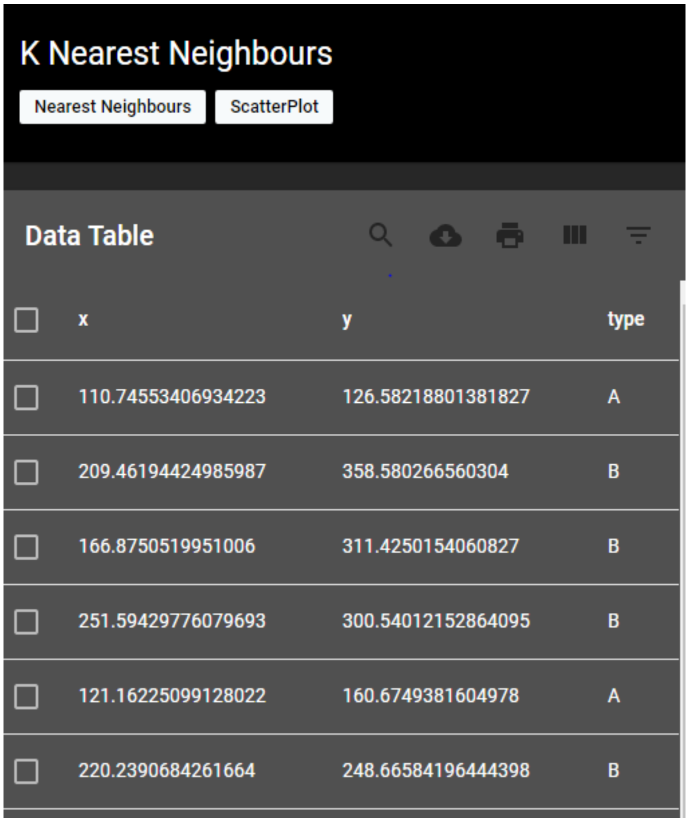
☑ x ☑ y

Select Threshold

0.5

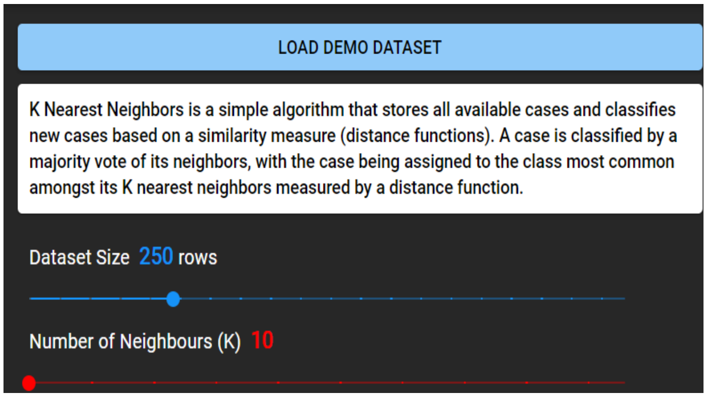


d. KNN



Data Table

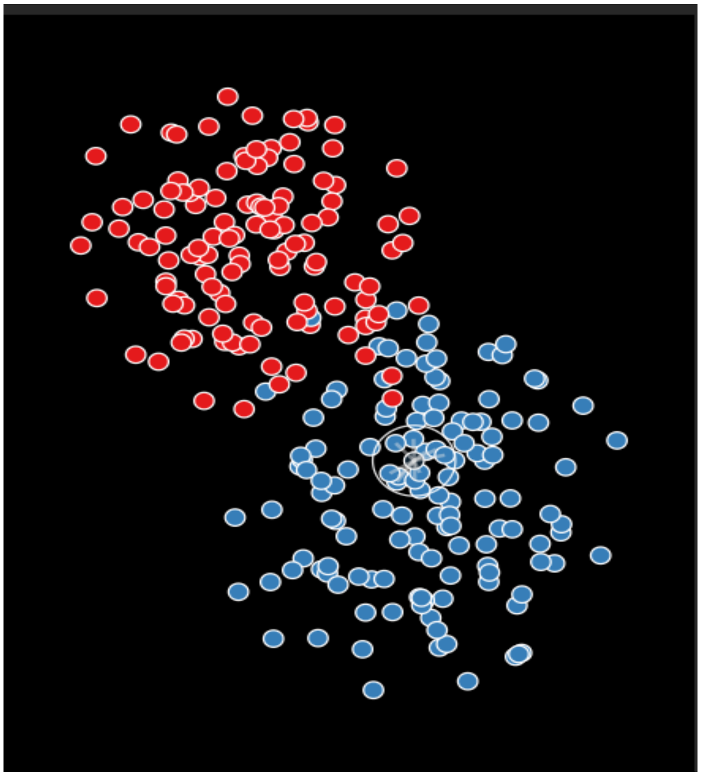
□	x	y	type
□	110.74553406934223	126.58218801381827	A
□	209.46194424985987	358.580266560304	B
□	166.8750519951006	311.4250154060827	B
□	251.59429776079693	300.54012152864095	B
□	121.16225099128022	160.6749381604978	A
□	220.2390684261664	248.66584196444398	B



