# Data Warehousing and Data Mining

## Tutorial 2


### Student Details

Name : Krunal Rank
Adm. No. : U18CO081


### What are the needs of pre-processing of data? Also explain them with proper examples.

Data Preprocessing is a technique that allows us to convert raw data into much cleaner information that is suitable for analysis and deducing conclusions.
It is a preliminary step that requires information to be organised, ordered and merged.
Some of the reasons why Data Preprocessing is required are as follows:

- Filling up missing values
- Get an insight on the summary of data and identify outliers that can hinder the progress of analysis in a long run
- Aggregating the raw data that can be successfully utilised in a machine learning pipeline

For example, let's say we have a certain data of the employees with some of the records missing age and gender.
Now, if we organise the data a little bit and analyse it thoroughly, we can eventually predict the age and gender values which are missing using the relevant nearby information (an idea that led to the KNN technique).

After fixing such issues, plotting the histograms of the data allows us to predict certain values that are quite odd and out of the blue, which can be detected as outliers and requires to be removed from the dataset.

Apart from that, we can also validate data into a certain range of values (normalization) so that the values become comparable.

### List the factors comprising data quality.

The factors that allow us to analyse data and its quality are as follows:

- Accuracy : This determines how accurate and validated the data is.
- Completeness : This determines how comprehensive the data is.
- Reliability : This determines the trustworthiness of the data.
- Relevance : This determines whether the data is actually necessary.
- Timeliness : This determines whether the data is up-to-date or not.

## What are the major tasks in Data Pre-processing? Explain in detail along with proper examples.

There are three steps in data preprocessing:

- Data Cleaning
    - This involves filling up or predicting missing values. For example, let's say a given dataset has some missing values. Then there are a few possible options:
        - Fill the data using mean, median or mode (If the data is continuous, then if it follows normal distribution, use mean, if it is skewed, use median or if the data is categorical, use mode)
        - Remove those records all together
        - For Categorical missing values, create a new category all together that symbolises these missing values.
    - Outlier Detection and Removal. For example, a given dataset may have some values that are quite odd and don't follow the usual characteristics set up by the rest of the data. These values need to be identified and removed. There are few options for this as well:
        - Create histograms and bin the data accordingly to identify the odd ones out. Histogram Based Outlier Score can also be used.
        - Regression techniques can be used to make the data more smooth by replacing values using boundary values of the bin.
        - Clustering algorithms such as K Means Clustering can be used to group data and find the outliers.
- Data Transformation
    - Normalisation allows us to map the data from a given range to our own required range by using the Mean and Standard Deviation of the data (Z Score Normalisation). This allows us to make columns comparable.
    - Attribute Selection allows us to identify which columns in the data are actually relevant. Also, it helps us remove columns that are closely related.
    - Discretization allows us to replace raw values of numeric attribute by interval levels
    - Concept Hierarchy Generation can be used to convert lower level column values to higher levels. For example, a city can be replaced with a country to group data with common countries.
- Data Reduction
    - Data Cube Aggregation can be used to create data cubes to analyse multidimensional data.
    - Attribute Subset Selection can be used with the help of p-values and correlation techniques to identify which attributes are required.
    - Numerosity Reduction is a further reduction in which we only store the model created from the data instead of the complete data. The model is usually a machine learning based model such as a regression or a logistic model.
    - Dimensionality Reduction can be used to reduce the size of the data with the help of techniques such as Wavelet transforms or Principal Component Analysis.

## Explain data cleaning method of filling in missing values.

- This involves filling up or predicting missing values. For example, let's say a given dataset has some missing values. Then there are a few possible options:

    - Fill the data using mean, median or mode (If the data is continuous, then if it follows normal distribution, use mean, if it is skewed, use median or if the data is categorical, use mode)
    - Remove those records all together
    - For Categorical missing values, create a new category all together that symbolises these missing values.

## Explain measures of central tendency with examples.

- A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data.
- As such, measures of central tendency are sometimes called measures of central location.
- They are also classed as summary statistics.
- Mean, median and Mode are the measures of central tendency but under various conditions, one of them is more appropriate than the rest.
- For data with values approximately following normal distribution, mean is an appropriate measure of central tendency.
- For data with highly skewed values (left skewed or right skewed), median is an appropriate measure of central tendency.
- For categorical data, mode is an appropriate measure of central tendency.

