A Iris Data Set

Shiza Mushtaq

June 27, 2021

A IRIS DATA SET

MS COMPUTER SCIENCE (2 Years)

SESSION 2021-2023

DEPARTMENT OF COMPUTER SCIENCE

THE ISLAMIA UNIVERSITY OF BAHAWALPUR

ASSIGNMENT SUPERVISOR:

DR.IMRAN SARWAR BAJWA

UNDERTAKEN BY:

SHIZA MUSHTAQ

ROLL NO:

S21BDOCS3E01044

ABSTRACAT

The Iris flower data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper the use of multiple measurements in taxonomic problems. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. The data set consists of 50 samples from each of three species of Iris (Iris Setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. This dataset became a typical test case for many statistical classification techniques in machine learning such as support vector machines. The dataset contains a set of 150 records under 5 attributes - Petal Length, Petal Width, Sepal Length, Sepal width and Class (Species).

Keywords: Selection of attributes; Replace Missing value; Normalization; K-means

INTRODUCTION

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician, eugenicist, and biologist Ronald Fisher in his 1936 paper *The use of multiple* measurements in taxonomic problems as an example of linear discriminant analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus". The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.

RELATED WORK

- Hashemite Kingdom of Jordan - 2009, IrisGuard deployed one of the world's first operational iris-enabled automated teller machine at Cairo Amman Bank, where bank customers can seamlessly withdraw cash from ATM's without a bank card or pin but simply by presenting their eye to the iris recognition camera on the ATM. Since June 2012, IrisGuard is also providing financial inclusion to UNHCR registered Syrian refugees in Jordan on ATM's. The system is designed to facilitate cash-supported interventions that help deliver financial assistance to refugees with speed and dignity while lowering overhead costs and boosting accountability.
- Aadhaar began operation in 2011 in India, whose government is enrolling the iris patterns (and other biometrics) of more than one billion residents for the Aadhaar scheme for entitlements distribution, run by the Unique Identification Authority of India (UIDAI). This programmer at its peak was enrolling about one million persons every day, across 36,000 stations operated by 83 agencies. By October 2015, the number of persons enrolled exceeded 926 million, with each new enrollee being compared to all existing ones for de-duplication checks (hence 926 trillion, i.e. 926 million-million, iris cross-comparisons per day). Its purpose is to issue residents a biometrically provable unique entitlement number (Aadhaar) by which benefits may be claimed, and social inclusion enhanced; thus the slogan of UIDAI is: "To give the poor an identity." Iris technology providers must be granted a STQC (Standardisation Testing and Quality Certification) certificate in order to supply iris scanners for the project. By far, there are providers such as: IriTech Inc. (dual iris scanner IriMagic 100BK), Cogent (CIS-202), Iris ID (icam TD 100), Iris Guard (IG-AD-100), etc

USED APPROACH

1) SELECTION OF ATTRIBUTES

Attribute selection is the process of identifying relevant information and removing as much of the irrelevant and redundant information as possible. Attribute selection is also defined as "the process of finding a best subset of features, from the original set of features in a given data set, optimal according to the defined goal and criterion of feature selection (a feature goodness criterion)".

2) REPLACE ALL MISSING VALUE

Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values.

3) NORMALIZATION

Normalization is used to avoid these problems by creating new values that maintain the general distribution and ratios in the source data while keeping values within a scale applied across all numeric columns used in the model. In this research, normalization was applied on all columns to change all values to a 0–1 scale in the dataset.

