Feven Belay Araya
ID- 20027
CS481 - Introduction to Data Science

HomeWork#1 and #2
Data- Mining Processes for Iris Dataset

1. Business Understanding(Short Description of iris plants)

The Iris flower is a genus that encompasses around 260–300 species within the Iridaceae family. It exhibits a wide variety of colors and patterns, making it not only a subject of admiration among gardeners and florists but also a significant model in botanical studies. Three species of Iris are often the focus of classification models:

- Iris Setosa: Known for its broad and blunt petals, the Iris Setosa is relatively simple to identify. It typically has a vibrant purple color with a pattern of deep veins and a distinctive pouch.
- Iris Versicolor: Also known as the Blue Flag Iris, it features narrower petals with a beautiful mix of purple and blue hues, often with some yellow and white patterns.
- Iris Virginica: Similar to the Versicolor but generally larger, the Iris Virginica has deeper blue and violet petals and is native to North America.

These species are often studied for their morphological variation and are particularly known for the differing lengths and widths of their sepals and petals, which are scientifically referred to as the flower's tepals.

In a data mining context, the Iris flower dataset is renowned for its use as a multivariate dataset introduced by the British statistician and biologist Ronald Fisher in 1936. It's often used for testing classification algorithms, including statistical, machine learning, and pattern recognition methods. The goal of an Iris classification model is to predict the species of an Iris plant based on the measurements of its sepals and petals, serving as a foundational example for teaching data classification concepts.

2. Data Understanding

The Iris dataset is a classic dataset in the field of machine learning and statistics, often used for demonstrating techniques in pattern recognition. Here's a short description covering the source, data format, and collection method:

- Source Link: The Iris dataset was originally introduced by the British statistician and biologist Ronald Fisher in his 1936 paper, "The use of multiple measurements in taxonomic problems." It is now publicly available on several data repositories, including the UCI Machine Learning Repository.
- Data Format: The dataset consists of 150 records, each with five attributes. There are four numerical features representing physical measurements: sepal length, sepal width, petal length, and petal width, all measured in centimeters. The fifth attribute is a categorical class label, which identifies the species of iris plant. There are three species included: Iris setosa, Iris virginica, and Iris versicolor, with 50 samples each.
- Data Collection Method: The data was collected from three species of Iris flowers (Iris setosa, Iris virginica, and Iris versicolor). The measurements of the physical characteristics (sepal length and width, petal length and width) were manually taken from

the flowers. Each class (species) in the dataset contains an equal number of instances (50 each), ensuring a balanced dataset for analysis.

3. Data Preparation

a) Exploratory Data Analysis (EDA) is a technique to analyze data using some visual Techniques. With this technique, we can get detailed information about the statistical summary of the data. We will also be able to deal with the duplicates values, outliers, and also see some trends or patterns present in the dataset.

b) Visual Inspection- Visual inspection in data mining is a critical and powerful technique used for exploring and analyzing data. In the context of data mining, visual inspection involves the use of visual representations to help detect patterns, outliers, trends, and relationships in large and complex datasets.

Refer Google collab file to see how implementation with commentary

4. Build a model

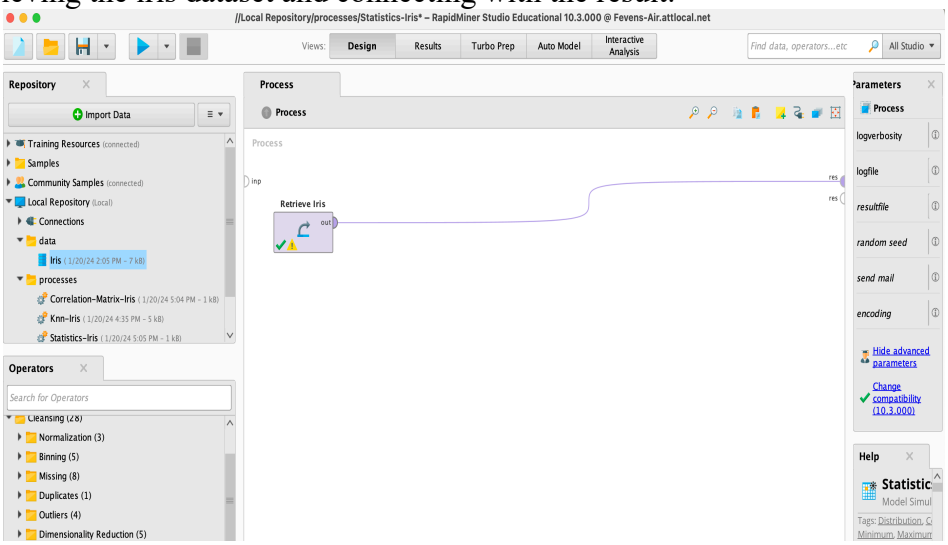
KNN is used at this case. The K-Nearest Neighbors (KNN) algorithm is a robust and intuitive machine learning method employed to tackle classification and regression problems. By capitalizing on the concept of similarity, KNN predicts the label or value of a new data point by considering its K closest neighbours in the training dataset. In this article, we will learn about a supervised learning algorithm (KNN) or the k – Nearest Neighbours, highlighting it's user-friendly nature