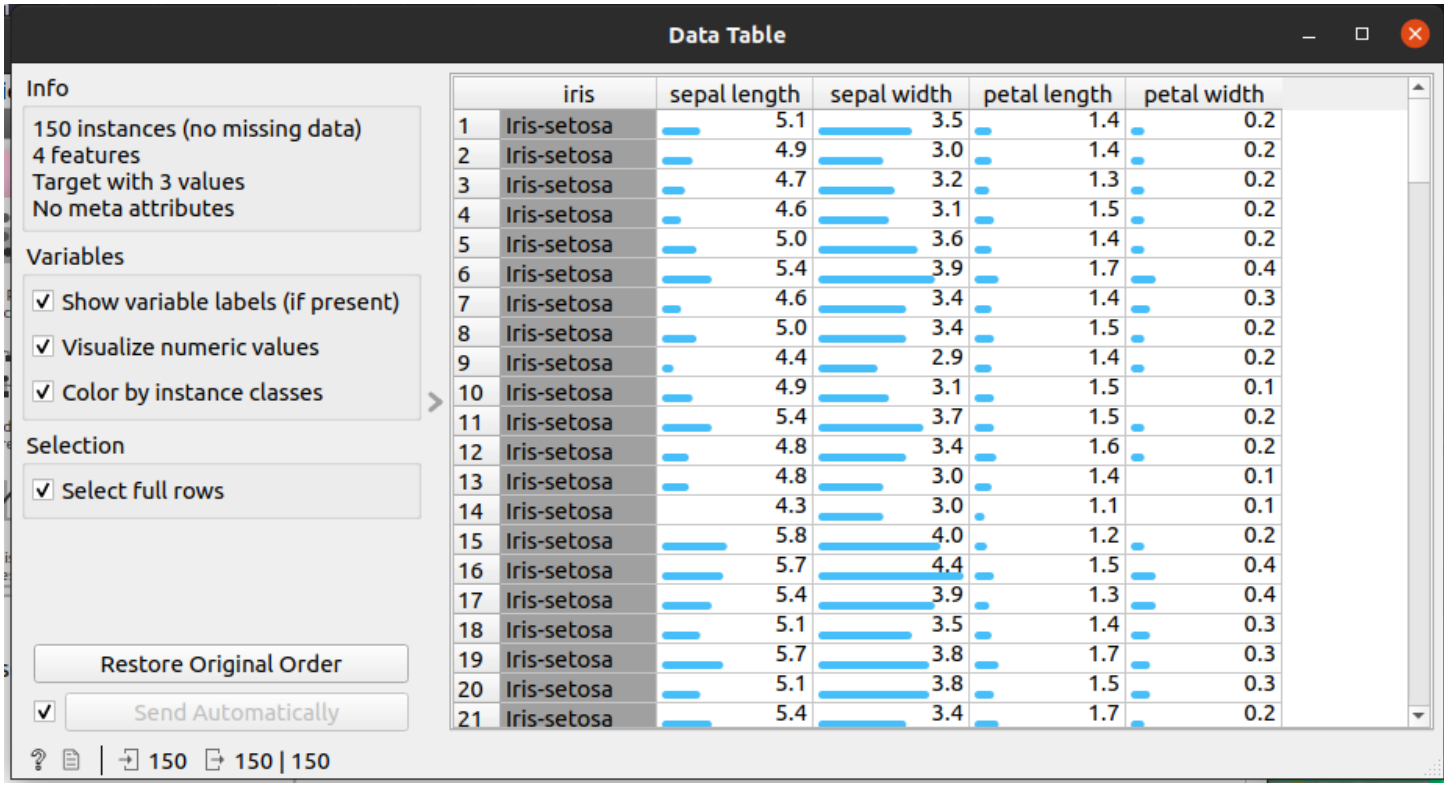
For example, let's say we have values : 10, 15 , 25, 30, 30 , 40, 50.

Then,
Mean = Equally weighted average = (10 + 15 + 25 + 30 + 30 + 40 + 50)/7 = 28.57
Median = Central value of the sorted data = $x_4$ = 30
Mode = Value which appeared most = 30

## Analyse Pre-Processing techniques which were discussed in class using the Orange Data Mining framework (Any Dataset can be utilized).