4) K-MEANS

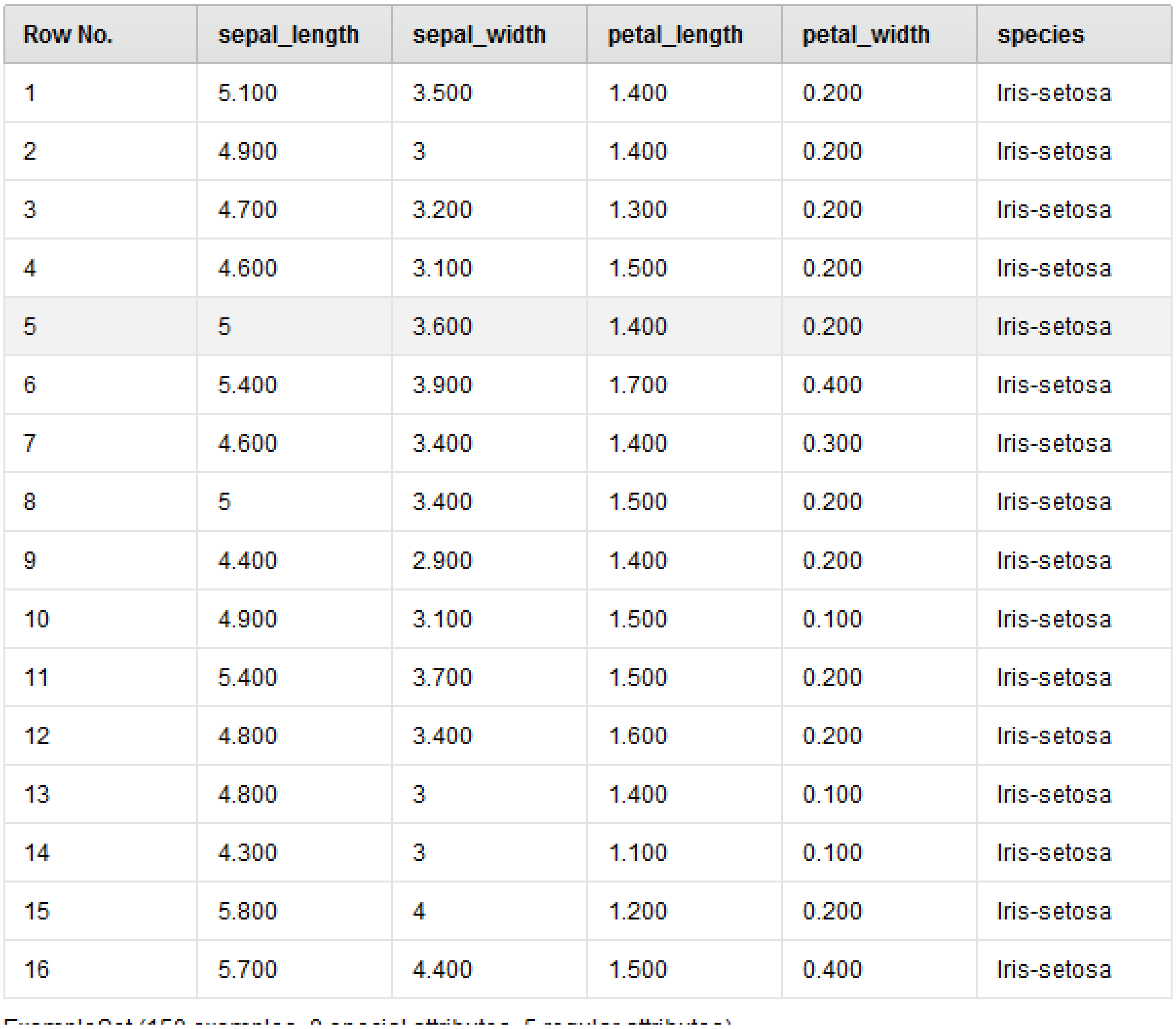
Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data. cluster, described in (Hartigan, 1975), is one of the most basic and commonly used clustering algorithms. Firstly, it requires specification in advance of how many clusters are to be generated. This makes it ideal for use with the Cluster Classifier. The number 7 of clusters is the parameter K for which the algorithm is named. K points are chosen at random to form the initial cluster centres. All data points are assigned to the nearest cluster by the Euclidean distance. The cluster centre is then recalculated by taking the mean attribute values of all data points in a cluster. The process is then repeated with these new cluster centres until there are no data points reassigned between two iterations. The clusters obtained by this process can be shown to be a local minimum of the sum of the Euclidean distance from each data point to it's cluster centre. There is no guarantee, however, of this result being the best cluster assignment that could have been obtained from this dataset since the result will often fall in a local minimum. The clusters that result from this clustered are highly dependent on the initial random selection process

EXPERIMENT AND RESULTS

Different classifiers were used to compare their statistical measures of performance with K-mean on the iris analysis dataset, data selection, Replace missing value, normalization , k-mean clustering. in order to obtain the best statistical measures of performance for each classifier in terms of sensitivity, specificity, and predictive value.

Table 1:

IRIS Data Analysis.