Refer Google collab file to see how implementation with commentary

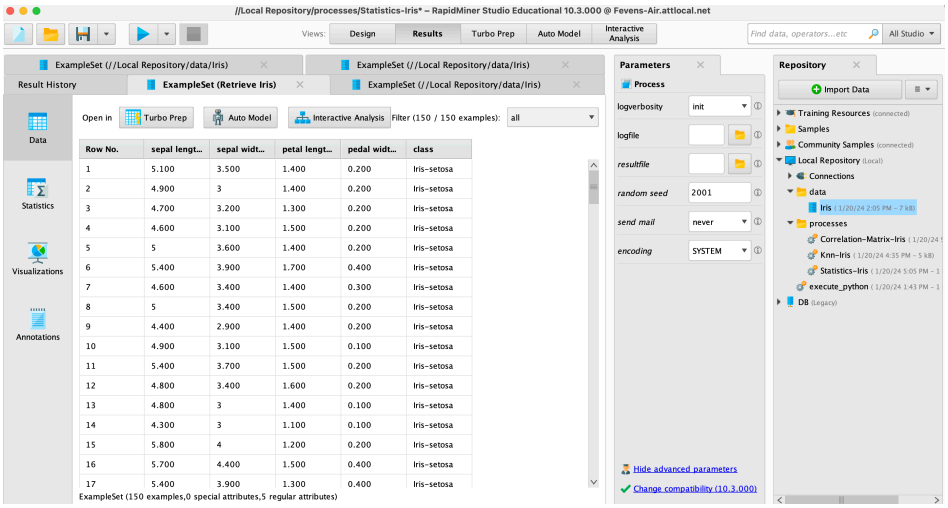
Extra- tried to work on RapidMiner as well

1. Exploratory Data Analysis

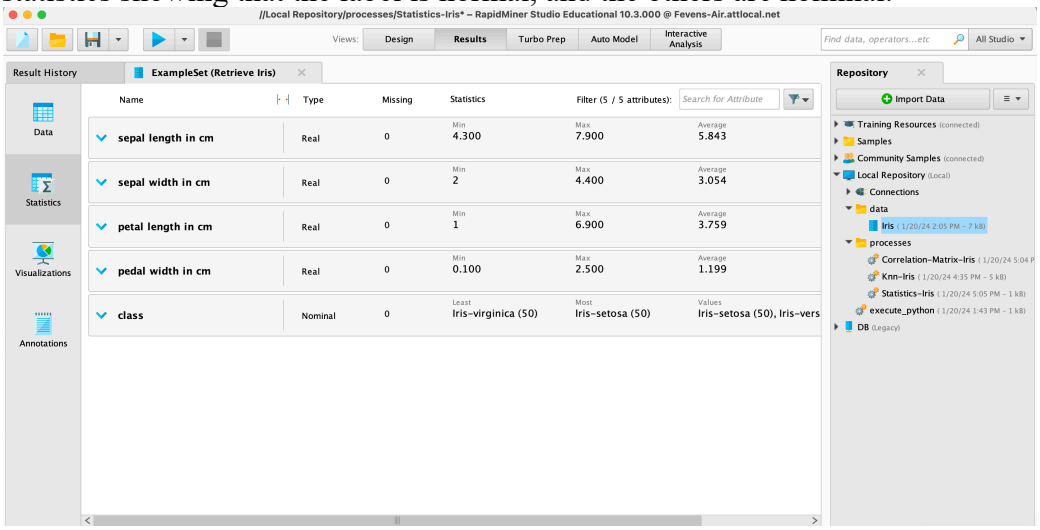
1. Retrieving the iris dataset and connecting with the result.



