A COMPARISON BETWEEN K-MEANS AND TREE ALGORITHM IN R SUPPORT VECTOR OF DATA MINING CLASSIFICATION PROBLEMS

Abdelatti A Blg[1]*, Muner Athaba[1], Sana Abouljam[1], Abdalla Mohamed Alasoud[2]

[1] Computer Science Department, Faculty of Science - Ajaylat, University of Zawia
[2] Computer Department, Faculty of Science, Al-Asmarya Islamic University
** Corresponding author: a.blg@uni.za.ly*

ABSTRACT

Correct analysis of the data to create a logical relationship that summarizes the data in a new way that is understandable and useful, the data will yield huge benefits. The process of analyzing and transforming data into knowledge is called knowledge discovery in the database. In the various steps of knowledge discovery in the database, the richness and precision of the algorithms make it difficult. When analyzing big data, effective user support is essential, and it is even more important now. Metadata is a necessary component to enhance user support [1]. In this paper, we will address problemsof data mining classification, four machine learning validation methods used to build and test 4 distinct datasets with the same amount of training and testing data for each predictive model. Besides, calculated the accuracy average and standard deviation of 20 trials, visualized the accuracy.

Keywords: K-MEANS, Data Mining, Machine Learning, KNIME, Tree algorithm, R.

1- INTRODUCTION

The target of the information extracting method is to extract data from a dataset and makeover it into a clear construction for additional use. this is a diagnostic method planned to scrutinize the information in seek of reliable patterns or organized associations connecting variables, and then to confirm the findings by applying the detected patterns. the focal point of this documents to concern a variety of categorization methods such as J48, kNN, CART, and SVM.

2- MATERIALS AND METHODS

Data mining is a method of exploring, identifying, and modeling large amounts of data that reveal unspecified patterns or relationships that lead to a true result.

There are several data classification algorithms available in DM. We will discuss some of the algorithms used later in this article.

2.1 KNIME Analytics

KNIME (Konstanz Information Miner) It is a modular computing environment that enables easy visual aggregation, interactive data analysis, and data manipulation. It is an open-source predictive analytics platform (published under the GNU General Public License v3) that is suitable for processing a wide variety of data formats, from simple CSV or XLSX files to more complex data structures such as XML, URL, and relational databases such as DB2, Oracle, and MySQL. Surprisingly, widespread use has not been found in the earth sciences.

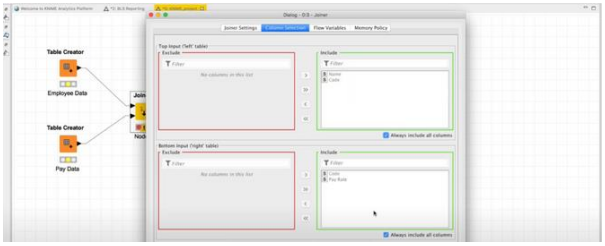


Figure 1: KNIME Analytics platform

KNIME is supported by a wide community of users and developers. Since KNIME is built on top of Eclipse, it shares the benefit of a plugin architecture that makes it easily extensible. Several custom-designed nodes

are available and are easily accessible through the community contributions area.

2.2 DecisionTree algorithm in R

A decision tree is a graph that shapes options and their outcomes in a tree format. A chart contract represents an event or a choice, margins, a graph, decision rules, or decisions. It is used mainly in machine learning and data mining with R.

Figure 2:Decision Tree algorithm in R

Moreover, R is a programming language and programming environment for statistical analysis, graphs, and reporting. R was created by Ross Ihca and Robert Gentleman at the University of Auckland, New Zealand, and is now being developed by the Core team in development.The core of R is an interpretable computer language that allows for branching and redundancy as well as modular programming using functions. R allows integration with procedures written in C, C ++, and. Net, Python, or FORTRAN to increase efficiency.

3- METHODOLOGY

With the increasing availability of information, the increase in data volume requires the use of data extraction techniques to collect useful information from the dataset. Data mining technology is described as paying special attention to classification technology as an important supervised learning

technique [3]. A Decision Tree algorithm has been used to train and test the datasets for this problem. Classification and prediction are techniques used to create categories of important data and predict potential trends. A decision tree is an important classification method for classifying data extraction. It is commonly used for marketing, surveillance, fraud detection, and scientific discovery [4]. This research will study two problems of data mining which hierarchical clustering and classification methods of data.

3.1 Problem 1: Hierarchical clustering:(single, average, complete, centroid and Ward linkages)

Hierarchical techniques often find clusters nested into groups or subdivisions. Clustering is where each data point begins with its group, and then a pair of similar groups are successively combined to form a hierarchical block. Instead, the group begins by dividing all the data points into one group, then repeatedly dividing each group into smaller groups. After splitting or merging, it will be irreversible, so the hierarchical group cannot be modified [5].

An agglomerative hierarchical clustering algorithm, the following steps are generally implemented [2]:

Step 1: Each observation is considered to be an initial cluster.

Step 2: Distances between clusters are computed.

Step 3: Two clusters that have minimum distance are combined and replaced by a single cluster. Step 4: Repeat Steps 2 and 3 until there is only a single cluster containing all observations.

The need to measure distance or proximity to determine the similarity between objects. The most common is the Euclidean distance. The output of

the hierarchical clustering algorithm is a tree diagram, which is a two-dimensional tree structure describing the arrangement of the nested groups[6].

-Single: Single linkage computes the smallest dissimilarity between two objects.

- Complete: Complete linkage, the opposite of Single linkage, computes the largest dissimilarity between two objects.

- Average: Average linkage is the intermediate between the maximum and minimum distance methods.

- Centroid: Centroid linkage is defined as the distance between centers of gravity (centroids) of two clusters.

-Ward: Ward linkage (or Ward minimum variance method) [2].

3.1.1 Task 1 KNIME Solution

The corresponding KNIME workflow and the R code have been used to calculate some statistics about the dataset "teeth.csv".

KNIME Solution:

The number of records and number of attributes has been calculated from data from this dataset as shown in figure 1.

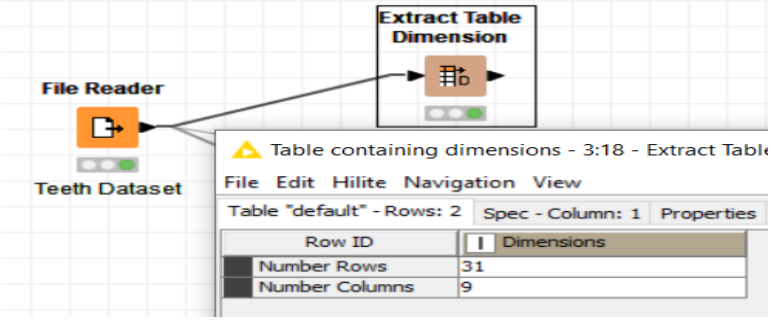


Figure3: Number of rows (records) and columns (attributes) as output table.