K Nearest Neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.

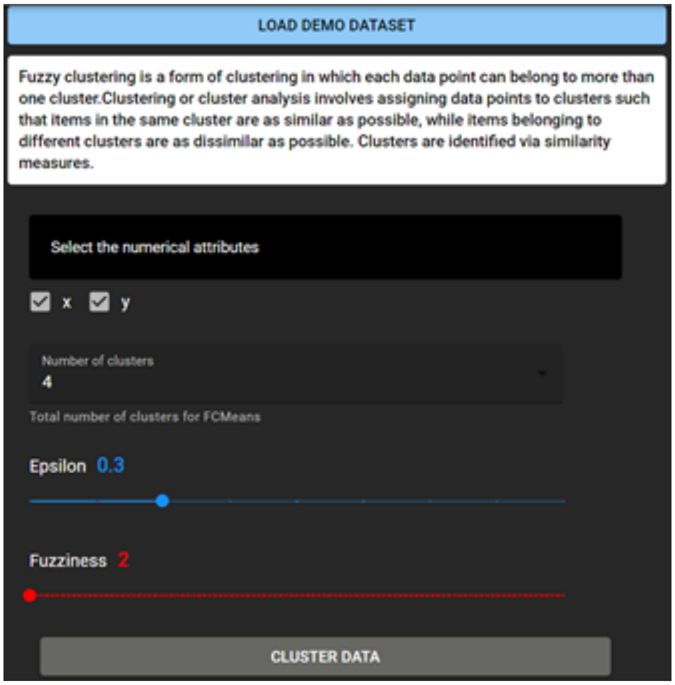
Dataset Size 250 rows

Number of Neighbours (K) 10



3. Clustering:

a. FC Means



Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster.Clustering or cluster analysis involves assigning data points to clusters such that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible. Clusters are identified via similarity measures.