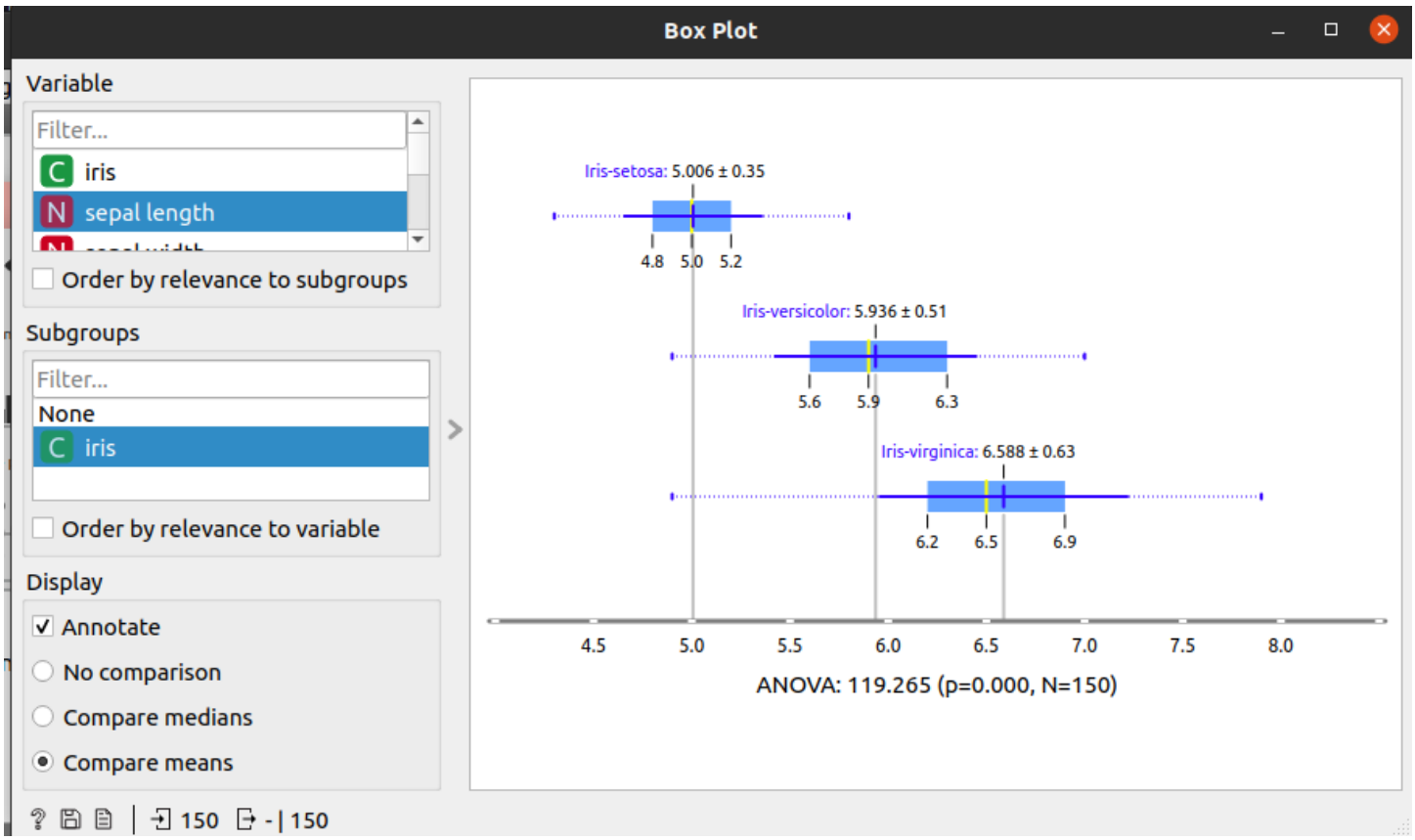
Below are the screenshots that show the analysis of the very famous Iris dataset that has 4 numeric attributes (sepal length, sepal width, petal length and petal width) based on which the target (iris type) is predicted.