Calculation of the range and mean value of each attribute has been done. Moreover plotted the histogram for each attribute of a dataset.

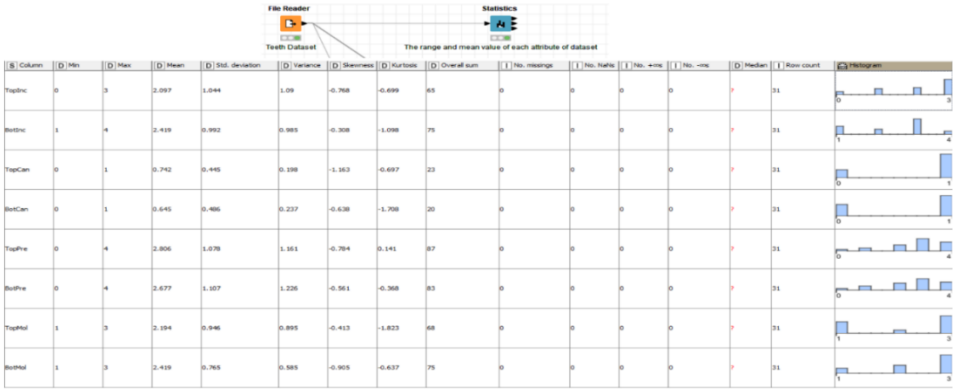


Figure 4: The range, mean value, and histogram of eachattribute

Command of R:

The calculation of records and the number of attributes can be done in different ways as shown in the following steps.

- Dimension(dim): number of rows/observations and columns/variables

  dim (teeth) 31 9

- Number of records (nrow) and number of attributes (ncol)

  nrow(teeth) 31 ncol(teeth) 9

The range and mean value of each attributes were calculated as shown in Table 1.

summary_csv = summary(teeth)

write.csv(summary_csv, file = 'teeth_Summary.csv')

TABLE 1: THE RANGE AND MEAN VALUE OF EACH ATTRIBUTE

Animal	TopInc	BotInc	TopCan	BotCan	TopPre	BotPre	TopMol	BotMol
Length:31	Min. :0.000	Min. :1.000	Min. :0.0000	Min. :0.0000	Min. :0.000	Min. :0.000	Min. :1.000	Min. :1.000
Class :character	1st Qu.:1.000	1st Qu.:1.500	1st Qu.:0.5000	1st Qu.:0.0000	1st Qu.:2.000	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:2.000
Mode :character	Median :2.000	Median :3.000	Median :1.0000	Median :1.0000	Median :3.000	Median :3.000	Median :3.000	Median :3.000
	Mean :2.097	Mean :2.419	Mean :0.7419	Mean :0.6452	Mean :2.806	Mean :2.677	Mean :2.194	Mean :2.419
	3rd Qu.:3.000	3rd Qu.:3.000	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000	3rd Qu.:3.500	3rd Qu.:3.000	3rd Qu.:3.000
	Max. :3.000	Max. :4.000	Max. :1.0000	Max. :1.0000	Max. :4.000	Max. :4.000	Max. :3.000	Max. :3.000

The following figure shows a histogram of each feature of the datasetusing the hist() function.

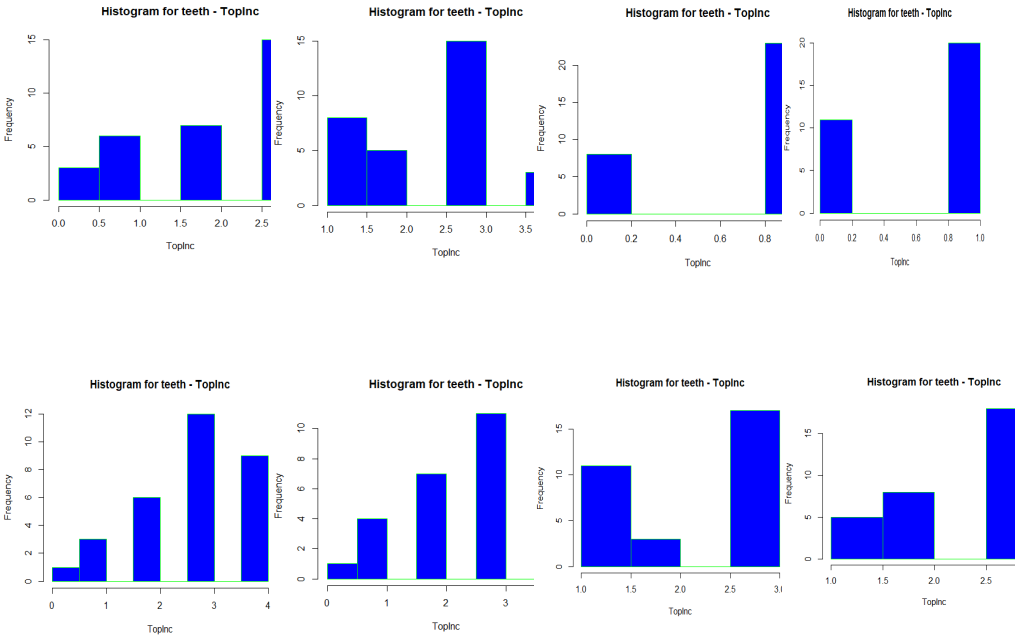


Figure 5: Histogram of each feature of the dataset

3.1.2 Task2 Clustering the teeth dataset

The generation of the dendrogram with KNIME.

Clustering the teeth dataset using hierarchical clustering.

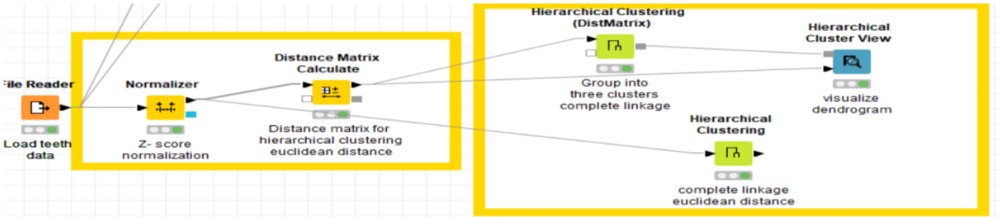


Figure 6: Clustering the teeth dataset using hierarchical clustering

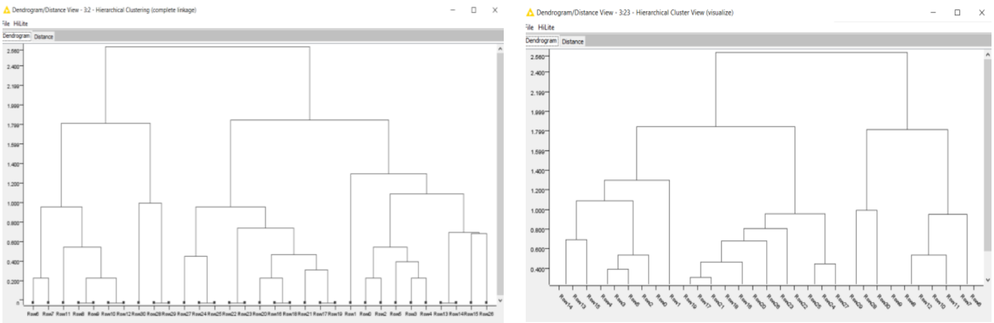


Figure7: Complete linkage Euclidean distance

3.1.3 Task dendrogram with R

The generation of the dendrogram with R.