Select the numerical attributes

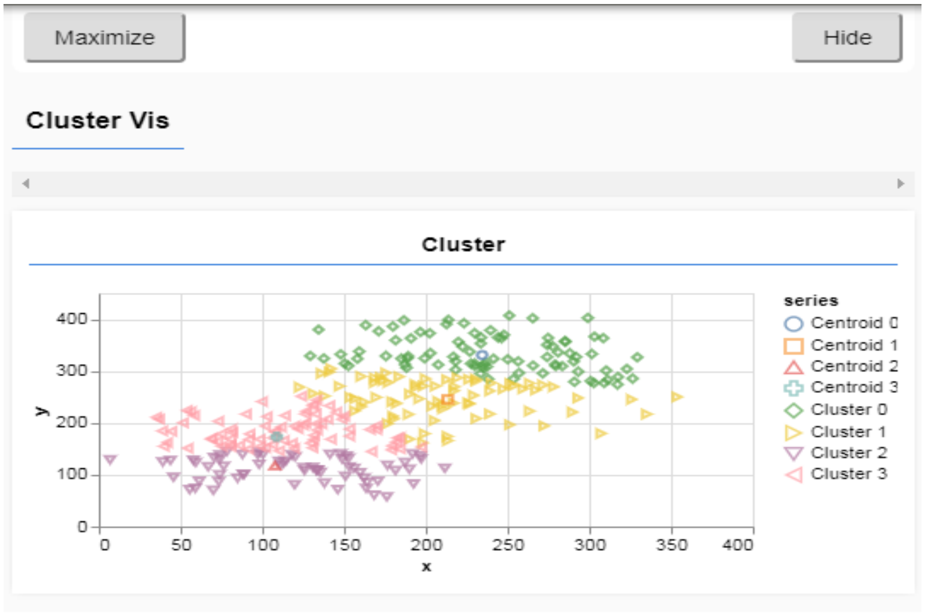
☑ x ☑ y

Number of clusters
4

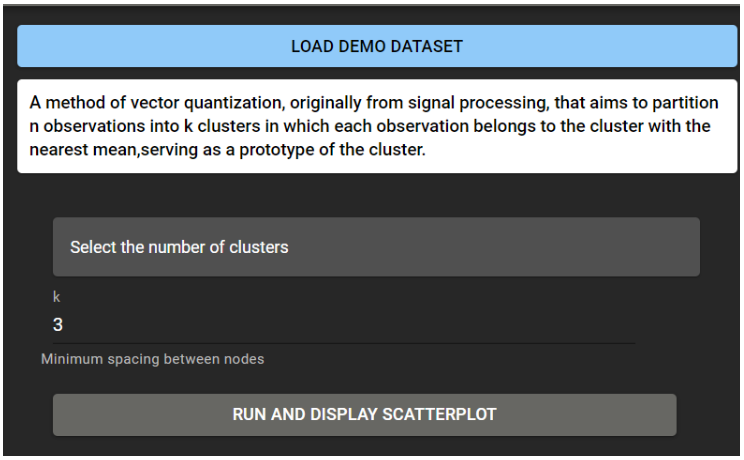
Total number of clusters for FCMeans

Epsilon 0.3

Fuzziness 2



b. K Means



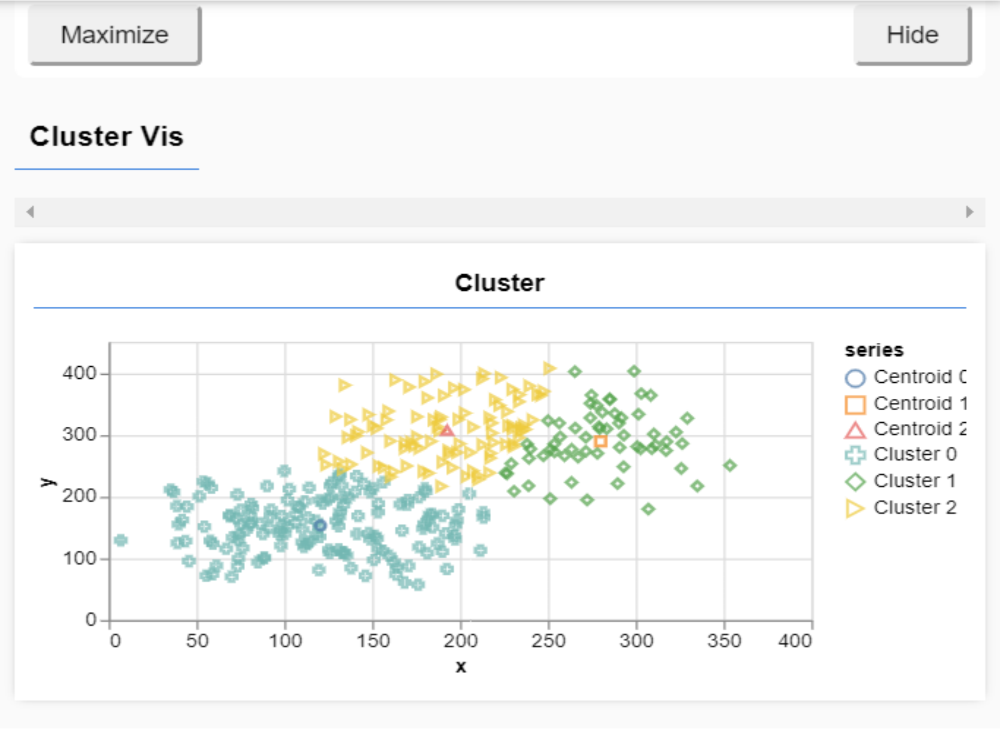
A method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean,serving as a prototype of the cluster.

