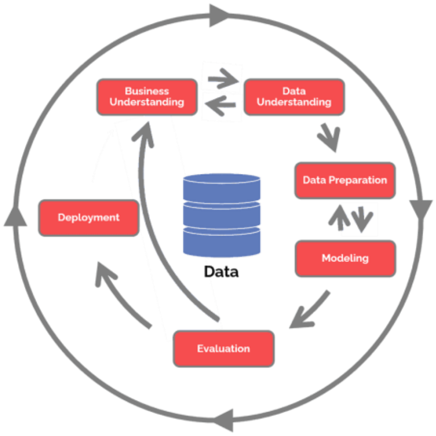
Titanic Dataset (CRISP-DM)



Business Understanding: Business Understanding is the first step in the process. In this step, we figure out what problem or question we want to solve or answer.

Data Understanding: The second step concerns the data used to resolve the business problem. This data must be collected, and its significance must be comprehended. For instance, if the data is stored in a table, it is crucial to understand the precise meaning of each column. The data comprehension may necessitate a return to the first step, i.e. business comprehension.

Data Preparation: In this step, the data get prepared for the next step, i.e. the modelling. Preparing the data includes, for example, a strategy to deal with missing values. In general, the data could be cleaner than it is. But since data of low quality can lead to wrong answers to business questions, the data preparation step also includes data cleaning.

Modelling: This step is one of the core steps where one tries to find a model which describes the data. In terms of our business problem of predicting the price of a car based on its features, modelling means finding a function that maps the features of a car to its price. It could be possible that the modelling step requires us to turn back to the data preparation step (for example, if we realise in the modelling step that the data still needs to be sufficiently cleaned).

Evaluation: Once the model is constructed, its performance must be evaluated. This occurs during the evaluation phase. For instance, we could predict the prices of automobiles based on a model for which we know the actual prices and then compare how closely the predicted values match the actual values.

Deployment: In the last step, if we are satisfied with the model's performance, we need to productionize it. For the car price prediction example, that could mean that we build an application that allows entering certain car features in a form and returns the price prediction to the user.

CRISP-DM Example Walkthrough for The Titanic Dataset

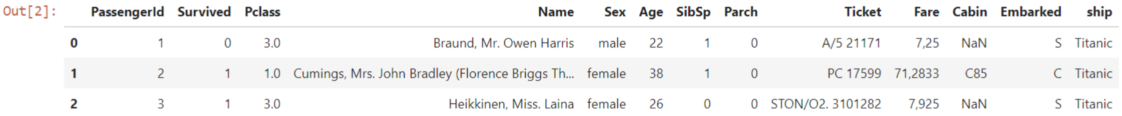
Business Problem

As mentioned above, the CRISP-DM process starts with understanding the business problem. There are some business questions regarding this dataset:

- What factors are crucial to surviving?
- Who has a higher chance to survive: Jack, as a young male, travelling alone as the passenger of 3rd class, or Rose, as a young female, travelling with family and passenger of 1st class?

Data Understanding and Data Preparation

The data is taken from the module. Below is data looks like:



Out[2]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	ship
0	1	0	3.0	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7,25	NaN	S	Titanic
1	2	1	1.0	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38	1	0	PC 17599	71,2833	C85	C	Titanic
2	3	1	3.0	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7,925	NaN	S	Titanic

The data above contains several problems, such as missing values, incorrect labelling data, outlier detection, and data duplication. This process is described in detail in the Jupyter logbook file. The resulting output is a new clean dataset in csv form. [Link to Data Cleaning]

Modelling

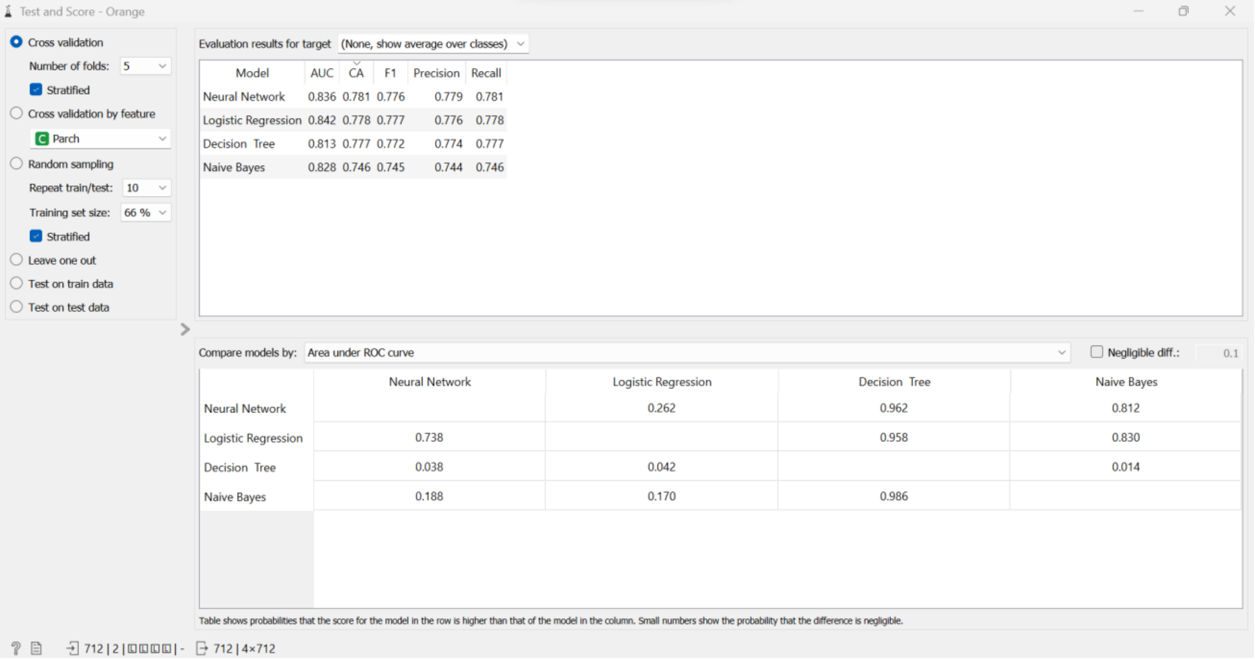
To generate a model to predict who has a high chance of survival and the determining factors in this case, we used several algorithms such as Neural Network, Logistic Regression, Naive Bayes, and Decision Tree.

- **Neural networks** classify and cluster. These add clustering and classification to our data. This classifies data when trained on a labelled dataset and groups unlabeled data by example input similarities. Deep neural networks can extract features fed to other algorithms for clustering and classification to be used as components of larger machine-learning applications involving reinforcement learning, type, and regression.
- **Logistic Regression** is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous, i.e., there could only be two possible classes (e.g., either the cancer is malignant or not). As a result, this technique is used while dealing with binary data.
- **The Naive Bayes** is fast and easy to implement, but their most significant disadvantage is the independent predictor requirement. In most real-life cases, the predictors are dependent, hindering the classifier’s performance.
- **Decision Trees** (DTs) is a non-parametric supervised classification and regression learning method. The objective is to develop a model capable of predicting the value of a target variable using simple decision rules inferred from the data features. A tree can be viewed as a piecewise constant approximation.

Evaluation

To evaluate the performance of these models, we use classification accuracy. Classification accuracy is a metric that summarises a classification model's performance by dividing the number of correct predictions by the total number of predictions. It is simple to calculate and intuitive to comprehend, making it the most popular metric for classifier model evaluation. This intuition only works when the distribution of examples among classes is grossly asymmetrical.

Below are the results:



Test and Score - Orange

● Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

Parch

○ Random sampling

Repeat train/test: 10

Training set size: 66 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
Neural Network	0.836	0.781	0.776	0.779	0.781
Logistic Regression	0.842	0.778	0.777	0.776	0.778
Decision Tree	0.813	0.777	0.772	0.774	0.777
Naive Bayes	0.828	0.746	0.745	0.744	0.746

Compare models by: Area under ROC curve

□ Negligible diff.: 0.1

	Neural Network	Logistic Regression	Decision Tree	Naive Bayes
Neural Network		0.262	0.962	0.812
Logistic Regression	0.738		0.958	0.830
Decision Tree	0.038	0.042		0.014
Naive Bayes	0.188	0.170	0.986	

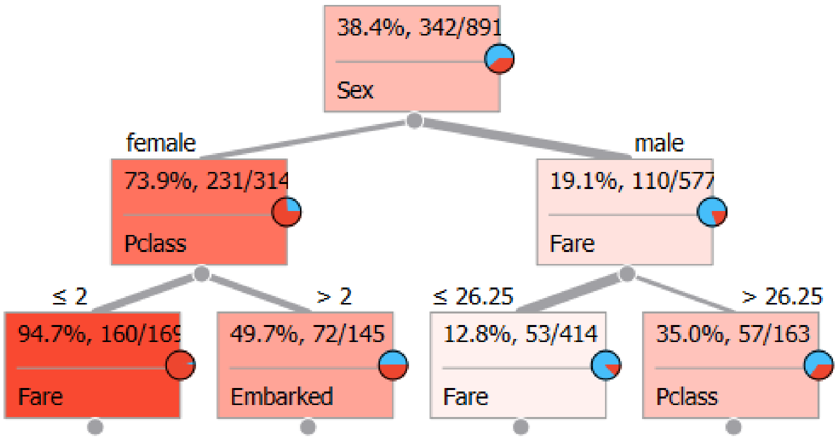
Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

712 | 2 | - 712 | 4×712

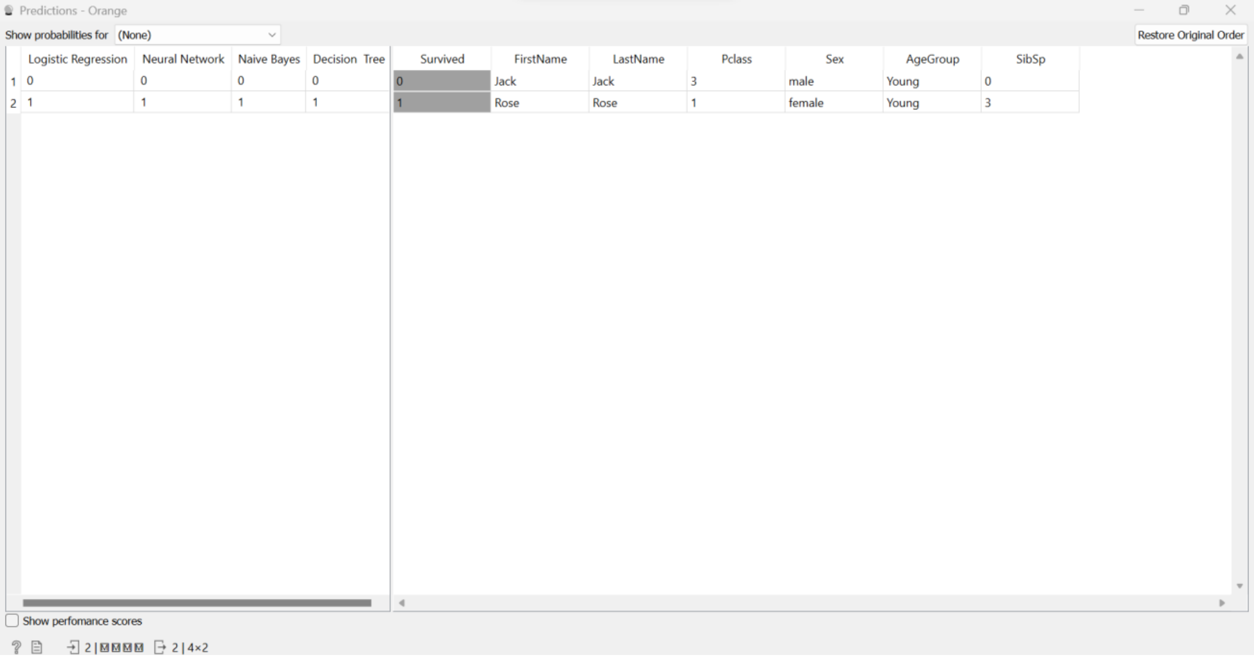
Regarding the results, Neural Network has a higher classification accuracy of 78.1%.

The answer to the Business Model:

1. The highest factors that affect survival rates are gender. Women have a higher chance of survival than men. 73.9% of women survive, and men only 19.1%. The second factor is the class of passengers; 94.7% of women in the first and second classes stayed than those in the third. For men, the fare has the most significant impact after gender, there 35% of the men passengers survived if they paid > 26.25.



2. According to the model created by the author, Rose is predicted to survive in four algorithms.



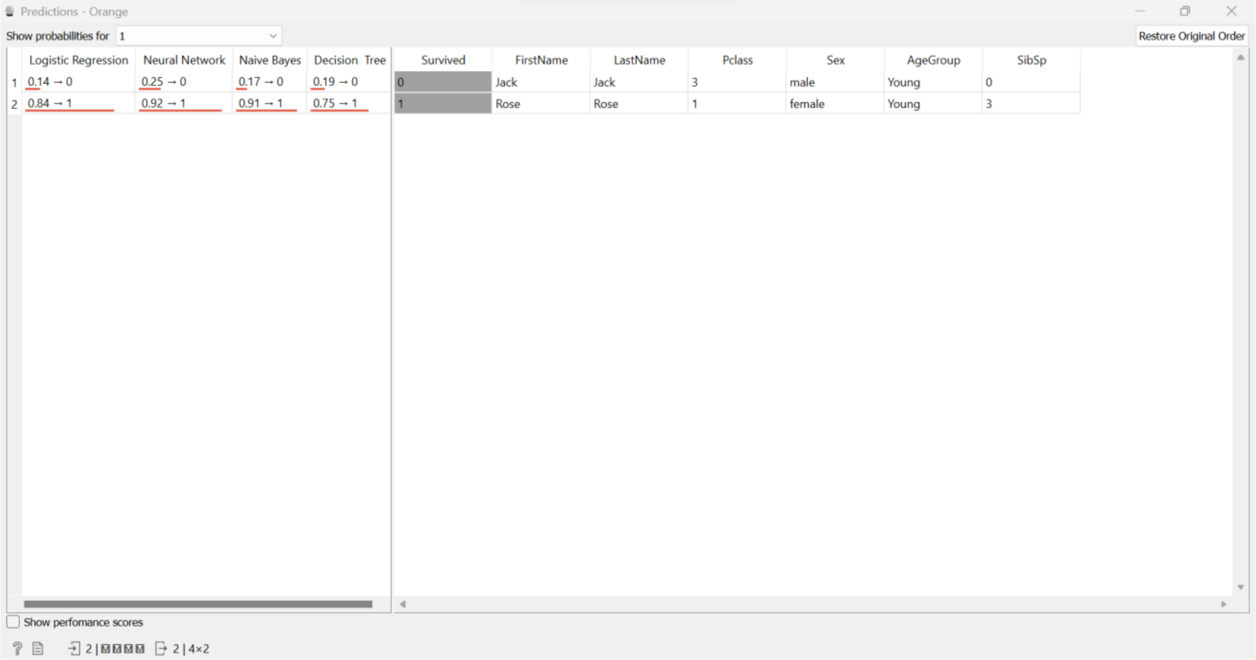
Show probabilities for (None)

	Logistic Regression	Neural Network	Naive Bayes	Decision Tree	Survived	FirstName	LastName	Pclass	Sex	AgeGroup	SibSp
1	0	0	0	0	0	Jack	Jack	3	male	Young	0
2	1	1	1	1	1	Rose	Rose	1	female	Young	3

□ Show perfomance scores

2 | 2 | 4×2

The following picture shows the survival percentage for two passengers (Rose and Jack).



Show probabilities for 1

	Logistic Regression	Neural Network	Naive Bayes	Decision Tree	Survived	FirstName	LastName	Pclass	Sex	AgeGroup	SibSp
1	0.14 → 0	0.25 → 0	0.17 → 0	0.19 → 0	0	Jack	Jack	3	male	Young	0
2	0.84 → 1	0.92 → 1	0.91 → 1	0.75 → 1	1	Rose	Rose	1	female	Young	3

The above picture illustrates that Rose has a bigger chance of surviving than Jack, with 84% for Logistic Regression, 92% for Neural Network, 91% for Naïve Bayes, and 75% for Decision tree.

Conclusion

To summarise, the factors that can improve survival in this problem are gender, passenger class, and fare. The female gender has a greater chance of survival than the male, and passengers in 1st and 2nd classes have a greater chance of survival than passengers in 3rd class. Thus, Jane is predicted to survive in all four methods used. Of all the methods, the method with the highest accuracy is the Neural network, with a percentage of 78.1%, and Naive bayes is in the bottom position with an accuracy rate of 74.6%.

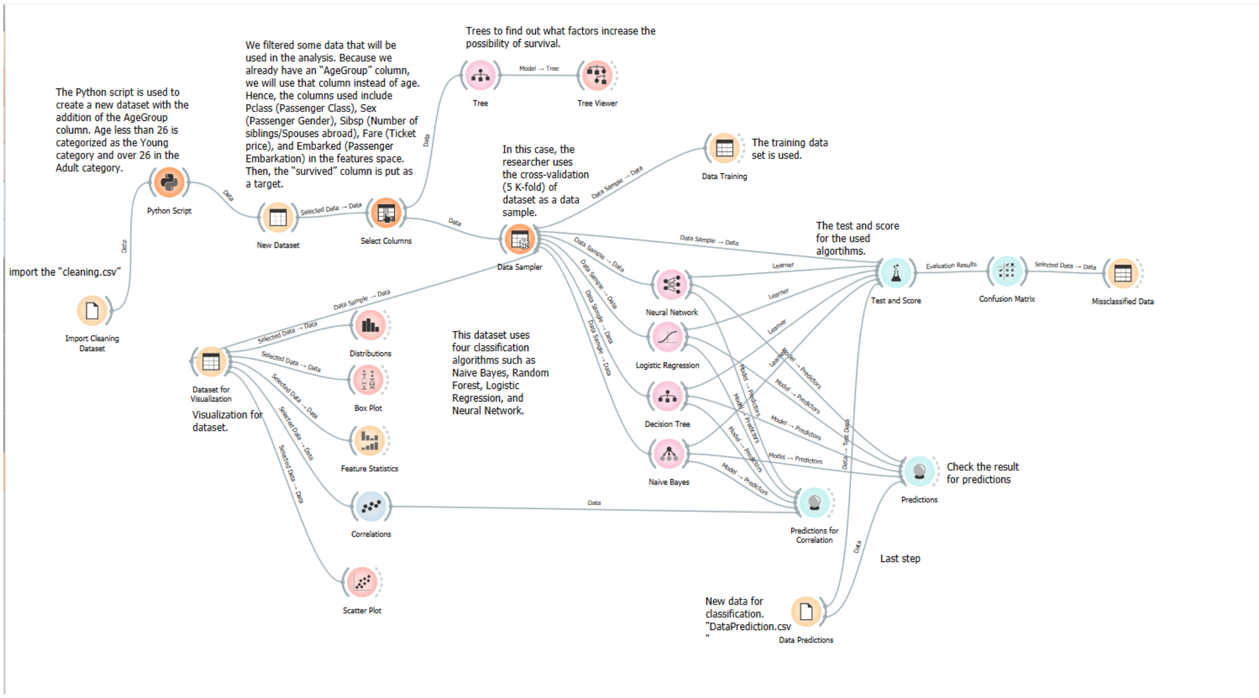
TUTORIAL

a. The cleaning Dataset – Titanic Dataset

In this step, we used Jupyter Notebook to deal with some issues. There are several problems found in this dataset, such as missing values, incorrect labelling data, outlier detection, and data duplication. This process is explained in detail in the Jupyter notebook file. The resulting output is a new dataset that is clean. [Link to Data Cleaning]

b. Preprocess data using Orange Data Mining Software

i. Start the orange software.



The Python script is used to create a new dataset with the addition of the AgeGroup column. Age less than 26 is categorized as the Young category and over 26 in the Adult category.

We filtered some data that will be used in the analysis. Because we already have an "AgeGroup" column, we will use that column instead of age. Hence, the columns used include Pclass (Passenger Class), Sex (Passenger Gender), Sibsp (Number of siblings/Spouses abroad), Fare (Ticket price), and Embarked (Passenger Embarkation) in the features space. Then, the "survived" column is put as a target.

Trees to find out what factors increase the possibility of survival.

In this case, the researcher uses the cross-validation (5 K-fold) of dataset as a data sample.

The training data set is used.

The test and score for the used algortihms.

import the "cleaning.csv"

This dataset uses four classification algorithms such as Naive Bayes, Random Forest, Logistic Regression, and Neural Network.

Visualization for dataset.

Check the result for predictions

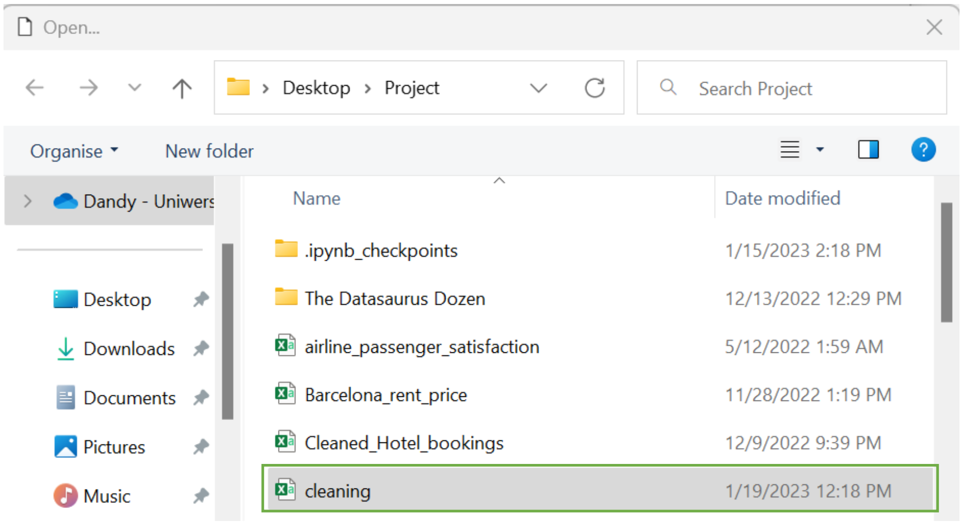
Last step

New data for classification. "DataPrediction.csv "

The picture above is the orange project for the titanic dataset. [Link to Download Orange Modelling]

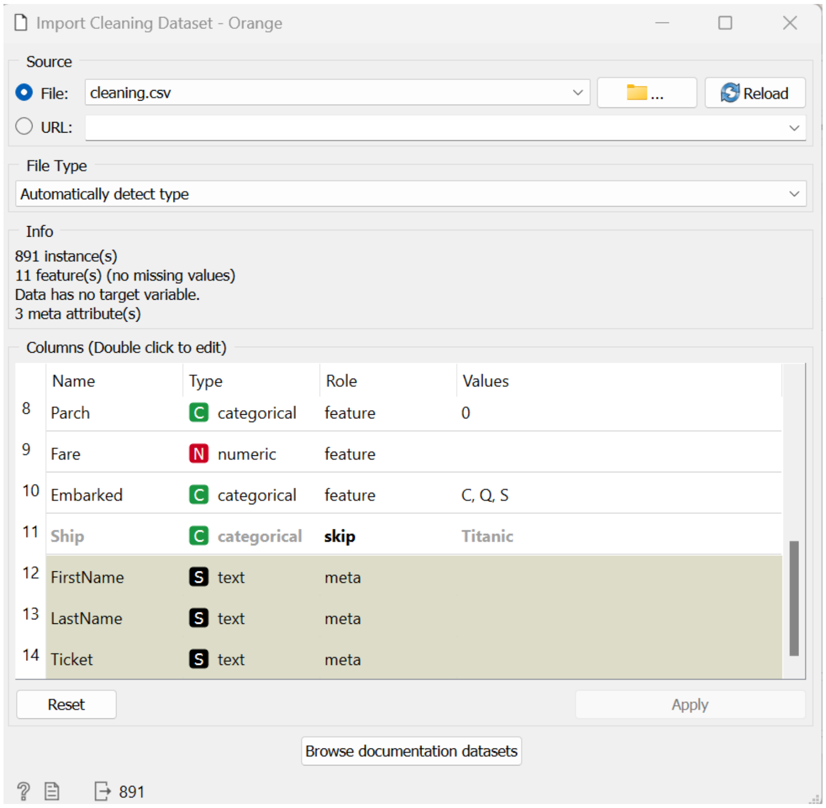
ii. Import Cleaning Dataset in the file

For the first step, we choose "import cleaning dataset." Then we import the "cleaning.csv" file.



Name	Date modified
.ipynb_checkpoints	1/15/2023 2:18 PM
The Datasaurus Dozen	12/13/2022 12:29 PM
airline_passenger_satisfaction	5/12/2022 1:59 AM
Barcelona_rent_price	11/28/2022 1:19 PM
Cleaned_Hotel_bookings	12/9/2022 9:39 PM
cleaning	1/19/2023 12:18 PM

For the next step, we close or skip it.


Import Cleaning Dataset - Orange
Source
Reload
Browse documentation datasets
Reset
Apply
891

● File: cleaning.csv

○ URL:

File Type

Automatically detect type

Info

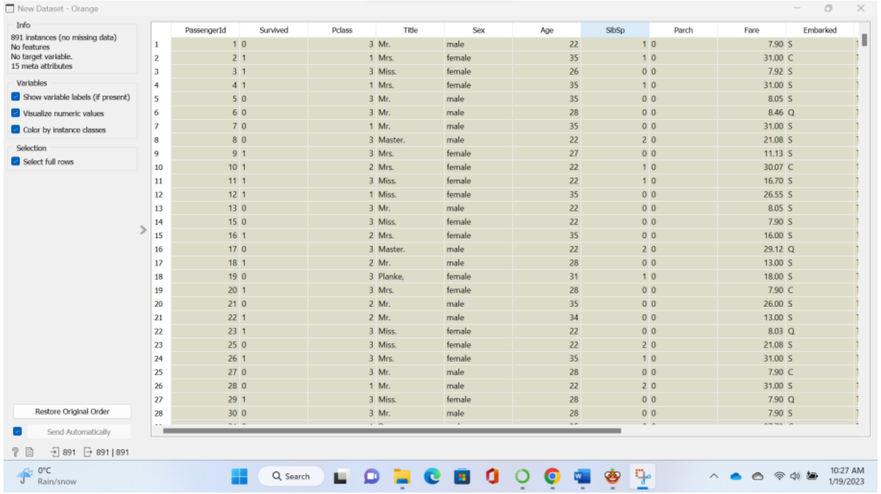
891 instance(s)
11 feature(s) (no missing values)
Data has no target variable.
3 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
8	Parch	C categorical	feature	0
9	Fare	N numeric	feature	
10	Embarked	C categorical	feature	C, Q, S
11	Ship	C categorical	skip	Titanic
12	FirstName	S text	meta	
13	LastName	S text	meta	
14	Ticket	S text	meta	

iii. Python Script and New Dataset

A Python script was used to create a new dataset with the addition of the AgeGroup column. Age less than 26 years old is categorised as young, and more than 26 years old in the adult category. We have a new dataset as below.



New Dataset - Orange

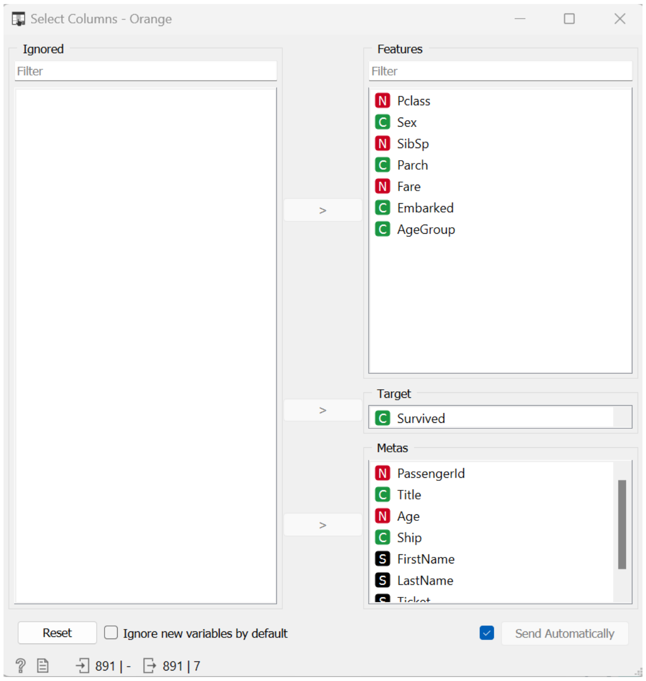
Info
891 instances (no missing data)
No features
No target variable.
15 meta attributes

Variables
☑ Show variable labels (if present)
☑ Visualize numeric values
☑ Color by instance classes

Selection
☑ Select full rows

	PassengerId	Survived	Pclass	Title	Sex	Age	SibSp	Parch	Fare	Embarked
1	1	0	3	Mr.	male	22	1	0	7.90	S
2	2	1	1	Mrs.	female	35	1	0	31.00	C
3	3	1	3	Miss.	female	26	0	0	7.92	S
4	4	1	1	Mrs.	female	35	1	0	31.00	S
5	5	0	3	Mr.	male	35	0	0	8.05	S
6	6	0	3	Mr.	male	28	0	0	8.46	Q
7	7	0	1	Mr.	male	35	0	0	31.00	S
8	8	0	3	Master.	male	22	2	0	21.08	S
9	9	1	3	Mrs.	female	27	0	0	11.13	S
10	10	1	2	Mrs.	female	22	1	0	30.07	C
11	11	1	3	Miss.	female	22	1	0	16.70	S
12	12	1	1	Miss.	female	35	0	0	26.55	S
13	13	0	3	Mr.	male	22	0	0	8.05	S
14	15	0	3	Miss.	female	22	0	0	7.90	S
15	16	1	2	Mrs.	female	35	0	0	16.00	S
16	17	0	3	Master.	male	22	2	0	29.12	Q
17	18	1	2	Mr.	male	28	0	0	13.00	S
18	19	0	3	Planke.	female	31	1	0	18.00	S
19	20	1	3	Mrs.	female	28	0	0	7.90	C
20	21	0	2	Mr.	male	35	0	0	26.00	S
21	22	1	2	Mr.	male	34	0	0	13.00	S
22	23	1	3	Miss.	female	22	0	0	8.03	Q
23	25	0	3	Miss.	female	22	2	0	21.08	S
24	26	1	3	Mrs.	female	35	1	0	31.00	S
25	27	0	3	Mr.	male	28	0	0	7.90	C
26	28	0	1	Mr.	male	22	2	0	31.00	S
27	29	1	3	Miss.	female	28	0	0	7.90	Q
28	30	0	3	Mr.	male	28	0	0	7.90	S

There are 891 instances without missing values and a new “Age group” column. Then we filter a new data set as below.



Select Columns - Orange

Ignored

Features
- N Pclass
- C Sex
- N SibSp
- C Parch
- N Fare
- C Embarked
- C AgeGroup

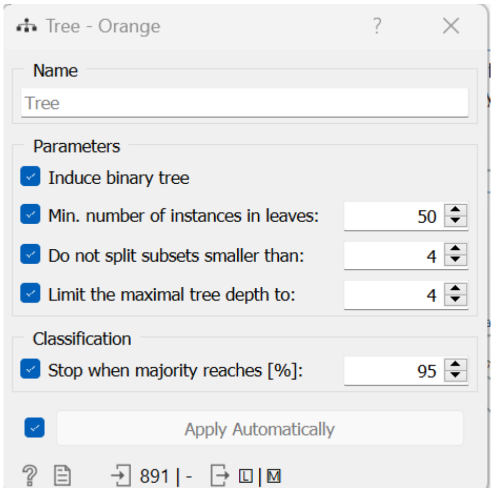
Target
- C Survived

Metas
- N PassengerId
- C Title
- N Age
- C Ship
- S FirstName
- S LastName
- S Ticket

We filtered some data that will be used in the analysis. Because we already have an “AgeGroup” column, we will use that column instead of age. Hence, the columns used include Pclass (Passenger Class), Sex (Passenger Gender), Sibsp (Number of siblings/Spouses abroad), Fare (Ticket price), and Embarked (Passenger Embarkation) in the features space. Then, the “survived” column is put as a target.

iv. Tree and Tree Viewer

In this step, we can know what factors affect the possibility of passengers surviving. We set the limit maximal for the tree as four, as below.



Name

Tree

Parameters

☑ Induce binary tree

☑ Min. number of instances in leaves: 50

☑ Do not split subsets smaller than: 4

☑ Limit the maximal tree depth to: 4

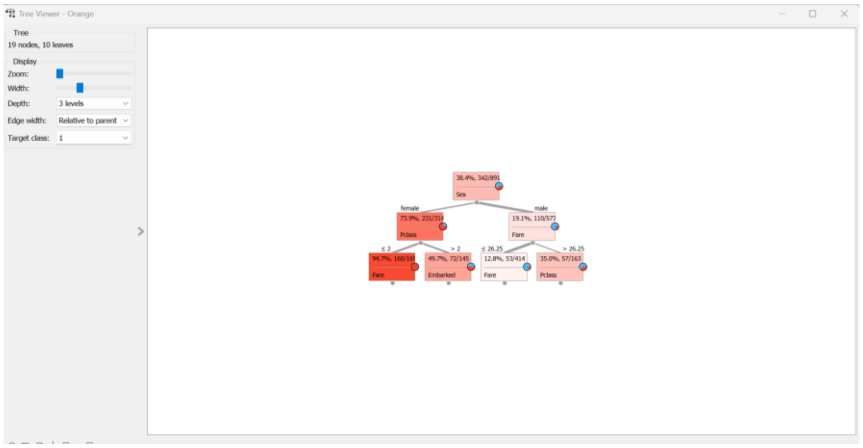
Classification

☑ Stop when majority reaches [%]: 95

☑ Apply Automatically

891 | -

After setting the tree, we can know the tree in Tree Viewer as below:



Tree

19 nodes, 10 leaves

Display

Zoom:

Width:

Depth: 3 levels

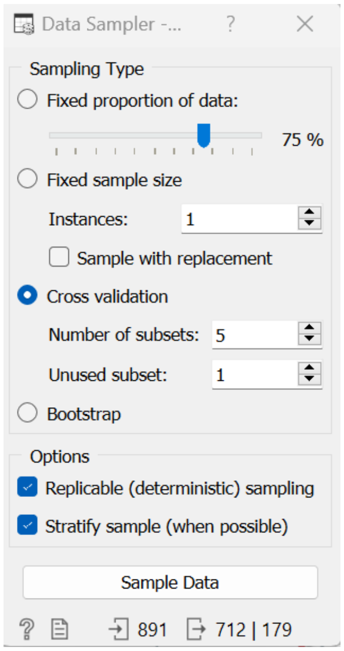
Edge width: Relative to parent

Target class: 1

The highest factors that affect survival rates are gender. Women have a higher chance of survival than men. 73.9% of women survive, and men only 19.1%. The second factor is class; 94.7% of women in the first and second classes stayed than those in the third.

v. Data Sampler

In this step, we choose a cross-validation sampling type and set your data sampler as below.



Data Sampler -... ? ×

Sampling Type

○ Fixed proportion of data:

75 %

○ Fixed sample size

Instances: 1

□ Sample with replacement

● Cross validation

Number of subsets: 5

Unused subset: 1

○ Bootstrap

Options

☑ Replicable (deterministic) sampling

☑ Stratify sample (when possible)

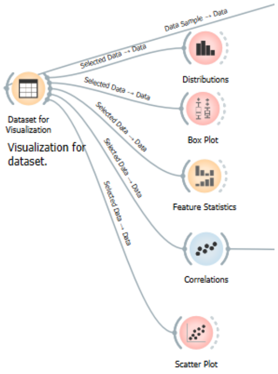
Sample Data

891 712 | 179

We usually use cross-validation to tune the hyperparameters of a given machine-learning algorithm to get good performance according to some suitable metric. Cross-validation, also known as rotation estimation or out-of-sample testing, is several similar model validation techniques used to evaluate how the results of a statistical analysis will generalise to an independent data set. Cross-validation is a resampling technique that employs distinct data subsets to test and train a model across multiple iterations. It is primarily employed in situations where prediction is objective, and one wishes to estimate how well a predictive model will perform in practice. In a typical prediction problem, a model is provided with a dataset of known data for training (training dataset) and a dataset of unknown data (or first-seen data) for testing (called the validation dataset or testing set).

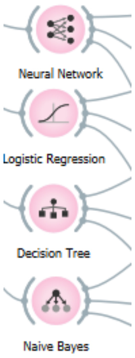
vi. Data Visualization

We can choose these visualisations for visualisation data, and the authors provide the visualisation in Jupyter Notebook. [Link to Data Visualization]



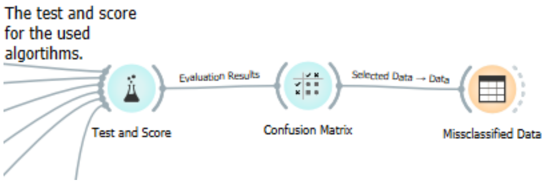
vii. **Algorithms**

In this step, the author provides four classification algorithms: the neural network, logistic regression, Decision Tree, and naïve bayes.



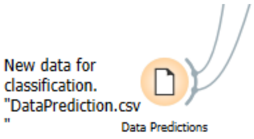
viii. **The Test, Result, and Misclassification Data.**

This step lets us know the test and score, confusion matrix, and Misclassification data.

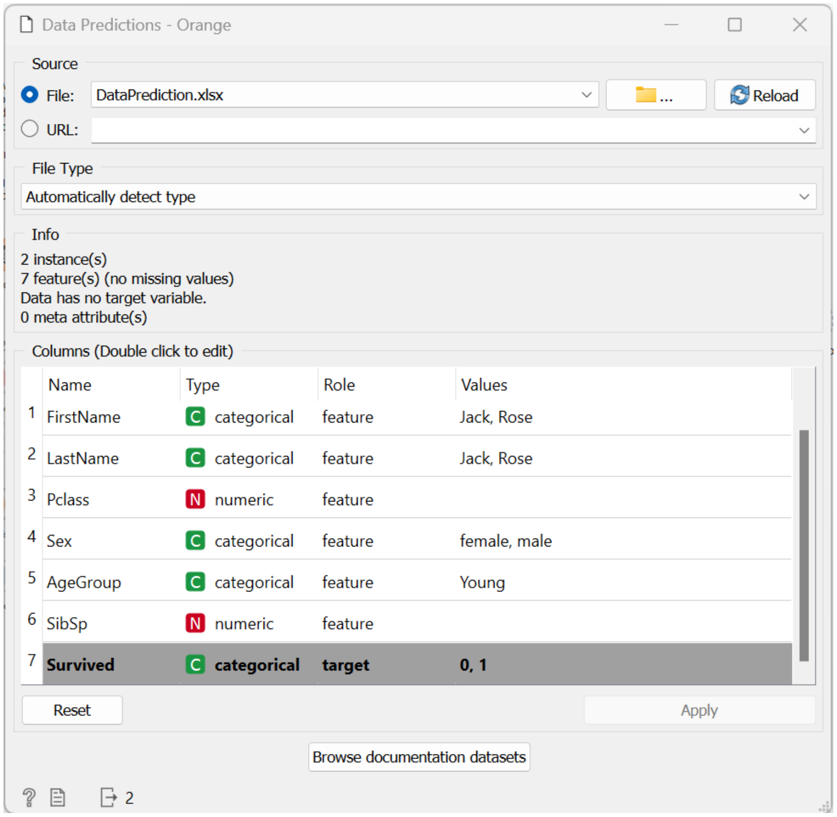


ix. **The Predictions**

We provide new data for classifying them. We import a new data set, “DataPrediction.csv,” into the first step’s data predictions.



In the next step, we set the dataprediction.csv like the following picture:



Data Predictions - Orange

Source

● File: DataPrediction.xlsx

○ URL:

File Type

Automatically detect type

Info

2 instance(s)
7 feature(s) (no missing values)
Data has no target variable.
0 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	FirstName	C categorical	feature	Jack, Rose
2	LastName	C categorical	feature	Jack, Rose
3	Pclass	N numeric	feature	
4	Sex	C categorical	feature	female, male
5	AgeGroup	C categorical	feature	Young
6	SibSp	N numeric	feature	
7	**Survived**	C **categorical**	**target**	**0, 1**

Then we check the results in the predictions.