We applied hierarchical clustering by normalizing all records and calculating the Euclidean distance between all records in the data set as shown in table2.

TABLE 2: ILLUSTRATE THE EUCLIDEAN DISTANCE MATRIX BETWEEN (DISSIMILARITY MATRIX) ALL THE ROWS (OBJECTS).

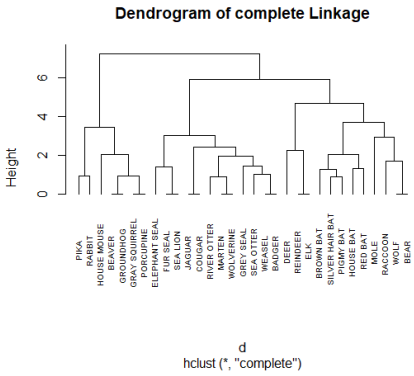
	BROWN BAT	MOLE	SILVER HA	PIGMY BA	HOUSE BA	RED BAT	PIKA	RABBIT	BEAVER	GROUNDH	GRAY SQU	HOUSE M	PORCUPIN	WOLF	BEAR
BROWN BAT	0	2.481878	0.927961	1.294952	2.064033	1.610609	3.87552	3.762784	4.287678	4.287678	4.57903	5.417751	4.57903	1.926616	1.926616
MOLE	2.481877582	0	2.649684	2.799396	3.227995	3.253935	2.943415	2.79331	3.723243	3.723243	4.055351	4.983009	4.055351	2.835004	2.835004
SILVER HAIR BAT	0.927960727	2.649684	0	0.903211	1.294952	1.316415	3.762784	3.87552	4.186057	4.186057	4.287678	5.004644	4.287678	2.50902	2.50902
PIGMY BAT	1.294951962	2.799396	0.903211	0	0.927961	0.957685	3.652774	3.768802	3.882745	3.882745	3.992094	4.57903	3.992094	2.95678	2.95678
HOUSE BAT	2.06403341	3.227995	1.294952	0.927961	0	1.333518	3.768802	4.097219	3.992094	3.992094	3.882745	4.287678	3.882745	3.612216	3.612216
RED BAT	1.610608689	3.253935	1.316415	0.957685	1.333518	0	3.77623	3.888576	3.762784	3.762784	3.87552	4.477763	3.87552	3.390285	3.390285
PIKA	3.87552016	2.943415	3.762784	3.652774	3.768802	3.77623	0	0.927961	1.316415	1.316415	1.610609	2.761297	1.610609	4.699501	4.699501
RABBIT	3.762784261	2.79331	3.87552	3.768802	4.097219	3.888576	0.927961	0	1.610609	1.610609	2.275389	3.454029	2.275389	4.416104	4.416104
BEAVER	4.28767824	3.723243	4.186057	3.882745	3.992094	3.762784	1.316415	1.610609	0	0	0.927961	2.064033	0.927961	5.377335	5.377335
GROUNDHOG	4.28767824	3.723243	4.186057	3.882745	3.992094	3.762784	1.316415	1.610609	0	0	0.927961	2.064033	0.927961	5.377335	5.377335
GRAY SQUIRREL	4.57903025	4.055351	4.287678	3.992094	3.882745	3.87552	1.610609	2.275389	0.927961	0.927961	0	1.294952	0	5.763791	5.763791
HOUSE MOUSE	5.417750543	4.983009	5.004644	4.57903	4.287678	4.477763	2.761297	3.454029	2.064033	2.064033	1.294952	0	1.294952	6.705191	6.705191
PORCUPINE	4.57903025	4.055351	4.287678	3.992094	3.882745	3.87552	1.610609	2.275389	0.927961	0.927961	0	1.294952	0	5.763791	5.763791
WOLF	1.926615896	2.835004	2.50902	2.95678	3.612216	3.390285	4.699501	4.416104	5.377335	5.377335	5.763791	6.705191	5.763791	0	0
BEAR	1.926615896	2.835004	2.50902	2.95678	3.612216	3.390285	4.699501	4.416104	5.377335	5.377335	5.763791	6.705191	5.763791	0	0
RACCOON	2.074516612	2.937519	2.6243	3.055212	3.69322	3.476464	4.762045	4.482605	5.432081	5.432081	5.8149	6.749175	5.8149	1.681472	1.681472
MARTEN	2.962224336	3.619416	3.370179	3.715572	4.255706	4.069024	5.210396	4.956299	5.829121	5.829121	6.187424	7.07266	6.187424	1.681472	1.681472
WEASEL	2.664183258	3.379833	2.821167	2.962224	3.370179	3.395033	4.702928	4.610468	5.226507	5.226507	5.468063	6.187424	5.468063	2.122321	2.122321
WOLVERINE	2.962224336	3.619416	3.370179	3.715572	4.255706	4.069024	5.210396	4.956299	5.829121	5.829121	6.187424	7.07266	6.187424	1.681472	1.681472
BADGER	2.664183258	3.379833	2.821167	2.962224	3.370179	3.395033	4.702928	4.610468	5.226507	5.226507	5.468063	6.187424	5.468063	2.122321	2.122321
RIVER OTTER	2.821167054	3.504908	3.246893	3.370179	3.957734	3.756273	4.969996	4.702928	5.468063	5.468063	5.848527	6.656725	5.848527	1.9087	1.9087
SEA OTTER	2.848360278	3.226141	2.995708	3.128907	3.517583	3.541403	4.367113	4.267384	4.926512	4.926512	5.182071	5.936191	5.182071	2.349368	2.349368
JAGUAR	3.611417779	4.167461	3.728733	3.617687	3.958661	3.979842	5.141052	5.05661	5.477014	5.477014	5.707978	6.271678	5.707978	3.475672	3.475672
COUGAR	3.611417779	4.167461	3.728733	3.617687	3.958661	3.979842	5.141052	5.05661	5.477014	5.477014	5.707978	6.271678	5.707978	3.475672	3.475672
FUR SEAL	3.862477616	4.148923	4.183547	4.466479	4.924936	4.764548	5.406575	5.162143	6.005121	6.005121	6.353506	7.218403	6.353506	2.995214	2.995214
SEA LION	3.862477616	4.148923	4.183547	4.466479	4.924936	4.764548	5.406575	5.162143	6.005121	6.005121	6.353506	7.218403	6.353506	2.995214	2.995214
GREY SEAL	2.1816945	2.656053	2.370844	2.537064	3.003336	3.0312	3.964633	3.854504	4.573528	4.573528	4.847731	5.64668	4.847731	2.098033	2.098033
ELEPHANT SEAL	4.128852758	4.375615	4.430661	4.698737	5.136505	4.79534	5.224807	4.971446	5.682844	5.682844	6.049816	6.952595	6.049816	3.596431	3.596431
REINDEER	2.985162579	3.509355	3.126069	3.253935	3.629246	2.799396	4.420118	4.321612	4.385582	4.385582	4.670831	5.495559	4.670831	4.036296	4.036296
ELK	2.985162579	3.509355	3.126069	3.253935	3.629246	2.799396	4.420118	4.321612	4.385582	4.385582	4.670831	5.495559	4.670831	4.036296	4.036296
DEER	3.7370501	4.167724	3.85054	3.955053	4.269166	3.590399	3.805666	3.690797	3.765499	3.765499	4.094181	5.01466	4.094181	4.620177	4.620177