Select the number of clusters

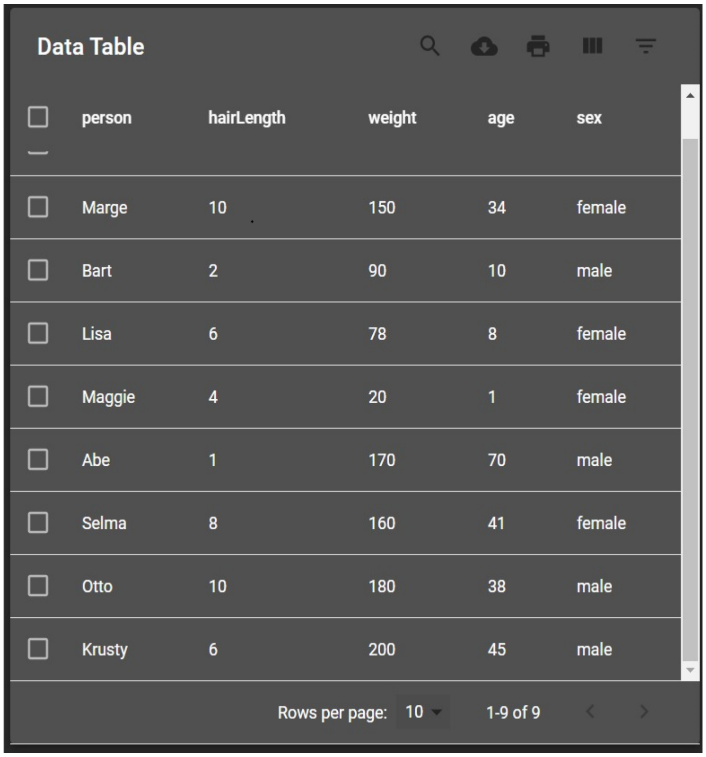
k

3

Minimum spacing between nodes



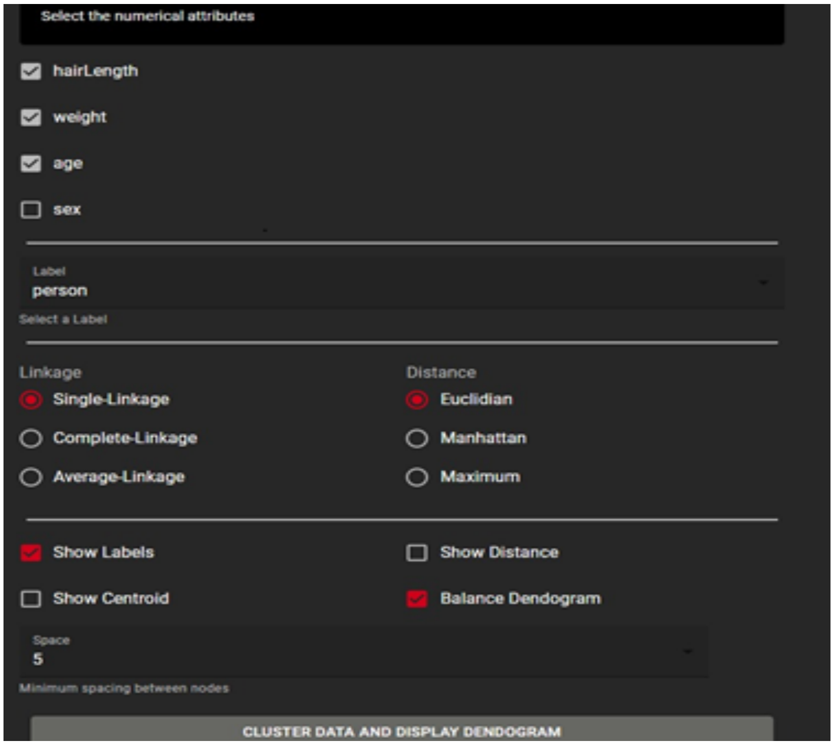
c. Hierarchical Clustering



Data Table

□	person	hairLength	weight	age	sex
□	Marge	10	150	34	female
□	Bart	2	90	10	male
□	Lisa	6	78	8	female
□	Maggie	4	20	1	female
□	Abe	1	170	70	male
□	Selma	8	160	41	female
□	Otto	10	180	38	male
□	Krusty	6	200	45	male

