Tutorial (Titanic Dataset)

Step 1: Define and understand the purpose – predict or classify? Step 2: Obtain data (may involve random sampling)

We know that we use the titanic dataset. In this case, we want to know the passengers who survived the titanic accident. Thus, the classification method is the best method for solving this problem.

Step 3: Explore, clean, and pre-process data.

a. The cleaning Dataset – Titanic Dataset

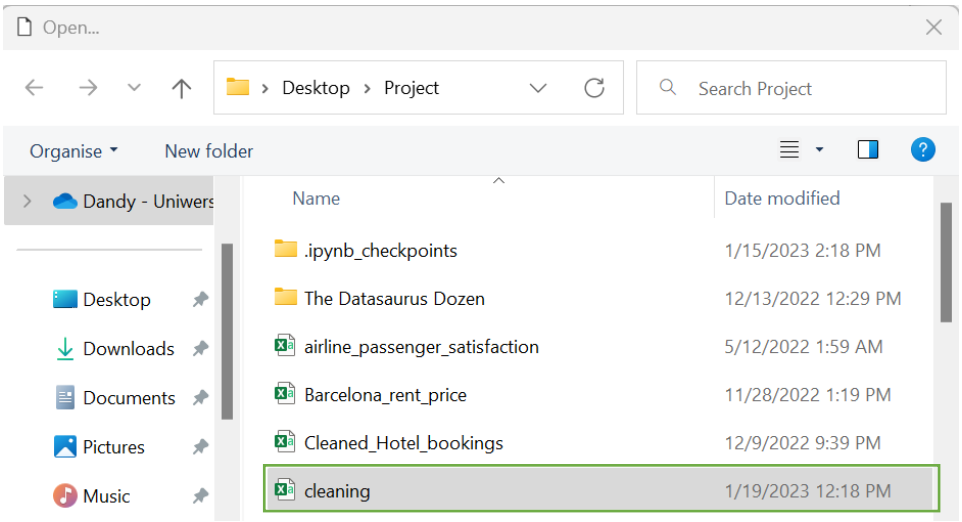
In this step, we used Jupyter Notebook to deal with some issues. There are several problems found in this dataset, such as missing values, incorrect labeling data, outlier detection, and data duplication. This process is explained in detail in the Jupyter notebook file. The resulting output is a new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

b. Preprocess data using Orange Data Mining Software

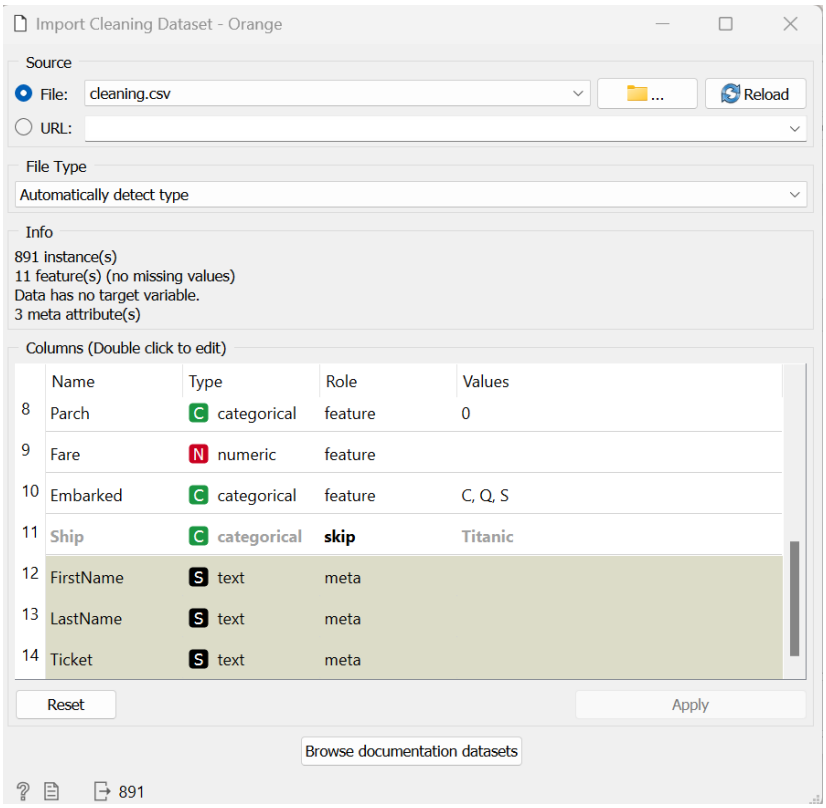
i. Start the orange software.