	RACCOON	MARTEN	WEASEL	WOLVERIN	BADGER	RIVER OTT	SEA OTTER	JAGUAR	COUGAR	FUR SEAL	SEA LION	GREY SEAL	ELEPHANT	REINDEER	ELK	DEER
BROWN BAT	2.074517	2.962224	2.664183	2.962224	2.664183	2.821167	2.84836	3.611418	3.611418	3.862478	3.862478	2.181694	4.128853	2.985163	2.985163	3.73705
MOLE	2.937519	3.619416	3.379833	3.619416	3.379833	3.504908	3.226141	4.167461	4.167461	4.148923	4.148923	2.656053	4.375615	3.509355	3.509355	4.167724
SILVER HAIR BAT	2.6243	3.370179	2.821167	3.370179	2.821167	3.246893	2.995708	3.728733	3.728733	4.183547	4.183547	2.370844	4.430661	3.126069	3.126069	3.85054
PIGMY BAT	3.055212	3.715572	2.962224	3.715572	2.962224	3.370179	3.128907	3.617687	3.617687	4.466479	4.466479	2.537064	4.698737	3.253935	3.253935	3.955053
HOUSE BAT	3.69322	4.255706	3.370179	4.255706	3.370179	3.957734	3.517583	3.958661	3.958661	4.924936	4.924936	3.003336	5.136505	3.629246	3.629246	4.269166
RED BAT	3.476464	4.069024	3.395033	4.069024	3.395033	3.756273	3.541403	3.979842	3.979842	4.764548	4.764548	3.0312	4.79534	2.799396	2.799396	3.590399
PIKA	4.762045	5.210396	4.702928	5.210396	4.702928	4.969996	4.367113	5.141052	5.141052	5.406575	5.406575	3.964633	5.224807	4.420118	4.420118	3.805666
RABBIT	4.482605	4.956299	4.610468	4.956299	4.610468	4.702928	4.267384	5.05661	5.05661	5.162143	5.162143	3.854504	4.971446	4.321612	4.321612	3.690797
BEAVER	5.432081	5.829121	5.226507	5.829121	5.226507	5.468063	4.926512	5.477014	5.477014	6.005121	6.005121	4.573528	5.682844	4.385582	4.385582	3.765499
GROUNDHOG	5.432081	5.829121	5.226507	5.829121	5.226507	5.468063	4.926512	5.477014	5.477014	6.005121	6.005121	4.573528	5.682844	4.385582	4.385582	3.765499
GRAY SQUIRREL	5.8149	6.187424	5.468063	6.187424	5.468063	5.848527	5.182071	5.707978	5.707978	6.353506	6.353506	4.847731	6.049816	4.670831	4.670831	4.094181
HOUSE MOUSE	6.749175	7.07266	6.187424	7.07266	6.187424	6.656725	5.936191	6.271678	6.271678	7.218403	7.218403	5.64668	6.952595	5.495559	5.495559	5.01466
PORCUPINE	5.8149	6.187424	5.468063	6.187424	5.468063	5.848527	5.182071	5.707978	5.707978	6.353506	6.353506	4.847731	6.049816	4.670831	4.670831	4.094181
WOLF	1.681472	1.681472	2.122321	1.681472	2.122321	1.9087	2.349368	3.475672	3.475672	2.995214	2.995214	2.098033	3.596431	4.036296	4.036296	4.620177
BEAR	1.681472	1.681472	2.122321	1.681472	2.122321	1.9087	2.349368	3.475672	3.475672	2.995214	2.995214	2.098033	3.596431	4.036296	4.036296	4.620177
RACCOON	0	2.11451	2.479527	2.11451	2.479527	2.299335	2.676441	3.210137	3.210137	2.682535	2.682535	1.951915	3.34051	4.108948	4.108948	4.683781
MARTEN	2.11451	0	1.294952	0	1.294952	0.903211	1.640788	2.415332	2.415332	1.65071	1.65071	1.951915	2.586089	4.621105	4.621105	5.138964
WEASEL	2.479527	1.294952	0	1.294952	0	0.927961	1.007613	1.589134	1.589134	2.098033	2.098033	1.460504	2.892189	4.435956	4.435956	4.973133
WOLVERINE	2.11451	0	1.294952	0	1.294952	0.903211	1.640788	2.415332	2.415332	1.65071	1.65071	1.951915	2.586089	4.621105	4.621105	5.138964
BADGER	2.479527	1.294952	0	1.294952	0	0.927961	1.007613	1.589134	1.589134	2.098033	2.098033	1.460504	2.892189	4.435956	4.435956	4.973133
RIVER OTTER	2.299335	0.903211	0.927961	0.903211	0.927961	0	1.369816	1.840234	1.840234	1.881657	1.881657	1.730371	2.739278	4.531977	4.531977	5.058969
SEA OTTER	2.676441	1.640788	1.007613	1.640788	1.007613	1.369816	0	1.881657	1.881657	1.840234	1.840234	1.057255	2.306275	4.766923	4.766923	5.270475
JAGUAR	3.210137	2.415332	1.589134	2.415332	1.589134	1.840234	1.881657	0	0	2.267058	2.267058	2.158337	3.017046	5.061835	5.061835	5.538639
COUGAR	3.210137	2.415332	1.589134	2.415332	1.589134	1.840234	1.881657	0	0	2.267058	2.267058	2.158337	3.017046	5.061835	5.061835	5.538639
FUR SEAL	2.682535	1.65071	2.098033	1.65071	2.098033	1.881657	1.840234	2.267058	2.267058	0	0	2.122321	1.390124	5.434072	5.434072	5.880772
SEA LION	2.682535	1.65071	2.098033	1.65071	2.098033	1.881657	1.840234	2.267058	2.267058	0	0	2.122321	1.390124	5.434072	5.434072	5.880772
GREY SEAL	1.951915	1.951915	1.460504	1.951915	1.460504	1.730371	1.057255	2.158337	2.158337	2.122321	2.122321	0	2.537064	4.401158	4.401158	4.942119
ELEPHANT SEAL	3.34051	2.586089	2.892189	2.586089	2.892189	2.739278	2.306275	3.017046	3.017046	1.390124	1.390124	2.537064	0	5.479029	5.479029	5.922339
REINDEER	4.108948	4.621105	4.435956	4.621105	4.435956	4.531977	4.766923	5.061835	5.061835	5.434072	5.434072	4.401158	5.479029	0	0	2.248188
ELK	4.108948	4.621105	4.435956	4.621105	4.435956	4.531977	4.766923	5.061835	5.061835	5.434072	5.434072	4.401158	5.479029	0	0	2.248188
DEER	4.683781	5.138964	4.973133	5.138964	4.973133	5.058969	5.270475	5.538639	5.538639	5.880772	5.880772	4.942119	5.922339	2.248188	2.248188	0