Rows per page: 10 1-9 of 9



Select the numerical attributes

☑ hairLength

☑ weight

☑ age

□ sex

Label
person
Select a Label

Linkage
● Single-Linkage
○ Complete-Linkage
○ Average-Linkage

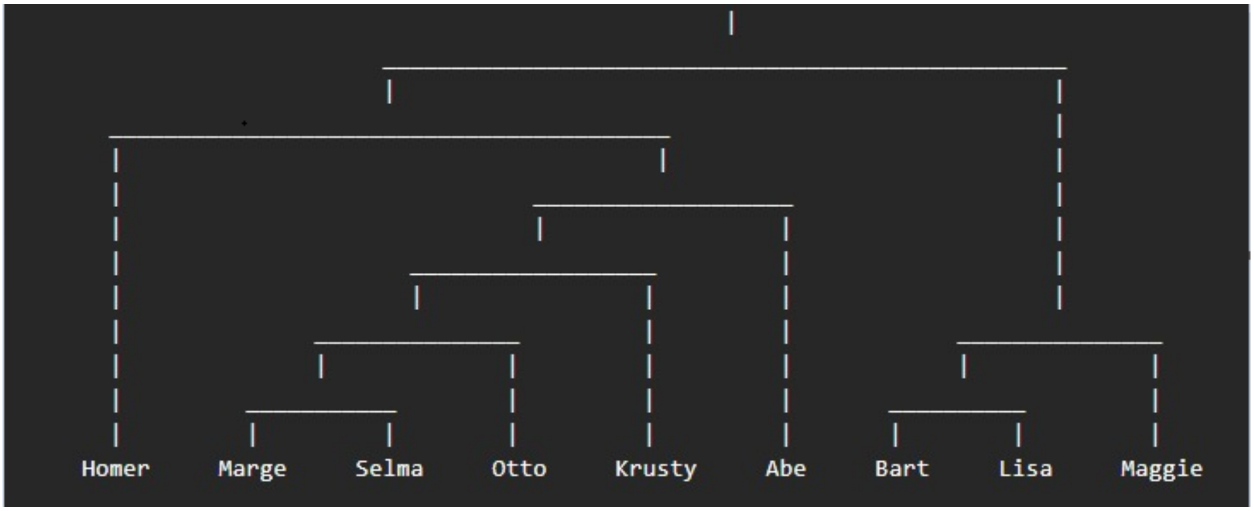
Distance
● Euclidian
○ Manhattan
○ Maximum

☑ Show Labels
□ Show Distance

□ Show Centroid
☑ Balance Dendogram

Space
5
Minimum spacing between nodes

Dendrogram:



```
                                                     |
                            ___________________________________________________
                            |                                                 |
        _________________________________________                             |
        |                                       |                             |
        |                              ____________________                   |
        |                              |                  |                   |
        |                     __________________          |                   |
        |                     |                |          |                   |
        |              _______________         |          |            _______________
        |              |             |         |          |            |             |
        |         ___________        |         |          |       __________         |
        |         |         |        |         |          |       |        |         |
      Homer     Marge     Selma     Otto     Krusty      Abe     Bart     Lisa     Maggie
```

CONCLUSION:

Traditional data visualization tools apply 2D and 3D plots, which were static and difficult to represent big data intuitively. To improve user experience, this project proposed a real-time interactive data mining and visualization system using programming technologies. To provide intuitive operation interfaces, we utilized the GUI methods for interactive data mining and visualization. The iteration of data mining was executed based on the interaction signals to satisfy the user's demands. This study is still ongoing, and we are continuously enhancing and developing common data mining algorithms and data visualization models in the system. Meanwhile, we will apply this system in big data researching areas, such as traffic, banking, and human-centric computing.