ii. Import Cleaning Dataset in the file

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

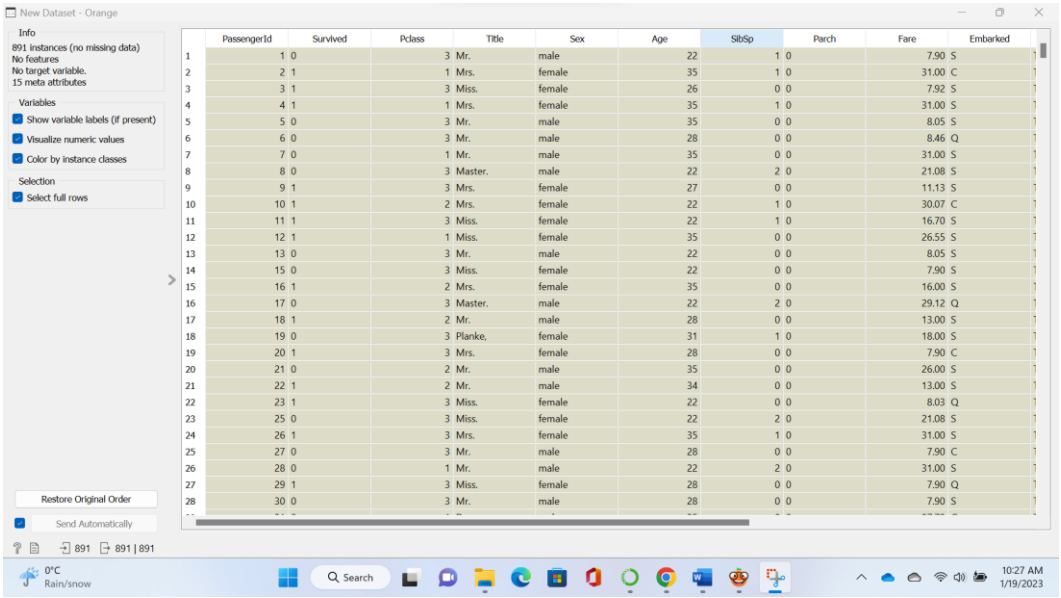


For the next step, we close or skip it.

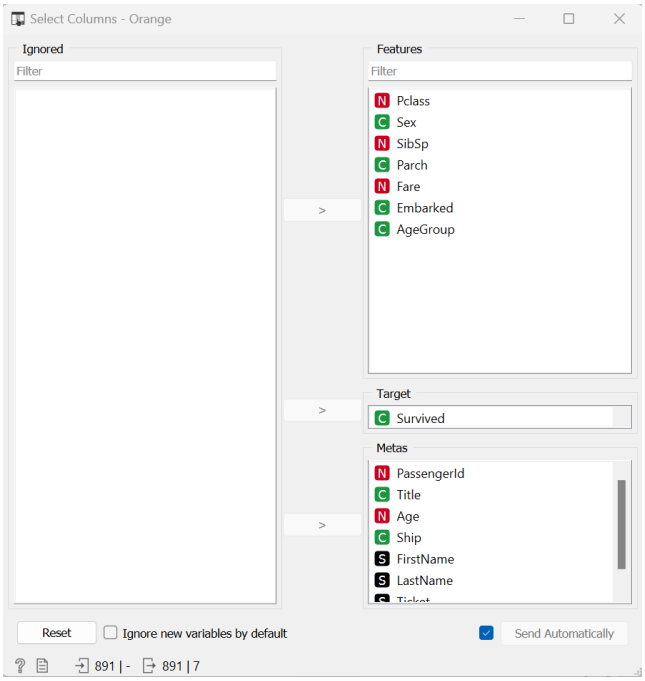


iii. Python Script and New Dataset

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



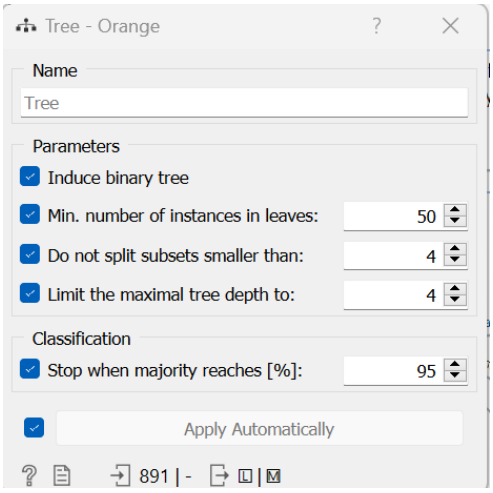
There are 891 instances without missing values and a new “Age group” column. Then we filter a new data set as below.



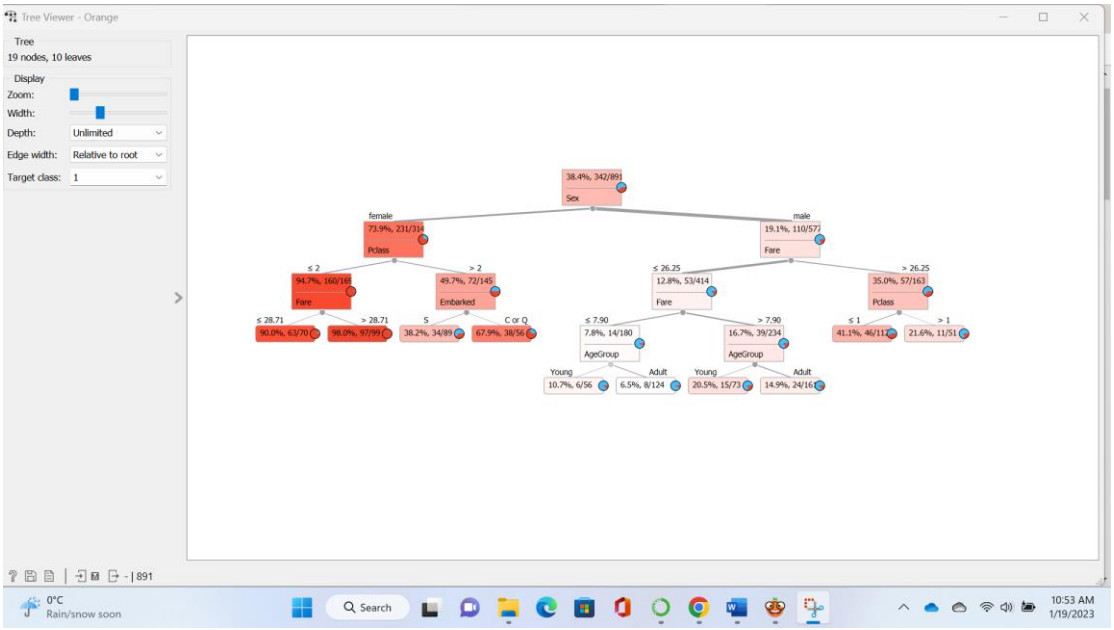
We filtered some data that will be used in the analysis. Because we already have an “AgeGroup” column, we will use that column instead of age. Hence, the columns used include Pclass (Passenger Class), Sex (Passenger Gender), Sibsp (Number of siblings/Spouses abroad), Fare (Ticket price), and Embarked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



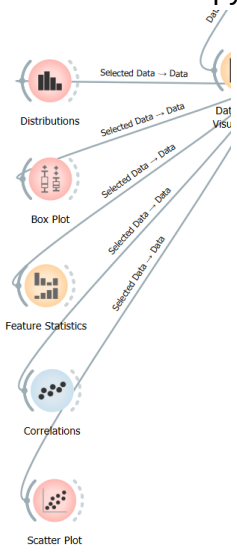
After setting the tree, we can know the tree in Tree Viewer as below:



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.

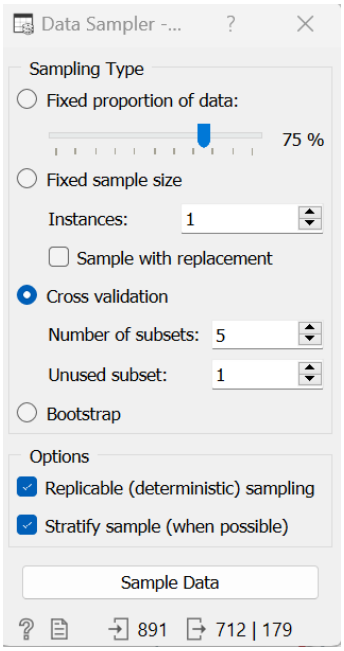
v. Data Visualization

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



vi. Data Sampler

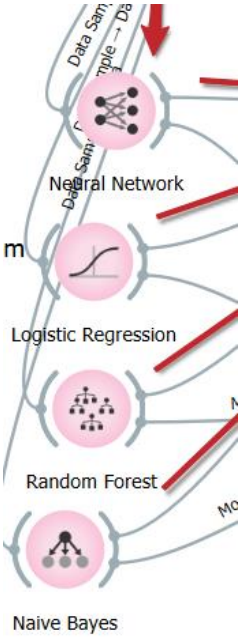
In this step, we choose a cross-validation sampling type and set your data sampler as below.



We usually use cross-validation to tune the hyperparameters of a given machine-learning algorithm to get good performance according to some suitable metric.

vii. Algorithms

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



Neural networks help us cluster and classify. You can think of them as a clustering and classification layer on top of the data you store and manage. They help to group unlabeled data according to similarities among the example inputs, and they classify

data when they have a labeled dataset to train on. (Neural networks can also extract features fed to other algorithms for clustering and classification so that you can think of deep neural networks as components of larger machine-learning applications involving algorithms for reinforcement learning, type, and regression.)

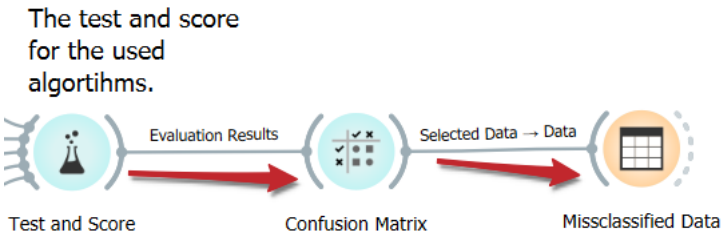
Logistic Regression is a classification technique used in machine learning. It uses a logistic function to model the dependent variable. The dependent variable is dichotomous, i.e., there could only be two possible classes (e.g., either the cancer is malignant or not). As a result, this technique is used while dealing with binary data.

Among all the available classification methods, random forests provide the highest accuracy. **The random forest** technique can also handle big data with numerous variables running into thousands. It can automatically balance data sets when a class is more infrequent than other classes in the data. The method also handles variables fast, making it suitable for complicated tasks.

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.

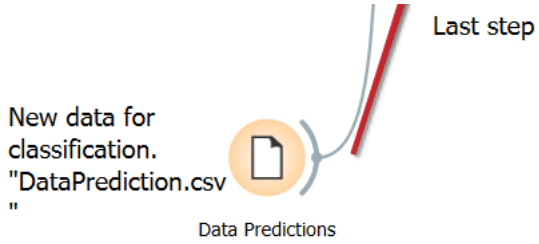
viii. The Test, Result, and Misclassification Data.

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

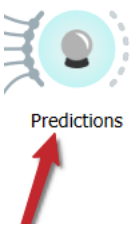


ix. The Predictions

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



Then we check the results in the predictions.



Check the result for predictions