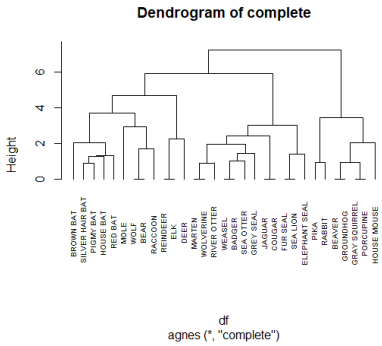
Hierarchical groups were applied to the dataset, and in Table 3, the methods used with generating a dendrogram by hierarchical clustering.

The data has been divided into three subgroups (k=3).

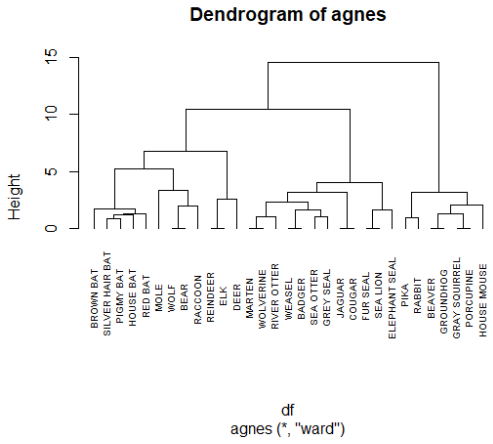
```
#Hierarchical clustering using complete
                Linkage.

hc1=hclust(d, method = "complete")

#plot the obtained dendrogram

plot(hc1, cex=0.6, hang=-1,
main="Dendrogram of complete
                Linkage")
```




```
#compute with agnes

hc2=agnes(df,method="complete")

#Agglomerative coefficient- strength of
clustering

pltree(hc2,cex=0.6,          hang=-1,
main="Dendrogram of complete")

hc2$ac

  0.9014727
```



```
hc3=agnes(df, method="ward")
pltree(hc3,cex=0.6, hang=-1, main="Dendrogram of agnes")
```



Divisive HC

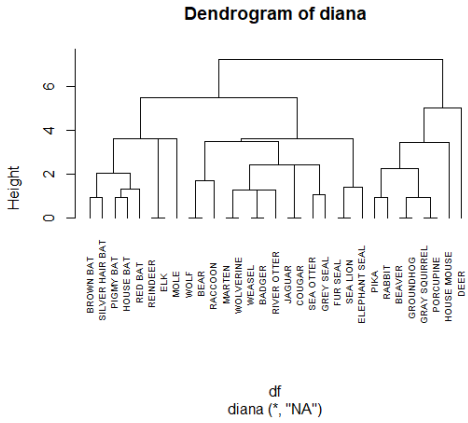
compute divisive hierarchical clustering

```
hc4=diana(df)
```

divise coefficient; amount of clustering structure found

```
pltree(hc4, cex=0.6, hang=-1, main="Dendrogram of diana")
hc4$dc
```

0.8861321



identify sub-groups

ward's method

```
hc5<- hclust(d, method="ward.D2")
```

cut tree into 3 groups

```
sub_grp<- cutree(hc5,k=3)# work with 3 clusters
```

Number of members in each cluster

Table(sub_grp) The number of members of each of the groups.

Group number	1	2	3
Number of members	12	7	12

library (tidyverse)

teeth%>%mutate(cluster=sub_grp)%>%head

plot(hc5, cex=0.6)

rect.hclust(hc5,k=5, border=2:5)

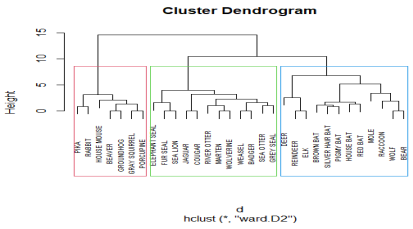
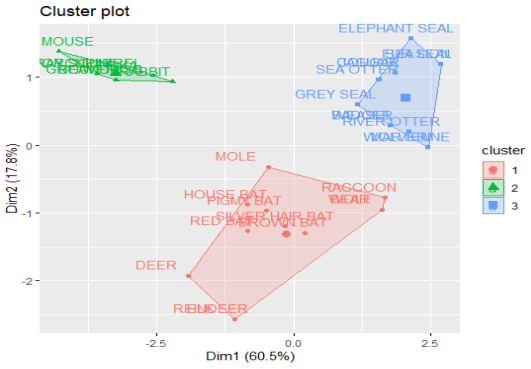


TABLE 3: DEMONSTRATE THE EVALUATION AND DETERMINE THE OPTIMAL CLUSTER NUMBER.