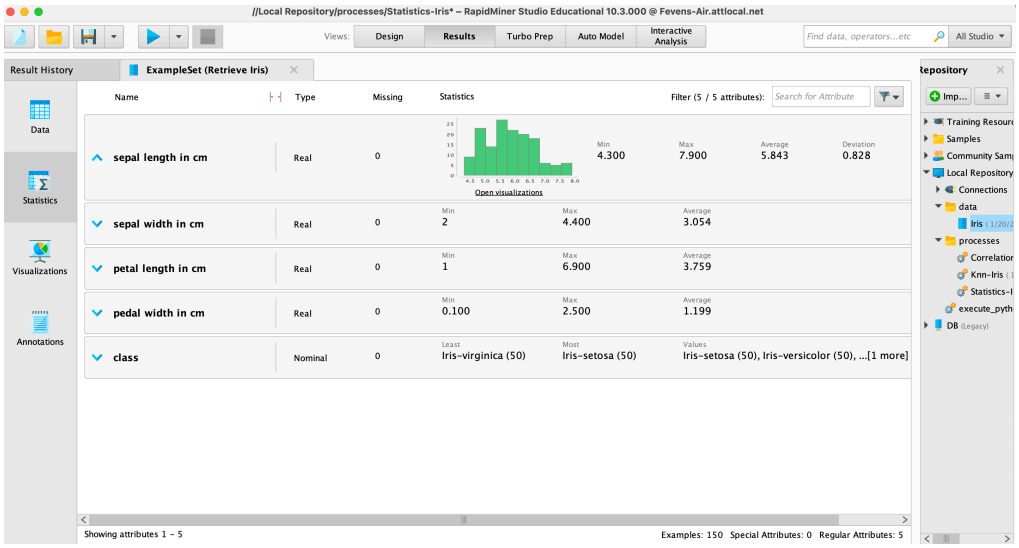
2. Click run on the RapidMiner. We have one labeled data named class and four regular attributes.



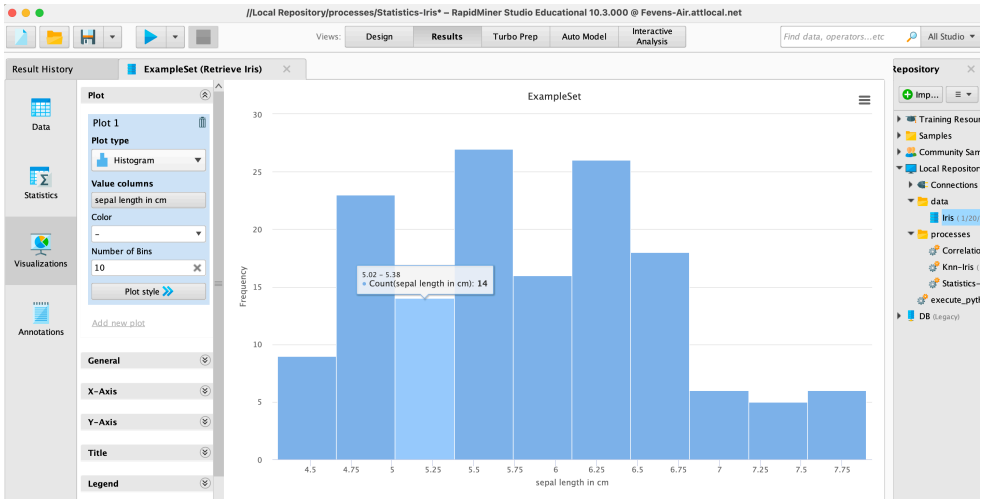
3. The statistics showing that the label is normal, and the others are nominal.



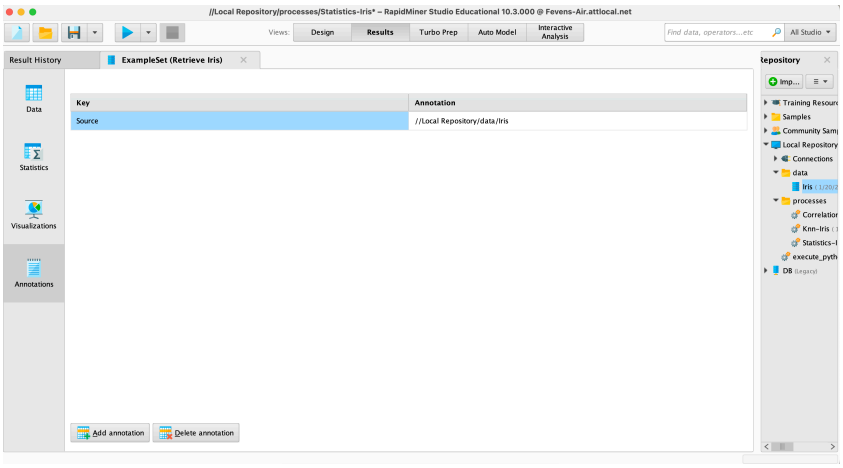
4. The visualization shows the distribution of the data points. For example, for sepal length in cm, the highest is 5.5 and the lowest is somewhere 7.5. The Min, Max and Average and deviation are also shown.



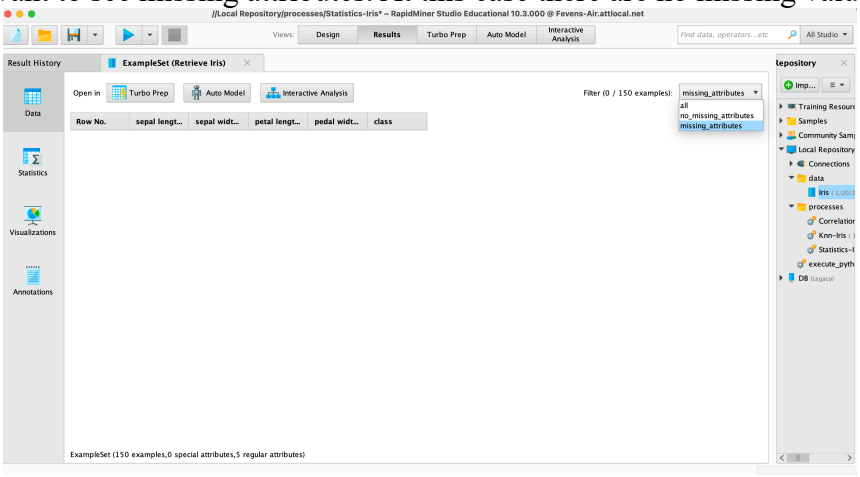
5. We can export the visualization and save it as svg file.