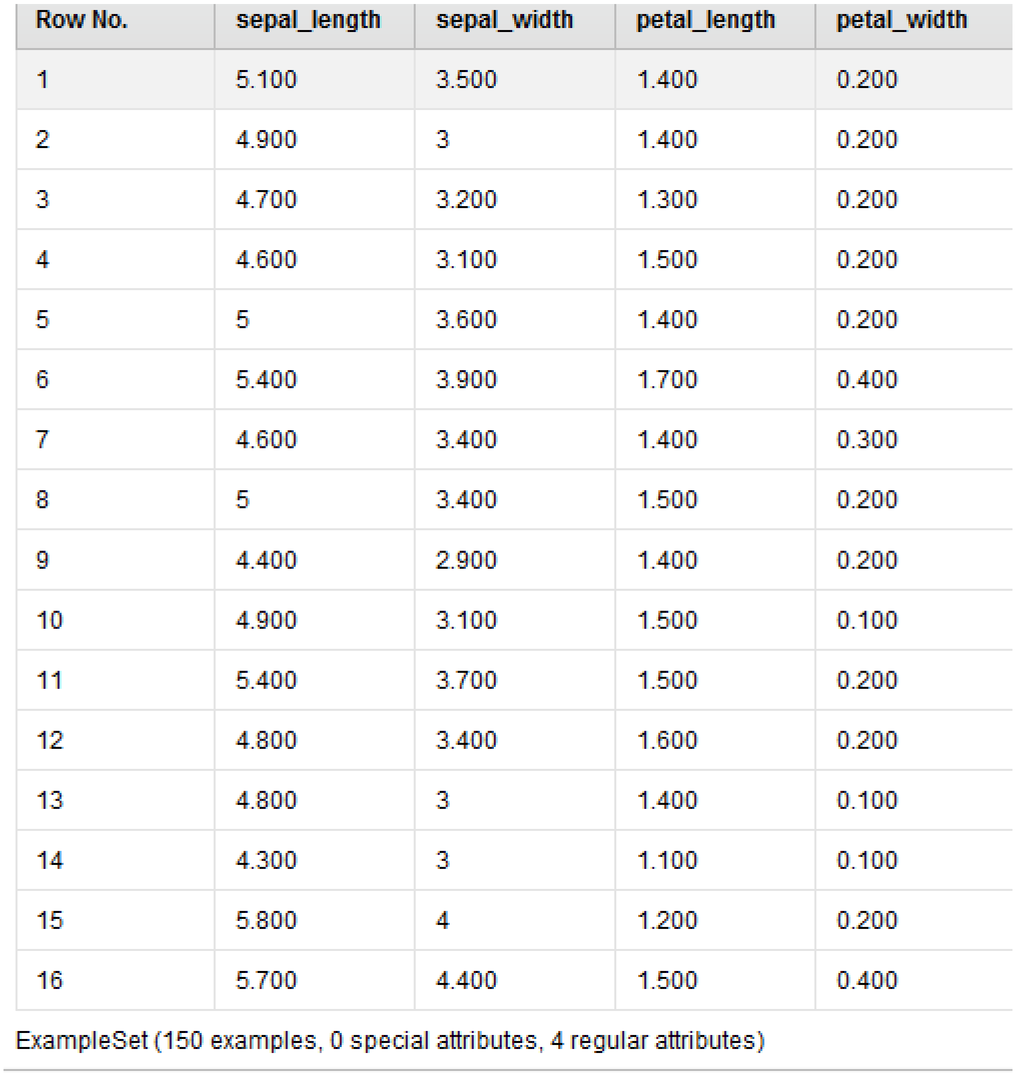


Row No.	sepal_length	sepal_width	petal_length	petal_width	species
1	5.100	3.500	1.400	0.200	Iris-setosa
2	4.900	3	1.400	0.200	Iris-setosa
3	4.700	3.200	1.300	0.200	Iris-setosa
4	4.600	3.100	1.500	0.200	Iris-setosa
5	5	3.600	1.400	0.200	Iris-setosa
6	5.400	3.900	1.700	0.400	Iris-setosa
7	4.600	3.400	1.400	0.300	Iris-setosa
8	5	3.400	1.500	0.200	Iris-setosa
9	4.400	2.900	1.400	0.200	Iris-setosa
10	4.900	3.100	1.500	0.100	Iris-setosa
11	5.400	3.700	1.500	0.200	Iris-setosa
12	4.800	3.400	1.600	0.200	Iris-setosa
13	4.800	3	1.400	0.100	Iris-setosa
14	4.300	3	1.100	0.100	Iris-setosa
15	5.800	4	1.200	0.200	Iris-setosa
16	5.700	4.400	1.500	0.400	Iris-setosa

ExampleSet (150 examples, 0 special attributes, 5 regular attributes)

Table 2:

Selection of Attribute.