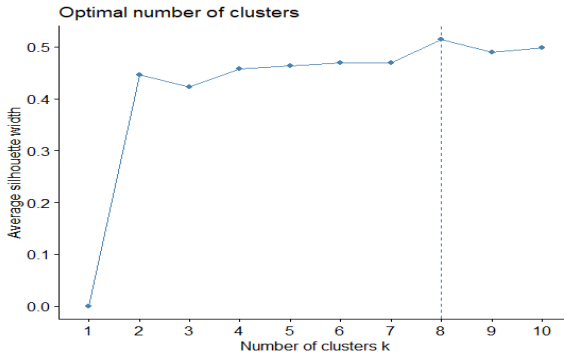
Plotted following plots for each cluster to evaluate.

fviz_cluster(list(data=df,cluster=sub_grp))



Determine the optimal cluster number(silhouette)

fviz_nbclust(df, FUN=hcut, method="silhouette")



Determine the optimal cluster number(wss)

fviz_nbclust(df, FUN=hcut, method="wss")

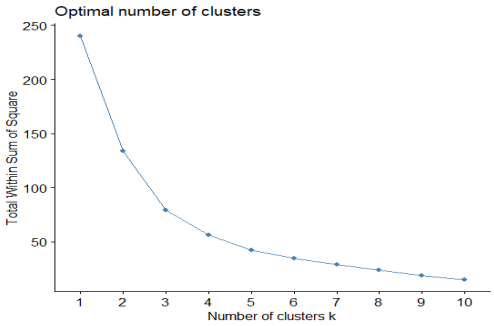


TABLE 4: EVALUATION AND DETERMINE THE OPTIMAL CLUSTER NUMBER

3.2 Problem #2 - Classification

Data mining is an interdisciplinary field, representing the integration of a range of disciplines, including database systems, statistics, machine learning, visualization, and information science.

Additionally, depending on the data extraction method used, techniques from other disciplines can be applied, such as neural networks, fuzzy and/or approximate set theory, knowledge representation, inductive logic programming, or high-performance computing.

3.2.1Task 1 KNIME Solution

The iris dataset is used to illustrate solutions adopted by the four assessment methods. Use decision tree nodes and loops to train and test data sets for more than 20 paths for each method to calculate mean accuracy, standard deviation, mean, and standard deviation of calculation accuracy.

Resub: Resubstitution error method, Use all data for training and testing for this method.

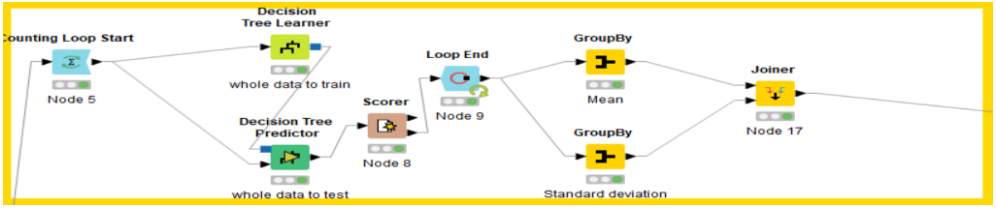


Figure 8: Resubstitution error method

hold-out-10%: hold-out method with 10% - 90% In this method, the partition split into 90% of the data for training and 10% for testing.

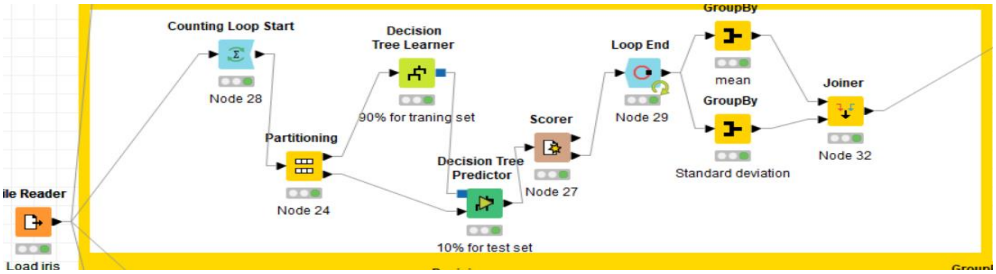


Figure 9: hold-out method

Val-10f: 10-fold cross-validation used the k-fold cross-validation method and needs to use the X-Partitioner and X-Aggregator nodes. Specifying thenumber of folds in the X-Partitioner node, k=10 used in

this case.

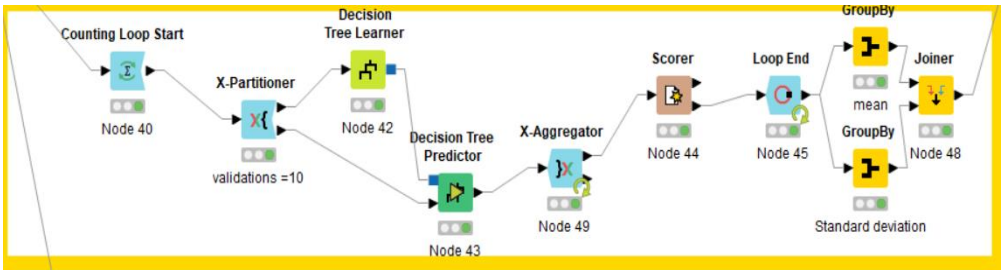


Figure10: 10-fold cross-validation

LOOCV: leave-one-out cross-validation method

LOOCV method was used in the same X-Partitioner and X-Aggregator nodes that used in the Val-10f method the distinction is selected the leave-one-out.

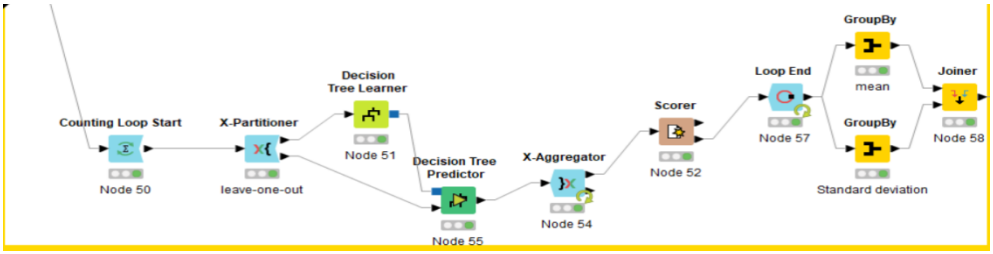
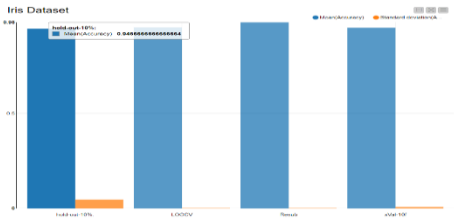
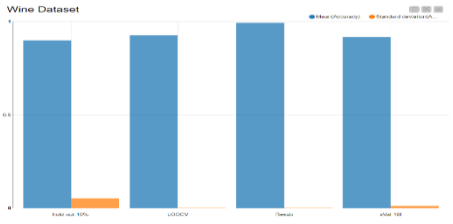


Figure 11: LOOCV method

The results are presented employing a bar chart for each dataset to compare the four methods.