Data Values

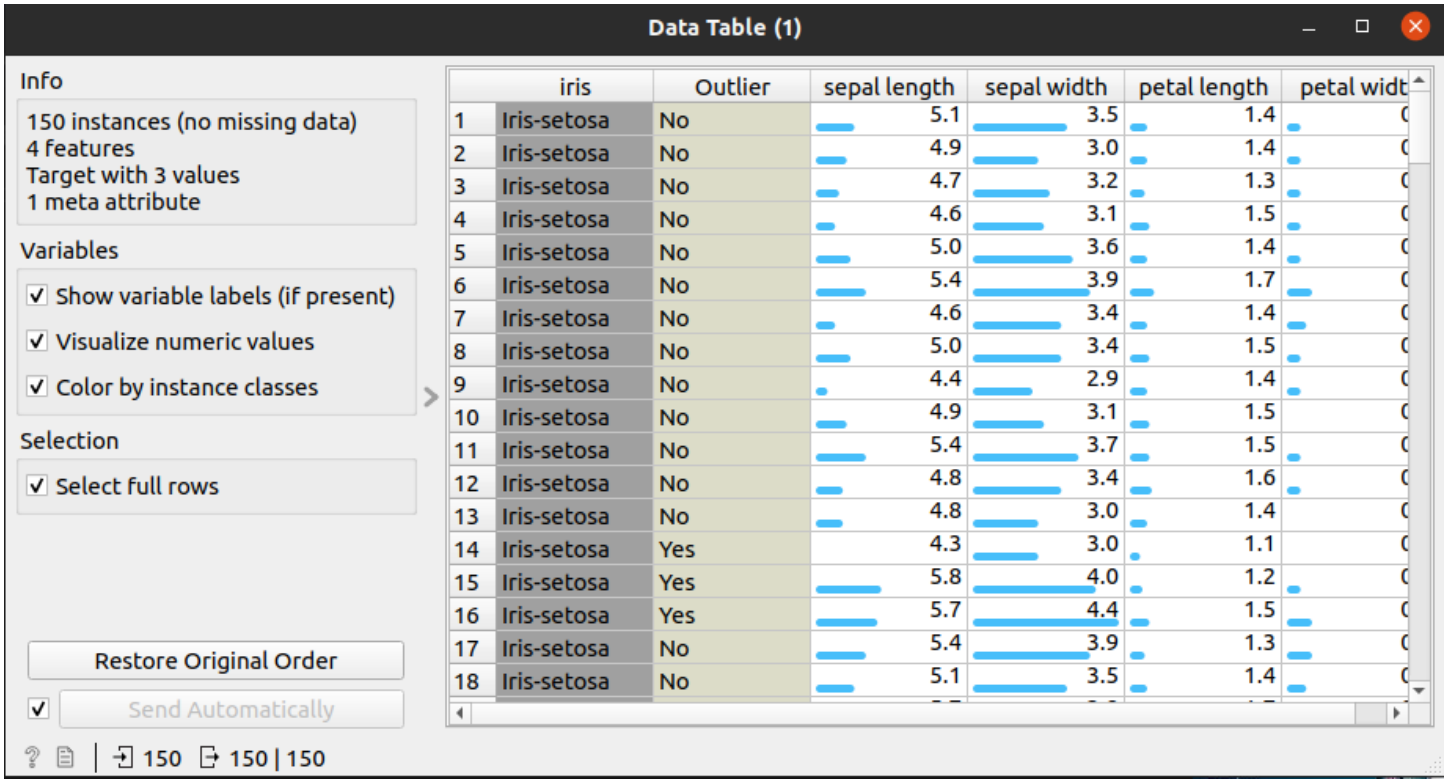



| | iris | sepal length | sepal width | petal length | petal width |
|---|---|---|---|---|---|
| 1 | Iris-setosa | 5.1 | 3.5 | 1.4 | 0.2 |
| 2 | Iris-setosa | 4.9 | 3.0 | 1.4 | 0.2 |
| 3 | Iris-setosa | 4.7 | 3.2 | 1.3 | 0.2 |
| 4 | Iris-setosa | 4.6 | 3.1 | 1.5 | 0.2 |
| 5 | Iris-setosa | 5.0 | 3.6 | 1.4 | 0.2 |
| 6 | Iris-setosa | 5.4 | 3.9 | 1.7 | 0.4 |
| 7 | Iris-setosa | 4.6 | 3.4 | 1.4 | 0.3 |
| 8 | Iris-setosa | 5.0 | 3.4 | 1.5 | 0.2 |
| 9 | Iris-setosa | 4.4 | 2.9 | 1.4 | 0.2 |
| 10 | Iris-setosa | 4.9 | 3.1 | 1.5 | 0.1 |
| 11 | Iris-setosa | 5.4 | 3.7 | 1.5 | 0.2 |
| 12 | Iris-setosa | 4.8 | 3.4 | 1.6 | 0.2 |
| 13 | Iris-setosa | 4.8 | 3.0 | 1.4 | 0.1 |
| 14 | Iris-setosa | 4.3 | 3.0 | 1.1 | 0.1 |
| 15 | Iris-setosa | 5.8 | 4.0 | 1.2 | 0.2 |
| 16 | Iris-setosa | 5.7 | 4.4 | 1.5 | 0.4 |
| 17 | Iris-setosa | 5.4 | 3.9 | 1.3 | 0.4 |
| 18 | Iris-setosa | 5.1 | 3.5 | 1.4 | 0.3 |
| 19 | Iris-setosa | 5.7 | 3.8 | 1.7 | 0.3 |
| 20 | Iris-setosa | 5.1 | 3.8 | 1.5 | 0.3 |
| 21 | Iris-setosa | 5.4 | 3.4 | 1.7 | 0.2 |