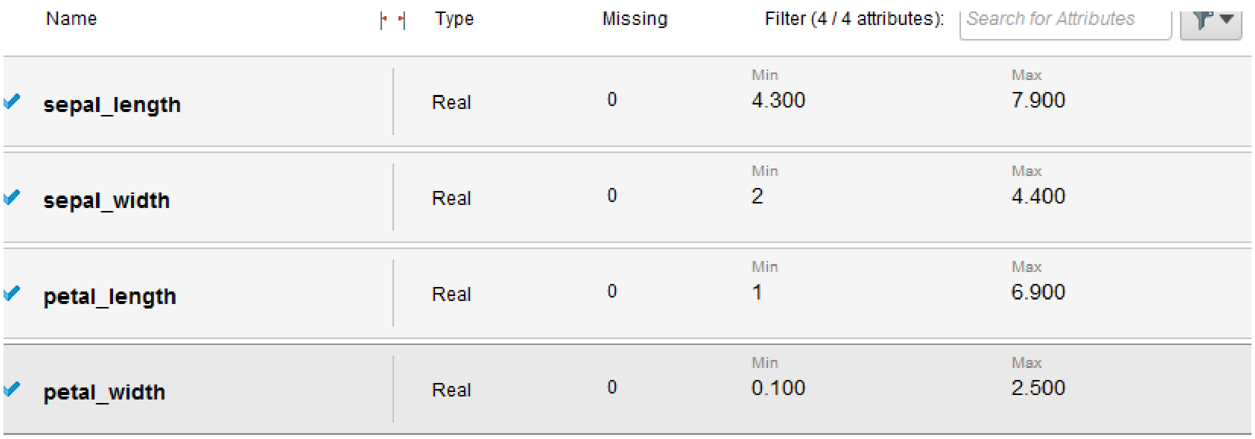


Row No.	sepal_length	sepal_width	petal_length	petal_width
1	5.100	3.500	1.400	0.200
2	4.900	3	1.400	0.200
3	4.700	3.200	1.300	0.200
4	4.600	3.100	1.500	0.200
5	5	3.600	1.400	0.200
6	5.400	3.900	1.700	0.400
7	4.600	3.400	1.400	0.300
8	5	3.400	1.500	0.200
9	4.400	2.900	1.400	0.200
10	4.900	3.100	1.500	0.100
11	5.400	3.700	1.500	0.200
12	4.800	3.400	1.600	0.200
13	4.800	3	1.400	0.100
14	4.300	3	1.100	0.100
15	5.800	4	1.200	0.200
16	5.700	4.400	1.500	0.400

ExampleSet (150 examples, 0 special attributes, 4 regular attributes)

Table 3:

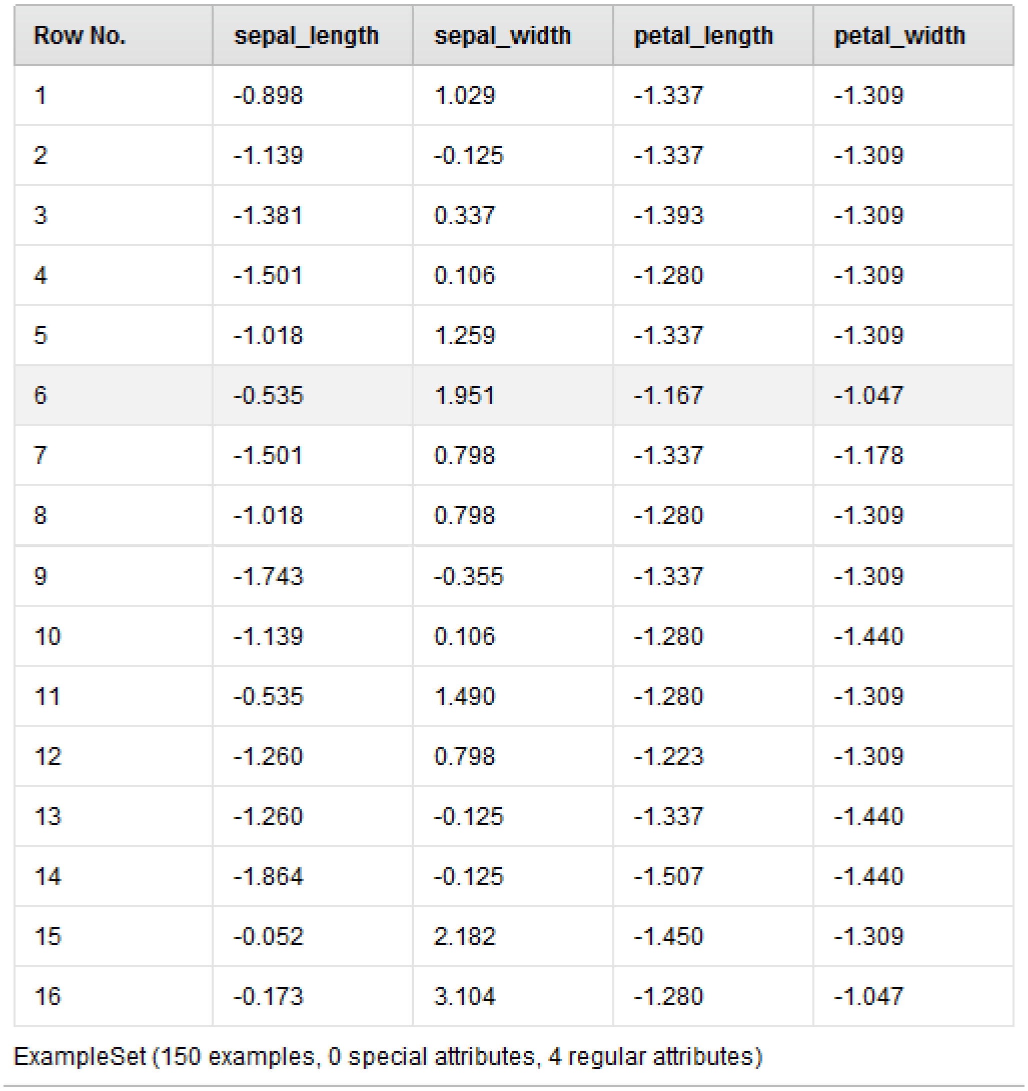
Replace All Missing Value.



	Name	Type	Missing	Min	Max
✓	**sepal_length**	Real	0	4.300	7.900
✓	**sepal_width**	Real	0	2	4.400
✓	**petal_length**	Real	0	1	6.900
✓	**petal_width**	Real	0	0.100	2.500

Table 4:

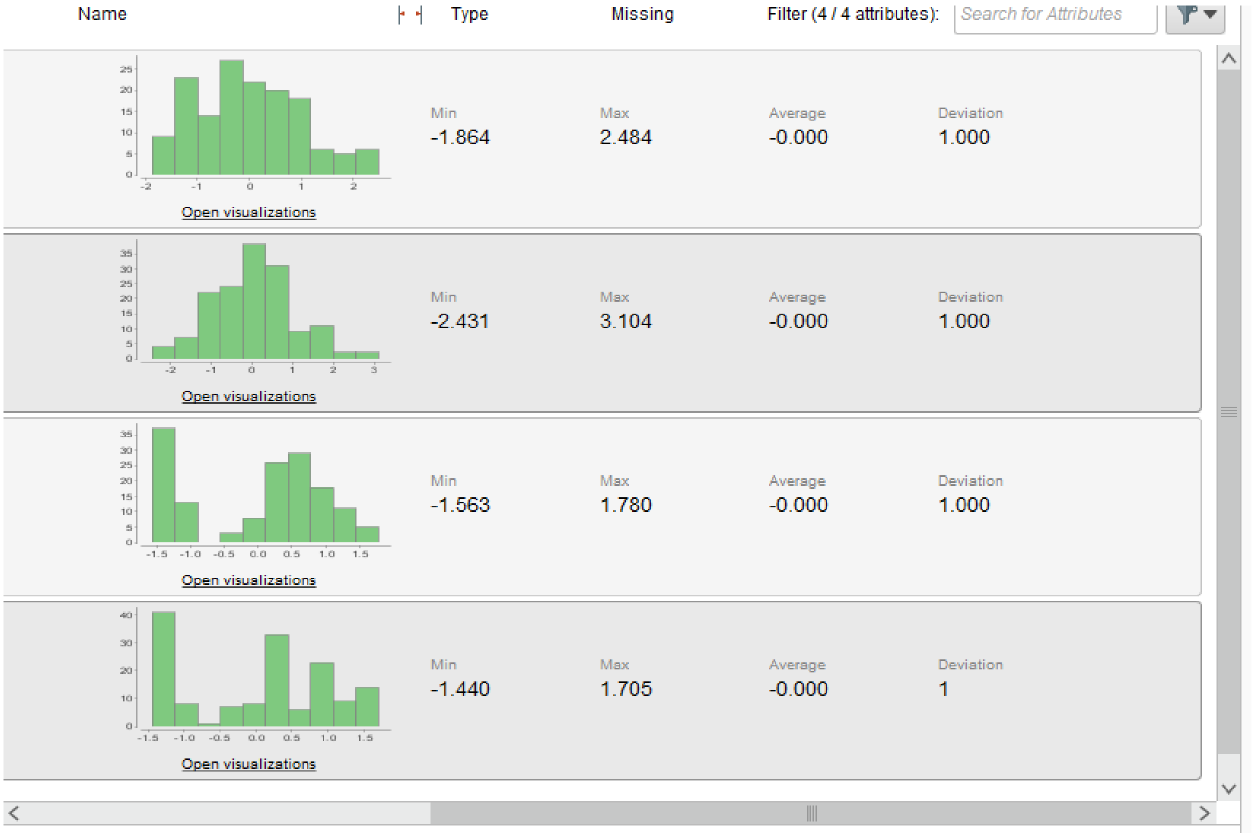
Normalization.