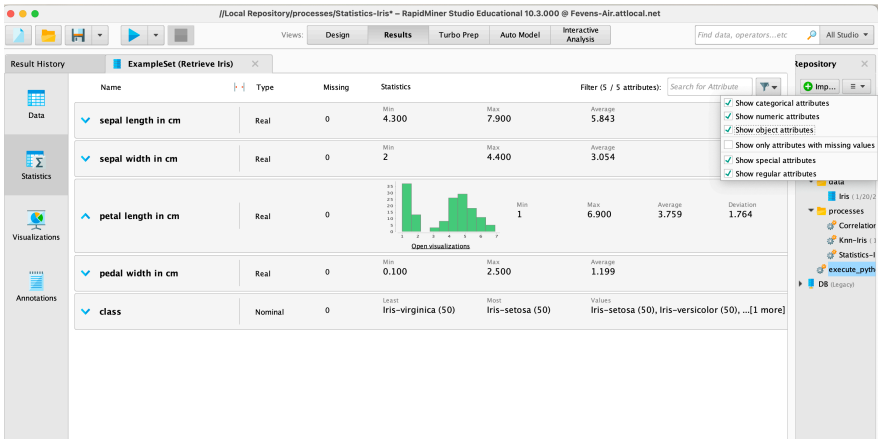


6. This shows the annotation of the iris file.



7. If you want to see missing attributes. At this case there are no missing values.

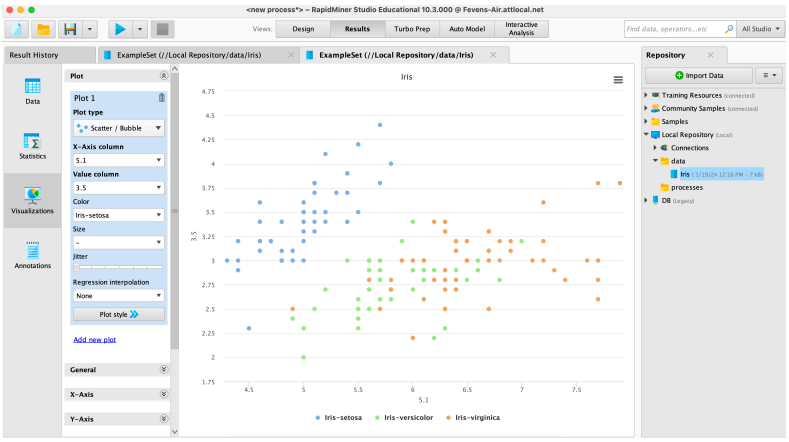


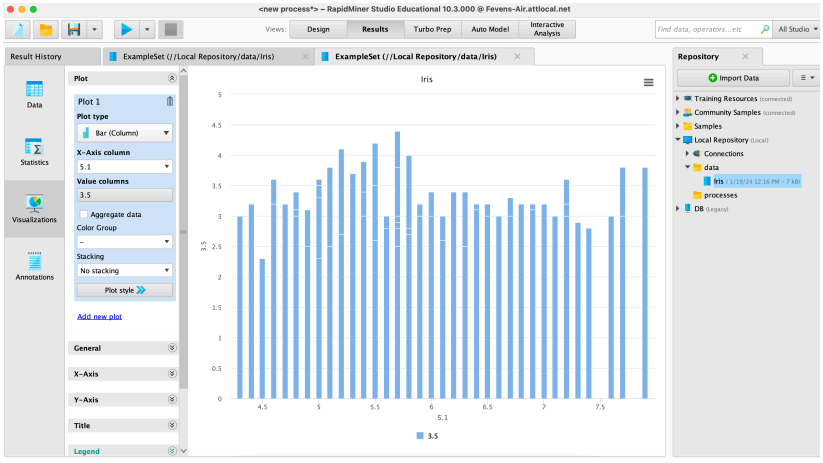


2. Data Visualization

Visualize data to understand distributions and relationships between variables.

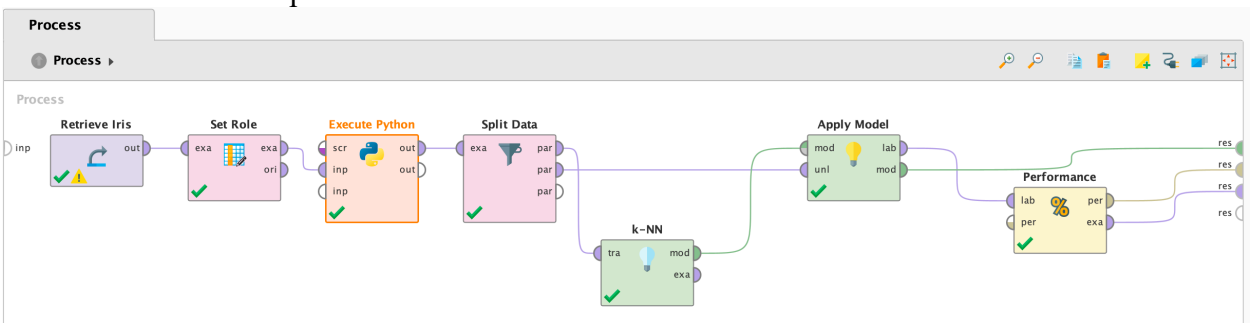
- Click on Virtualization from the Menu and select the Plot type you wanna view as follows.



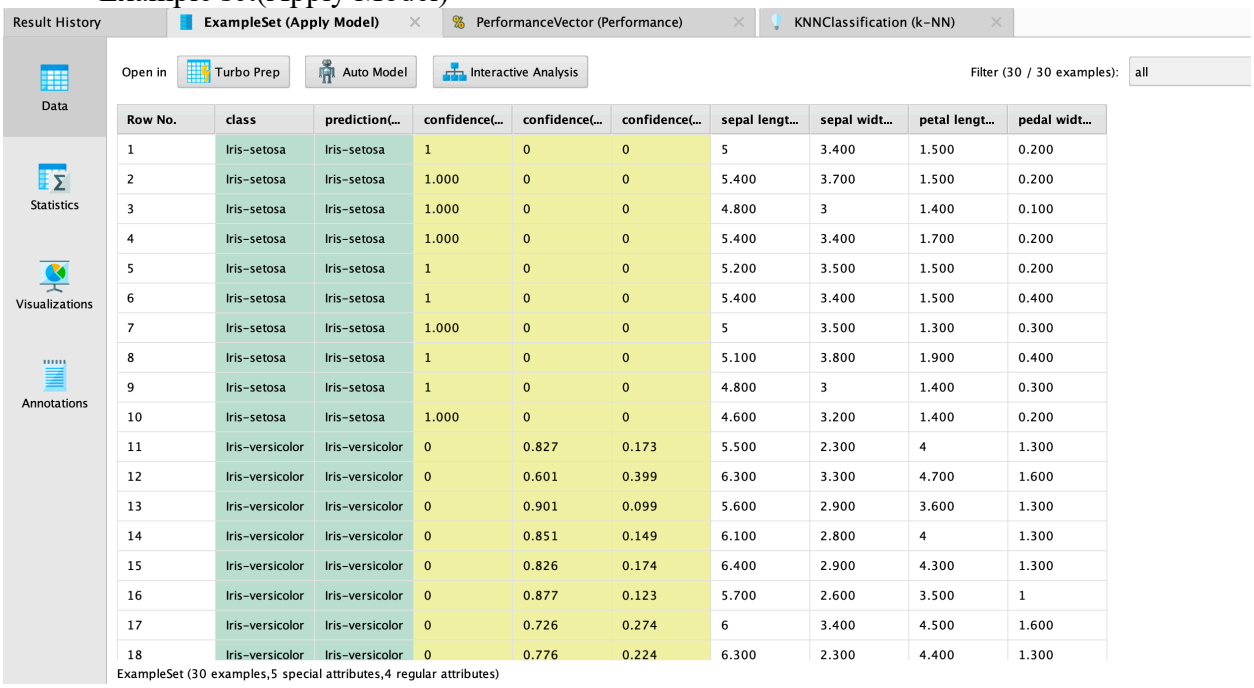


3. KNN classification

- Connected the processes as follows



- Example set(Apply Model)



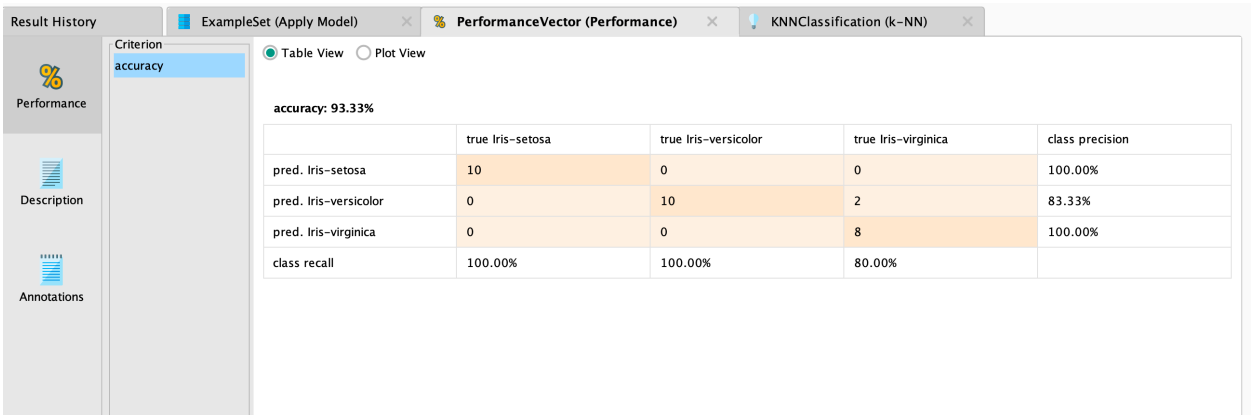
Result History | ExampleSet (Apply Model) | PerformanceVector (Performance) | KNNClassification (k-NN)

Open in: Turbo Prep | Auto Model | Interactive Analysis — Filter (30 / 30 examples): all

Row No.	class	prediction(...	confidence(...	confidence(...	confidence(...	sepal lengt...	sepal widt...	petal lengt...	pedal widt...
1	Iris-setosa	Iris-setosa	1	0	0	5	3.400	1.500	0.200
2	Iris-setosa	Iris-setosa	1.000	0	0	5.400	3.700	1.500	0.200
3	Iris-setosa	Iris-setosa	1.000	0	0	4.800	3	1.400	0.100
4	Iris-setosa	Iris-setosa	1.000	0	0	5.400	3.400	1.700	0.200
5	Iris-setosa	Iris-setosa	1	0	0	5.200	3.500	1.500	0.200
6	Iris-setosa	Iris-setosa	1	0	0	5.400	3.400	1.500	0.400
7	Iris-setosa	Iris-setosa	1.000	0	0	5	3.500	1.300	0.300
8	Iris-setosa	Iris-setosa	1	0	0	5.100	3.800	1.900	0.400
9	Iris-setosa	Iris-setosa	1	0	0	4.800	3	1.400	0.300
10	Iris-setosa	Iris-setosa	1.000	0	0	4.600	3.200	1.400	0.200
11	Iris-versicolor	Iris-versicolor	0	0.827	0.173	5.500	2.300	4	1.300
12	Iris-versicolor	Iris-versicolor	0	0.601	0.399	6.300	3.300	4.700	1.600
13	Iris-versicolor	Iris-versicolor	0	0.901	0.099	5.600	2.900	3.600	1.300
14	Iris-versicolor	Iris-versicolor	0	0.851	0.149	6.100	2.800	4	1.300
15	Iris-versicolor	Iris-versicolor	0	0.826	0.174	6.400	2.900	4.300	1.300
16	Iris-versicolor	Iris-versicolor	0	0.877	0.123	5.700	2.600	3.500	1
17	Iris-versicolor	Iris-versicolor	0	0.726	0.274	6	3.400	4.500	1.600
18	Iris-versicolor	Iris-versicolor	0	0.776	0.224	6.300	2.300	4.400	1.300

ExampleSet (30 examples,5 special attributes,4 regular attributes)

- Performance Vector



Result History | ExampleSet (Apply Model) | PerformanceVector (Performance) | KNNClassification (k-NN)

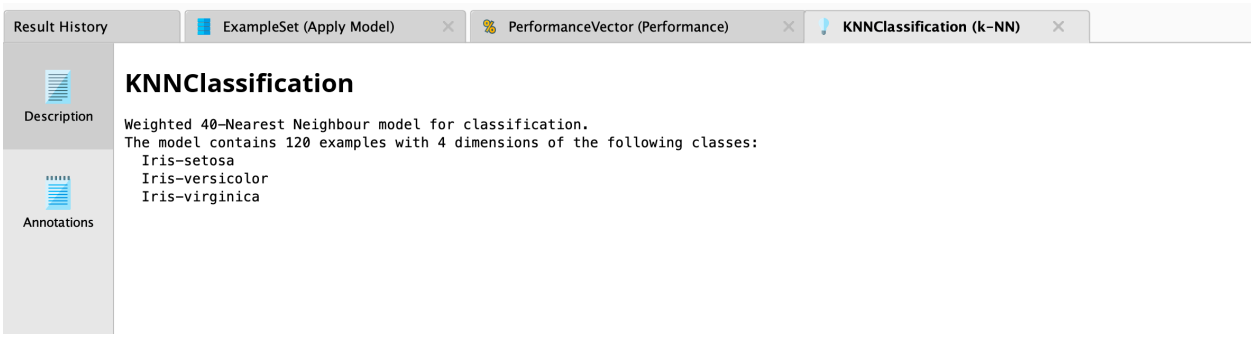
Criterion: accuracy

Table View | Plot View

accuracy: 93.33%

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	10	0	0	100.00%
pred. Iris-versicolor	0	10	2	83.33%
pred. Iris-virginica	0	0	8	100.00%
class recall	100.00%	100.00%	80.00%	

- KNN classification


Result History
ExampleSet (Apply Model)
PerformanceVector (Performance)
KNNClassification (k-NN)
Description
Annotations
KNNClassification
Weighted 40-Nearest Neighbour model for classification.
The model contains 120 examples with 4 dimensions of the following classes:
Iris-setosa
Iris-versicolor
Iris-virginica