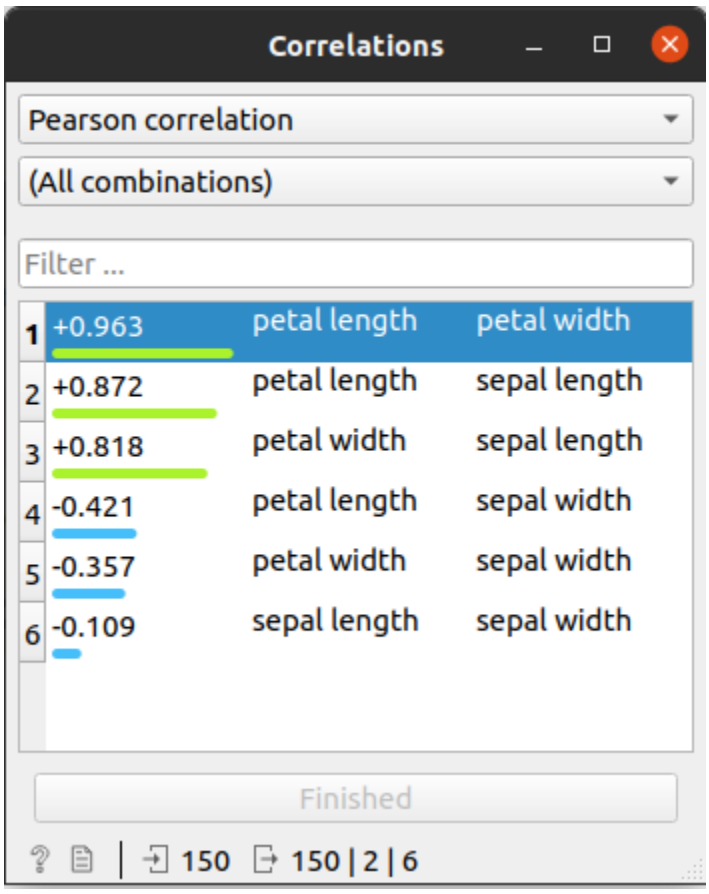
Box Plot of sepal length



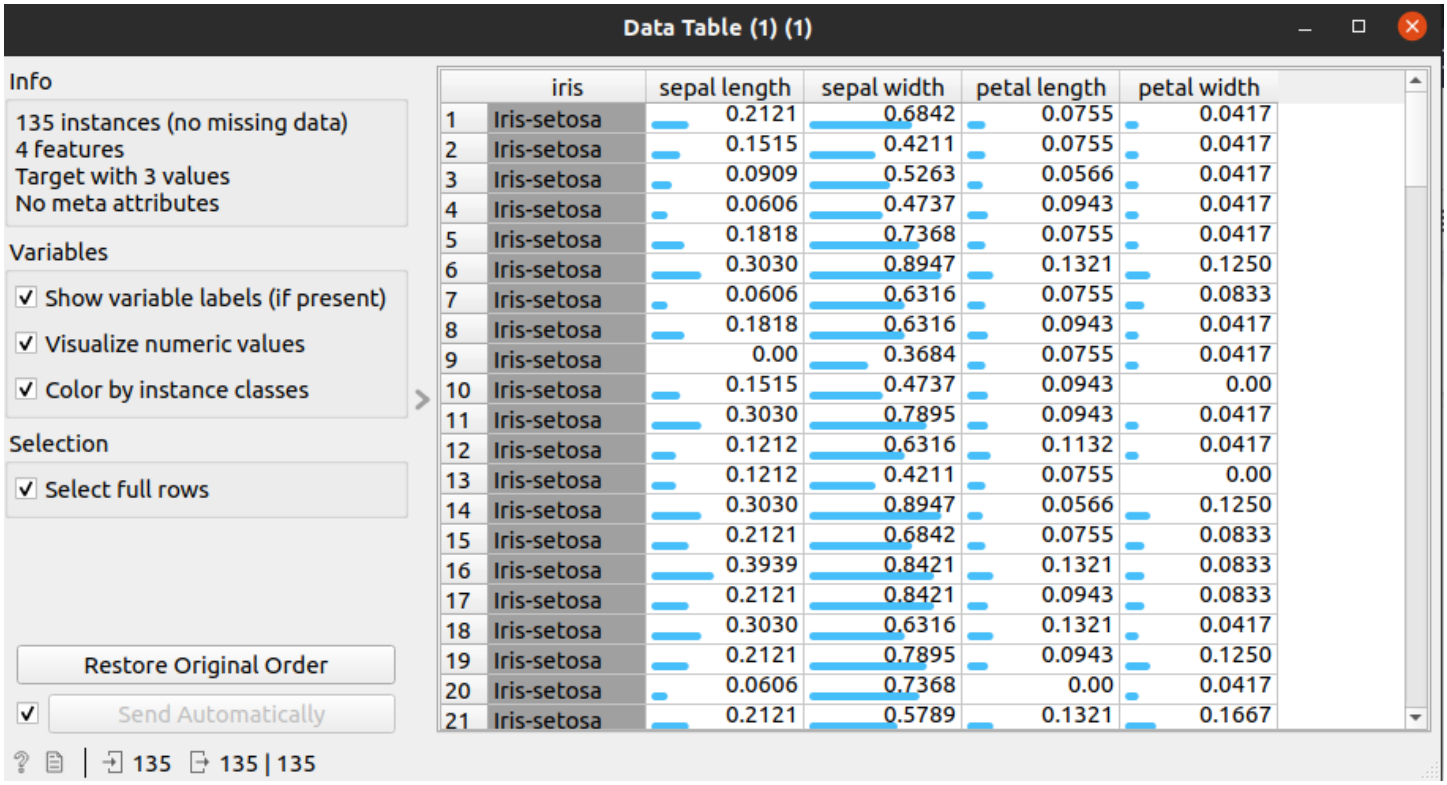
Outlier Detection




Pairwise Attribute Correlation

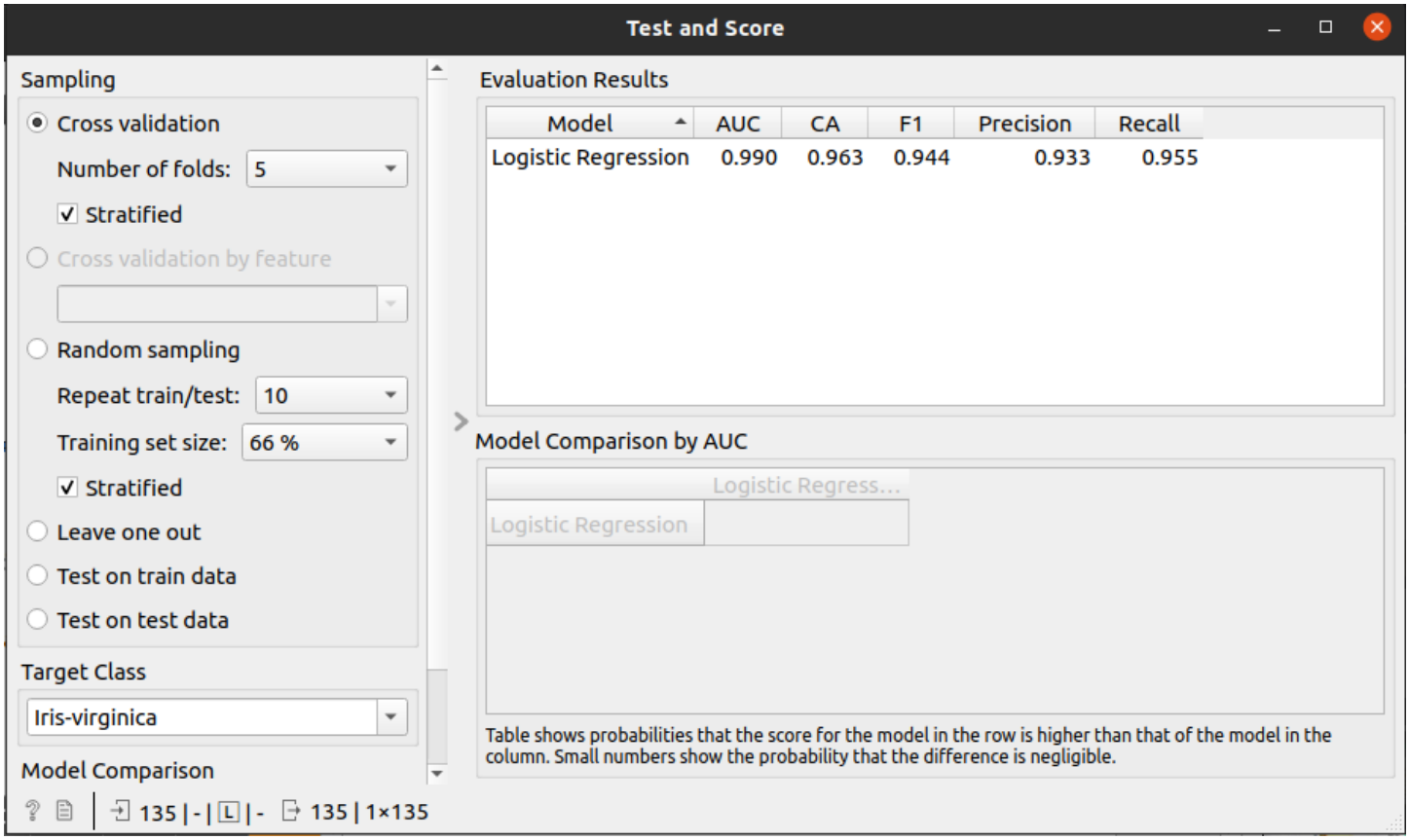


Normalised Data to [0,1]




| | iris | sepal length | sepal width | petal length | petal width |
|---|---|---|---|---|---|
| 1 | Iris-setosa | 0.2121 | 0.6842 | 0.0755 | 0.0417 |
| 2 | Iris-setosa | 0.1515 | 0.4211 | 0.0755 | 0.0417 |
| 3 | Iris-setosa | 0.0909 | 0.5263 | 0.0566 | 0.0417 |