Row No.	sepal_length	sepal_width	petal_length	petal_width
1	-0.898	1.029	-1.337	-1.309
2	-1.139	-0.125	-1.337	-1.309
3	-1.381	0.337	-1.393	-1.309
4	-1.501	0.106	-1.280	-1.309
5	-1.018	1.259	-1.337	-1.309
6	-0.535	1.951	-1.167	-1.047
7	-1.501	0.798	-1.337	-1.178
8	-1.018	0.798	-1.280	-1.309
9	-1.743	-0.355	-1.337	-1.309
10	-1.139	0.106	-1.280	-1.440
11	-0.535	1.490	-1.280	-1.309
12	-1.260	0.798	-1.223	-1.309
13	-1.260	-0.125	-1.337	-1.440
14	-1.864	-0.125	-1.507	-1.440
15	-0.052	2.182	-1.450	-1.309
16	-0.173	3.104	-1.280	-1.047

ExampleSet (150 examples, 0 special attributes, 4 regular attributes)

Table 5:

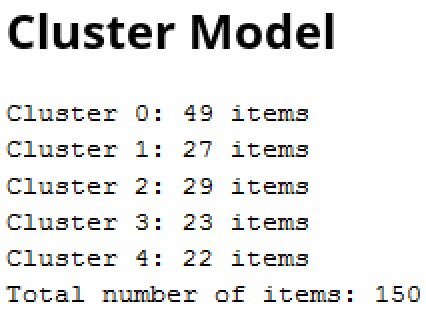


Name Type Missing Filter (4 / 4 attributes):

Min	Max	Average	Deviation
-1.864	2.484	-0.000	1.000
-2.431	3.104	-0.000	1.000
-1.563	1.780	-0.000	1.000
-1.440	1.705	-0.000	1

Table 6:

K-Mean

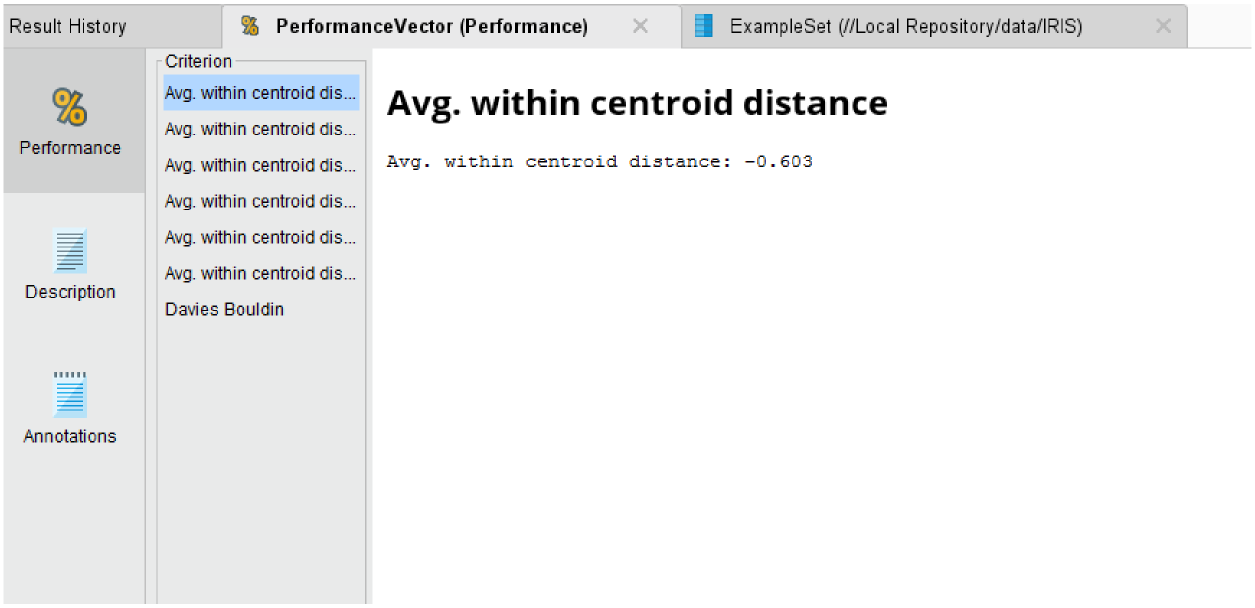


Cluster Model

```
Cluster 0: 49 items
Cluster 1: 27 items
Cluster 2: 29 items
Cluster 3: 23 items
Cluster 4: 22 items
Total number of items: 150
```

Table 7:

Cluster Performance Distance.


Result History
PerformanceVector (Performance)
ExampleSet (//Local Repository/data/IRIS)
Performance
Description
Annotations
Criterion
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Avg. within centroid dis...
Davies Bouldin

Avg. within centroid distance

```
Avg. within centroid distance: -0.603
```

Conclusion

1. Check for identifiable clusters in the data - We were able to use K-Means clustering to identify patterns and the algorithm tended to group the data based on their species.
2. Assess the accuracy of these clusters based on species - We were able to assess the accuracy of the clustering methods by looking at the number of correctly categorized data points against the whole population.
3. Asses the accuracy of the model's predictions based on actual species - We were able to assess the accuracy of the models using the same approach as we did the clustering methods.

REFRENCES

1. *R. A. Fisher (1936). "The use of multiple measurements in taxonomic problems". Annals of Eugenics. **7** (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x. hdl:2440/15227.*
2. *^ Edgar Anderson (1936). "The species problem in Iris". Annals of the Missouri Botanical Garden. **23** (3): 457–509. doi:10.2307/2394164. JSTOR 2394164.*
3. *^ Edgar Anderson (1935). "The irises of the Gaspé Peninsula". Bulletin of the American Iris Society. **59**: 2–5.*
4. ^ Jump up to:[a] [b] A. N. Gorban, A. Zinovyev. Principal manifolds and graphs in practice: from molecular biology to dynamical systems, International Journal of Neural Systems, Vol. 20, No. 3 (2010) 219–232.
5. *^ "UCI Machine Learning Repository: Iris Data Set". archive.ics.uci.edu. Retrieved 2017-12-01.*

6. ^ *Ines Färber, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller, Erich Schubert, Thomas Seidl, Arthur Zimek (2010). "On Using Class-Labels in Evaluation of Clusterings" (PDF). In Xiaoli Z. Fern; Ian Davidson; Jennifer Dy (eds.). MultiClust: Discovering, Summarizing, and Using Multiple Clusterings. ACM SIGKDD.*
7. ^ A.N. Gorban, N.R. Sumner, and A.Y. Zinovyev, Topological grammars for data approximation, Applied Mathematics Letters Volume 20, Issue 4 (2007), 382-386.
8. ^ *Bezdek, J.C. and Keller, J.M. and Krishnapuram, R. and Kuncheva, L.I. and Pal, N.R. (1999). "Will the real iris data please stand up?". IEEE Transactions on Fuzzy Systems.* ***7*** *(3): 368–369. doi:10.1109/91.771092*
9. Wikipedia, https://www.wikipedia.org
10. Wikipedia, http://axiomic.net/primers/primer2-p-value.pdf