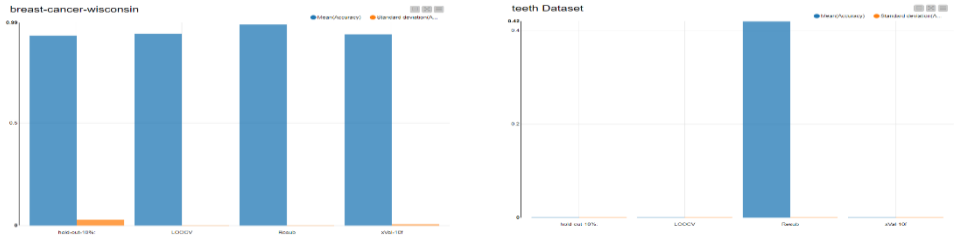


Figure 12 chart for each dataset to compare the four methods

Iris dataset:

Mean(A...	Standa...	Metho
0.98	0	Resub
0.947	0.046	hold-out-1
0.951	0.007	xVal-10f
0.953	0	LOOCV

Wine dataset:

Mean(A...	Standa...	Metho
0.994	0	Resub
0.925	0.063	hold-out-1
0.921	0.015	xVal-10f
0.927	0	LOOCV

breast-cancer-Wisconsin dataset:

Mean(A...	Standa...	Metho
0.981	0	Resub
0.926	0.029	hold-out-1
0.932	0.006	xVal-10f
0.936	0	LOOCV

Teeth dataset:

Mean(A...	Standa...	Method
0.419	0	Resub
0	0	hold-out-10...
0	0	xVal-10f
0	0	LOOCV

TABLE 5 RESULT OF DATASET TO COMPARE THE FOUR METHODS

3.2.2Task 2 Decision of Tree algorithm in R

resub: resubstitution error method for this method.

Iris dataset

```
holdout_acc<-c()
for (n in 1:20) {
#H= holdout(1:5,ratio=0.9, mode="random",seed=NULL)
#print (H)
# random stratified holdout
```

```
H=holdout(iris$class,ratio=0.9,mode="random",seed=NULL)
train_data_iris<-(table(iris[H$tr,]$class))
test_data_iris<-(table(iris[H$ts,]$class))
test_Y_iris<- (table(iris[H$ts,]$class))
train_data_iris
test_data_iris
test_Y_iris
iris_model<- train(class~ ., method="rpart",data=train_data_iris)
y_iris=predict(iris_model, newdata=test_data_iris)
y1_iris<-as.character(y_iris)
print (y1_iris)
holdout_acc$iris[n]<-mean(test_Y_iris==y1_iris)
}
```

Teeth dataset

```
for (n in 1:20){
train_data_teeth=teeth
train_test_teeth=teeth
test_Y_teeth=teeth$Animal
teeth_model<- train(Animal~ ., method="rpart",data=train_data_teeth)
y_teeth=predict(teeth_model, newdata=train_test_teeth)
y1_teeth<-as.character(y_teeth)
resub_AccuracyData$teeth[n]<- mean(test_Y_teeth==y1_teeth)
}
```

Wine dataset

```
for (n in 1:20){
train_data_wine=wine
```

```
 train_test_wine=wine
 test_Y_wine=wine$Alcohol
 wine_model<- train(Alcohol~ ., method="rpart",data=train_data_wine)
 y_wine=predict(wine_model, newdata=train_test_wine)
 y1_wine<-as.character(y_wine)
 resub_AccuracyData$wine[n]<- mean(test_Y_wine==y1_wine)
}
```

breast-cancer-wisconsin dataset

```
for (n in 1:20){
 train_data_breast-cancer-wisconsin=breast-cancer-wisconsin
 train_test_breast-cancer-wisconsin=breast-cancer-wisconsin
 test_Y_breast-cancer-wisconsin=breast-cancer-wisconsin$class
 breast-cancer-wisconsin_model<- train(class~ ., method="rpart",data=train_data_breast-cancer-wisconsin)
 y_breast-cancer-wisconsin=predict(breast-cancer-wisconsin_model, newdata=train_test_breast-cancer-wisconsin)
 y1_breast-cancer-wisconsin<-as.character(y_breast-cancer-wisconsin)
 resub_AccuracyData$breast-cancer-wisconsin[n]<- mean(test_Y_breast-cancer-wisconsin==y1_breast-cancer-wisconsin)
}
```

hold-out-10%: hold-out method with 10% - 90%

Teeth dataset

```
holdout_acc<-c()
for (n in 1:20) {
 H=holdout(teeth$Animal,ratio=0.9,mode="random",seed=NULL)
 train_data_teeth<-(table(teeth[H$tr,]$Animal))
```

```
 test_data_teeth<-(table(teeth[H$ts,]$Animal))
 test_Y_teeth<- (table(teeth[H$ts,]$Animal))
 train_data_teeth
 test_data_teeth
 test_Y_teeth
 teeth_model<- train(Animal~ ., method="rpart",data=train_data_teeth)
 y_teeth=predict(teeth_model, newdata=test_data_teeth)
 y1_teeth<-as.character(y_teeth)
 print (y1_teeth)
 holdout_acc$teeth[n]<-mean(test_Y_teeth==y1_teeth)
}
holdout_acc<-c()
for (n in 1:20) {
 H=holdout(wine$Alcohol,ratio=0.9,mode="random",seed=NULL)
 train_data_wine<-(table(wine[H$tr,]$Alcohol))
 test_data_wine<-(table(wine[H$ts,]$Alcohol))
 test_Y_wine<- (table(wine[H$ts,]$Alcohol))
```

Wine dataset

```
 train_data_wine
 test_data_wine
 test_Y_wine
 wine_model<- train(Alcohol~ ., method="rpart",data=train_data_wine)
 y_wine=predict(wine_model, newdata=test_data_wine)
 y1_wine<-as.character(y_wine)
 print (y1_wine)
```

```
 holdout_acc$wine[n]<-mean(test_Y_wine==y1_wine)
}
holdout_acc<-c()
for (n in 1:20) {
  H=holdout(breast-cancer-wisconsin$class,ratio=0.9,mode="random",seed=NULL)
 train_data_breast-cancer-wisconsin<-(table(breast-cancer-wisconsin[H$tr,]$class))
 test_data_breast-cancer-wisconsin<-(table(breast-cancer-wisconsin[H$ts,]$class))
 test_Y_breast-cancer-wisconsin<- (table(breast-cancer-wisconsin[H$ts,]$class))
```

breast-cancer-wisconsin dataset

```
 train_data_breast-cancer-wisconsin
 test_data_breast-cancer-wisconsin
 test_Y_breast-cancer-wisconsin
 breast-cancer-wisconsin_model<- train(class~ ., method="rpart",data=train_data_breast-cancer-wisconsin)
 y_breast-cancer-wisconsin=predict(breast-cancer-wisconsin_model, newdata=test_data_breast-cancer-wisconsin)
 y1_breast-cancer-wisconsin<-as.character(y_breast-cancer-wisconsin)
 print (y1_breast-cancer-wisconsin)
 holdout_acc$breast-cancer-wisconsin[n]<-mean(test_Y_breast-cancer-wisconsin==y1_breast-cancer-wisconsin)
}
```

xVal-10f: 10-fold cross-validation method.

Iris dataset	**xVal_acc<-c()** **for(n in 1:20){** **iris_model<train(class ~., data=iris_data, method="report",parms=list(split="information", tr.control=trctrl))** **accuracy<iris_model[["resample"]][["Accuracy"]]** **xVal_acc$iris[n]<- accuracy** **}**
Teeth dataset	**xVal_acc<-c()** **for(n in 1:20){** **teeth_model<train(Animal ~., data=teeth_data, method="report",parms=list(split="information", tr.control=trctrl))** **accuracy<teeth_model[["resample"]][["Accuracy"]]** **xVal_acc$teeth[n]<- accuracy** **}**
Wine dataset	**xVal_acc<-c()** **for(n in 1:20){** **wine_model<train(class ~., data=wine_data, method="report",parms=list(split="information", tr.control=trctrl))** **accuracy<wine_model[["resample"]][["Accuracy"]]** **xVal_acc$wine[n]<- accuracy** **}**
breast-cancer-wisconsin dataset	**xVal_acc<-c()** **for(n in 1:20){** **breast-cancer-wisconsin_model<train(Alcohol ~., data=breast-cancer-wisconsin_data, method="report",parms=list(split="information", tr.control=trctrl))** **accuracy<breast-cancer-wisconsin_model[["resample"]][["Accuracy"]]** **xVal_acc$breast-cancer-wisconsin[n]<- accuracy** **}**

LOOCV: leave-one-out cross-validation method

Iris dataset	**trctrl<- trainControl(method = "LOOCV")** **for (n in 1:20){** **iris_model<- train(class~., data=iris_data, method="report", parms=list(split="information"),tr.control=trctrl)** **accuracy<-iris_model[["results"]][[Accuracy]][1]** **loocv_acc$iris[n]<-accuracy** **}**
Teeth dataset	**trctrl<- trainControl(method = "LOOCV")** **for (n in 1:20){** **teeth_model<- train(Animal~., data=teeth_data, method="report", parms=list(split="information"),tr.control=trctrl)** **accuracy<-teeth_model[["results"]][[Accuracy]][1]** **loocv_acc$teeth[n]<-accuracy** **}**
Wine dataset	**trctrl<- trainControl(method = "LOOCV")** **for (n in 1:20){** **wine_model<- train(class~., data=wine_data, method="report", parms=list(split="information"),tr.control=trctrl)** **accuracy<-wine_model[["results"]][[Accuracy]][1]** **loocv_acc$wine[n]<-accuracy** **}**
breast-cancer-wisconsin dataset	**for (n in 1:20){** **breast-cancer-wisconsin_model<- train(Alcohol~., data=breast-cancer-wisconsin_data, method="report", parms=list(split="information"),tr.control=trctrl)** **accuracy<-breast-cancer-wisconsin_model[["results"]][[Accuracy]][1]** **loocv_acc$breast-cancer-wisconsin[n]<-accuracy}**

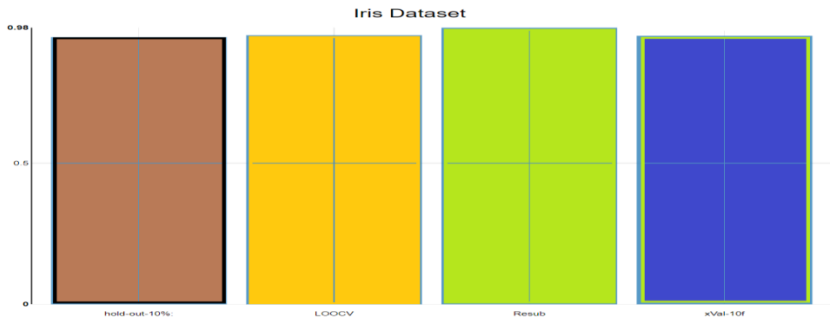


Figure13: chart of Iris Dataset

4- CONCLUSION

In this paper, we used the practical differences of the four validation methodsusing 4 distinct datasets Iris dataset, the Wine dataset, the Breast-Cancer-Wisconsin dataset, and the Teeth dataset are analyzed, the Classification and prediction data to find the optimal solution for data classifying. the performance indicators accuracy, specificity, sensitivity, precision, the error rate is calculated for the given dataset. Accusation besides a proper data preprocessing technique can get better the accuracy of the classifier. The function of data normalization had a noticeable impact on categorization performance and considerably enhanced the performance of the Iris dataset technique. The performance of the Teeth dataset method has minimum accuracy. Based on the analysis, the performances of the validation methods are analyzed. The results show that the performance of the Iris dataset technique is significantly superior to the other three techniques for the classification of data. To improve the overall accuracy, it is necessary to use more data set with a large number of attributes and use the best feature selection method in the future. Future works may also include hybrid classification models by combining some of the dataset techniques.

REFERENCES

[1]. Bilalli, B., Abelló, A., Aluja-Banet, T. and Wrembel, R., 2016, April. Towards Intelligent Data Analysis: The Metadata Challenge. In IoTBD (pp. 331-338).

[2]. Govender, P. and Sivakumar, V., 2020. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). Atmospheric Pollution Research, 11(1), pp.40-56.

[3]. David, S.K., Saeb, A.T., Rafiullah, M. and Rubeaan, K., 2019. Classification Techniques and Data Mining Tools Used in Medical Bioinformatics. In Big Data Governance and Perspectives in Knowledge Management (pp. 105-126). IGI Global.

[4]. Brijain, M., Patel, R., Kushik, M. and Rana, K., 2014. A survey on decision tree algorithm for classification.

[5]. Liao, T.W., 2005. Clustering of time series data—a survey. Pattern recognition, 38(11), pp.1857-1874.

[6]. Dubes, R. and Jain, A.K., 1976. Clustering techniques: the user's dilemma. Pattern Recognition, 8(4), pp.247-260.

[7]. Sanakal Ravi and Jayakumari T, 2014. Prognosis of Diabetes Using Data Mining Approach-Fuzzy C Means Clusteringand Support Vector Machine. International Journal of Computer Trends and Technology.

[8]. Sharma Arvind and Gupta PC, 2012. Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool. International Journal of Communication and Computer Technologies. (01),pp.6-10

[9]. Salama GI, Abdelhalim MB, Zeid MA 2012. Experimental comparison of classification for breast cancer diagnosis. International Conference on Computer Engineering &Systems. pp, (98-180-5)