| 4 | Iris-setosa | 0.0606 | 0.4737 | 0.0943 | 0.0417 |
| 5 | Iris-setosa | 0.1818 | 0.7368 | 0.0755 | 0.0417 |
| 6 | Iris-setosa | 0.3030 | 0.8947 | 0.1321 | 0.1250 |
| 7 | Iris-setosa | 0.0606 | 0.6316 | 0.0755 | 0.0833 |
| 8 | Iris-setosa | 0.1818 | 0.6316 | 0.0943 | 0.0417 |
| 9 | Iris-setosa | 0.00 | 0.3684 | 0.0755 | 0.0417 |
| 10 | Iris-setosa | 0.1515 | 0.4737 | 0.0943 | 0.00 |
| 11 | Iris-setosa | 0.3030 | 0.7895 | 0.0943 | 0.0417 |
| 12 | Iris-setosa | 0.1212 | 0.6316 | 0.1132 | 0.0417 |
| 13 | Iris-setosa | 0.1212 | 0.4211 | 0.0755 | 0.00 |
| 14 | Iris-setosa | 0.3030 | 0.8947 | 0.0566 | 0.1250 |
| 15 | Iris-setosa | 0.2121 | 0.6842 | 0.0755 | 0.0833 |
| 16 | Iris-setosa | 0.3939 | 0.8421 | 0.1321 | 0.0833 |
| 17 | Iris-setosa | 0.2121 | 0.8421 | 0.0943 | 0.0833 |
| 18 | Iris-setosa | 0.3030 | 0.6316 | 0.1321 | 0.0417 |
| 19 | Iris-setosa | 0.2121 | 0.7895 | 0.0943 | 0.1250 |
| 20 | Iris-setosa | 0.0606 | 0.7368 | 0.00 | 0.0417 |
| 21 | Iris-setosa | 0.2121 | 0.5789 | 0.1321 | 0.1667 |

Using Logistic Regression :



Confusion Matrix: