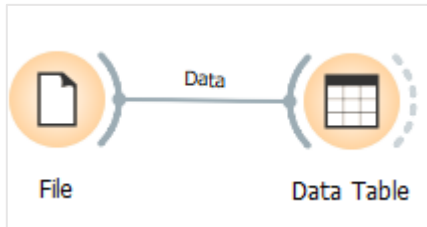


1. Gathering the data collection



Data Table - Orange

Info
1384 instances
127 features (26.9 % missing data)
No target variable.
No meta attributes

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

Selection
☒ Select full rows

Restore Original Order

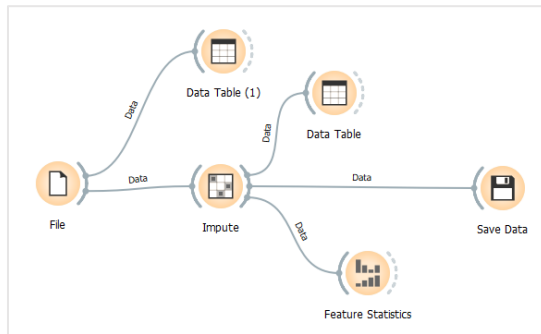
☒ Send Automatically

	covid19_statuses	last.status	age.splits	ender_concept_n
1	positive	discharged	(74,90]	FEMALE
2	positive	deceased	(74,90]	MALE
3	positive	deceased	(74,90]	MALE
4	positive	deceased	(74,90]	FEMALE
5	positive	deceased	(74,90]	FEMALE
6	positive	deceased	(74,90]	FEMALE
7	positive	discharged	(59,74]	FEMALE
8	positive	discharged	(74,90]	FEMALE
9	positive	discharged	(59,74]	MALE
10	positive	discharged	(59,74]	FEMALE
11	positive	discharged	[18,59]	FEMALE
12	positive	deceased	(59,74]	FEMALE
13	positive	deceased	(74,90]	MALE
14	positive	discharged	(74,90]	MALE
15	positive	discharged	[18,59]	FEMALE
16	positive	discharged	(59,74]	MALE
17	positive	discharged	[18,59]	FEMALE

1384 | 1384 | 1384

- By linking file data to the Data Table widget, can better understand the data and select the attributes as a wish. As a result, there are 26.9% missing values, 127 characteristics, and 1383 instances in total.

2. Data Cleaning



File - Orange

Source

File: deidentified_overlap_tcia.csv.cleaned.csv_20210806.csv

Info

1384 instance(s)
127 feature(s) (26.9% missing values)
Regression; numerical class (14.9% missing values)
2 meta attribute(s)

Columns (Double click to edit)

Name	Type	Role	Values
covid19_statuses	categorical	skip	positive
last.status	categorical	target	deceased, discharged
age.splits	categorical	feature	(59,74], (74,90], [18,59]
gender_concept...	categorical	feature	FEMALE, MALE
visit_start_dat...	categorical	skip	1/1/1901, 1/2/1901...
visit_concept...	categorical	skip	Emergency Room ...
is_icu	categorical	feature	FALSE, TRUE

Reset Apply

Browse documentation datasets

1384

- Use for selected or drop any features that aren't relevant in the File widget for preparing datasets, and choose target variable in the column role. Then, using the Impute widget, replace the noisy or missing data with "Average/Most Frequent" and restore the data.

Impute - Orange

Default Method

☐ Don't impute

☒ Average/Most frequent

☐ As a distinct value

☐ Fixed values; numeric variables: 0, time: 1970-01-01 08:00:00

Individual Attribute Settings

Filter...

age.splits

gender_concept_name

is_icu

was_ventilated

length_of_stay

Urine.protein

hf_ef_v

ckd_v

malignancies_v

other_lung_disease_v

Default (above)

Don't impute

Average/Most frequent

As a distinct value

Model-based imputer (simple tree)

Random values

Remove instances with unknown values

Fixed value

Restore All to Default

Apply Automatically

1384 | - 1384

Data Table - Orange

Info

1384 instances (no missing data)
11 features
Target with 2 values
No meta attributes

Variables

☒ Show variable labels (if present)

☒ Visualize numeric values

☒ Color by instance classes

Selection

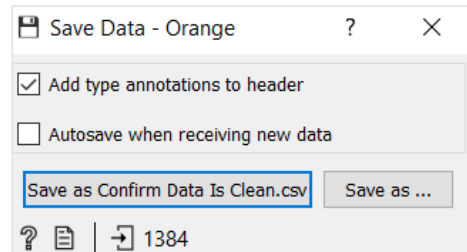
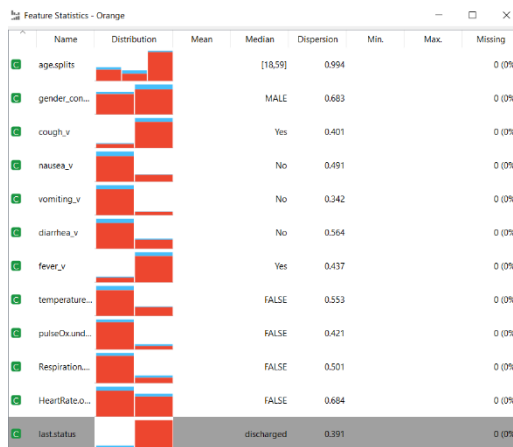
☒ Select full rows

Restore Original Order

Send Automatically

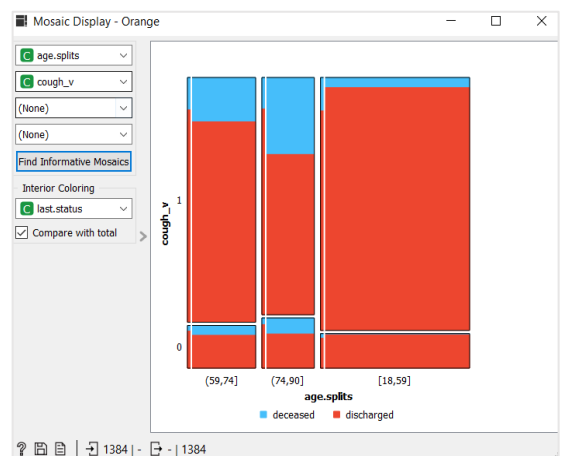
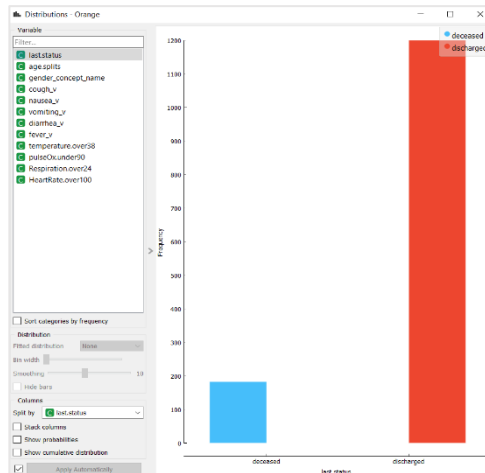
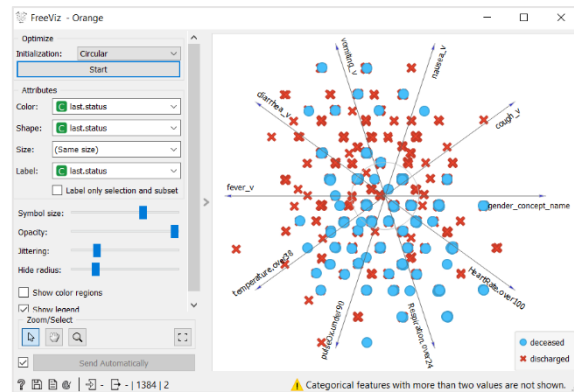
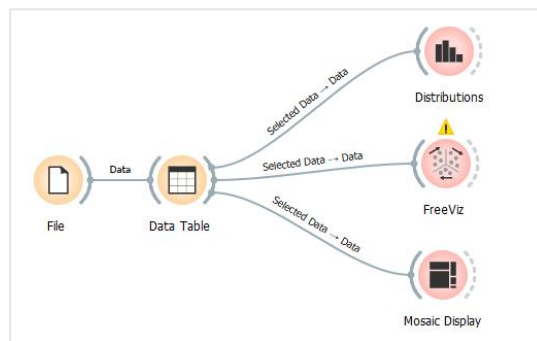
	temperature.over38	pulseOx.under90	Respiration.over24	HeartRate.over1
1	FALSE	FALSE	FALSE	FALSE
2	FALSE	FALSE	TRUE	TRUE
3	FALSE	TRUE	TRUE	FALSE
4	FALSE	FALSE	FALSE	FALSE
5	FALSE	FALSE	FALSE	FALSE
6	FALSE	FALSE	TRUE	FALSE
7	FALSE	FALSE	FALSE	FALSE
8	FALSE	FALSE	TRUE	FALSE
9	FALSE	FALSE	TRUE	FALSE
10	FALSE	FALSE	FALSE	FALSE
11	FALSE	FALSE	FALSE	FALSE
12	FALSE	TRUE	FALSE	FALSE
13	FALSE	FALSE	TRUE	FALSE
14	FALSE	FALSE	FALSE	FALSE
15	FALSE	FALSE	FALSE	FALSE
16	TRUE	TRUE	TRUE	FALSE
17	FALSE	FALSE	FALSE	FALSE

1384 | 1384



- After utilising the Impute Widget, the Feature Statistics widget displays that no missing data was found.
- After ensuring that the data is clean and that the target variable has been selected, save the new data collection using the Save Data widget so that it can be utilised in the next step.

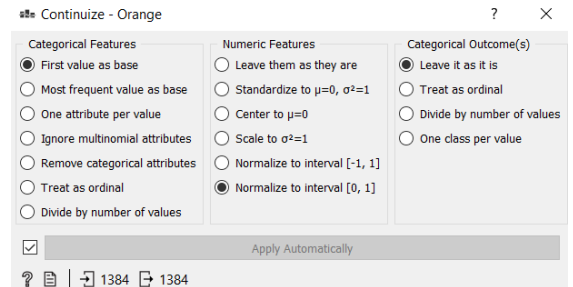
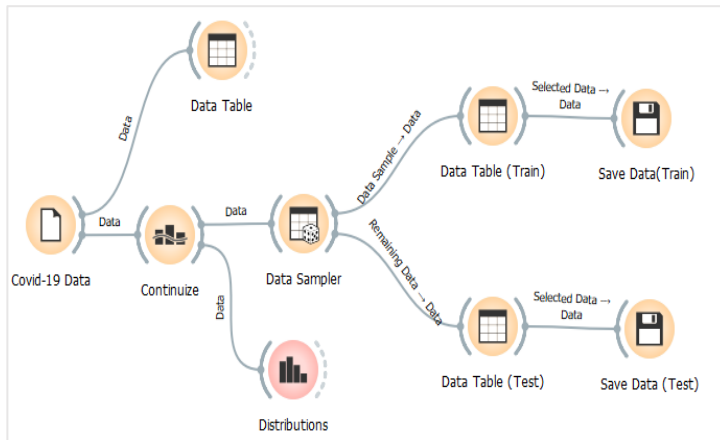
3. Data Preparation



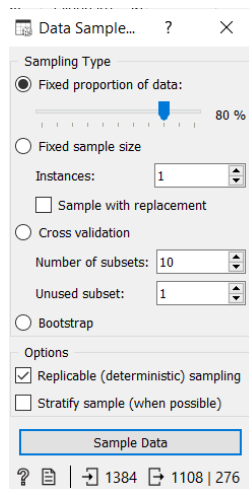
- The Distribution widget, FreeViz widget, and Mosaic Display widget can be used to visualise the features specified for a better understanding.
- The display from the FreeViz widget shows there is no missing data with 11 features selected and 1384 instances. The target with two values represents the last status feature, which can be "Discharged" or "Deceased."
- The Feature Statistics widget provides an explanation of the visualisation and allows to examine the distribution, meaning, media, and dispersion of each feature name. Then, in the next step, connect the Impute widget to the Save Data widget.

4. Choosing A Model For Training and Testing Data

a) Splitting the data



- First and foremost, the dataset must be split. In this stage, convert a categorical feature to a numerical one, such as [Yes:1, No:0]. By clicking "Normalize to interval [0,1]", can alter the Continue widget.



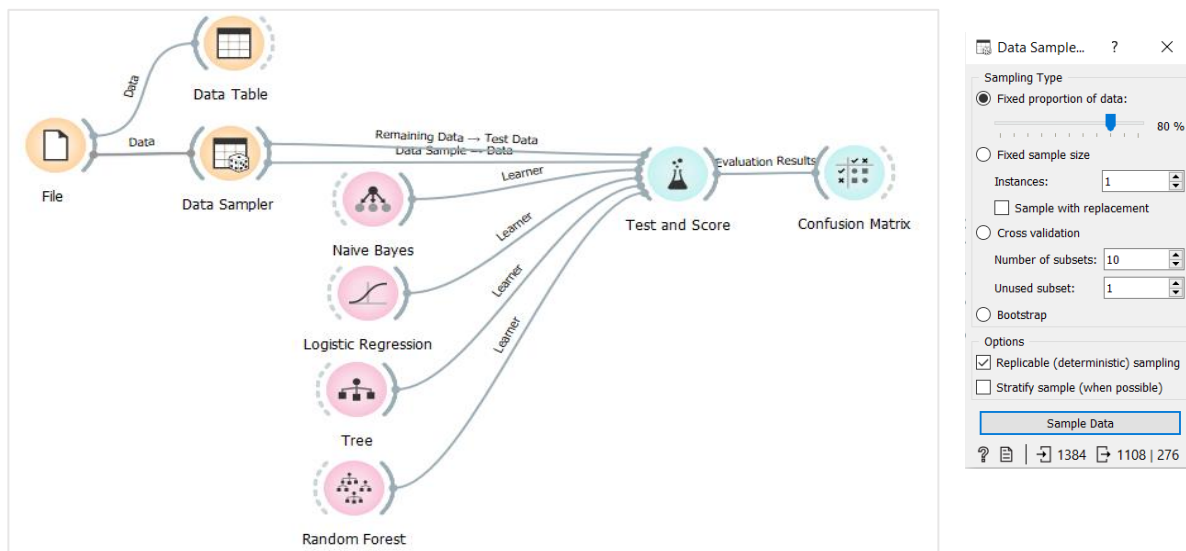
- Then divide the data into 80:20 training and testing data (1108) (276).

	last_status	age_splits=[74,90]	age_splits=[18,59]	er_concept_name
1	discharged	0	1	
2	discharged	0	0	
3	discharged	0	1	
4	discharged	0	1	
5	discharged	0	1	
6	discharged	0	0	
7	deceased	1	0	
8	discharged	0	1	
9	discharged	0	1	
10	deceased	0	0	
11	discharged	0	1	
12	discharged	0	0	
13	discharged	1	0	
14	discharged	0	1	
15	discharged	0	1	
16	discharged	0	0	

	last_status	age_splits=[74,90]	age_splits=[18,59]	er_concept_name
1	discharged	0	1	
2	discharged	0	1	
3	discharged	1	0	
4	discharged	0	1	
5	deceased	0	0	
6	discharged	0	1	
7	discharged	0	1	
8	discharged	0	1	
9	discharged	0	0	
10	discharged	0	1	
11	discharged	0	0	
12	discharged	1	0	
13	deceased	0	0	
14	discharged	0	1	
15	discharged	0	1	
16	discharged	1	0	

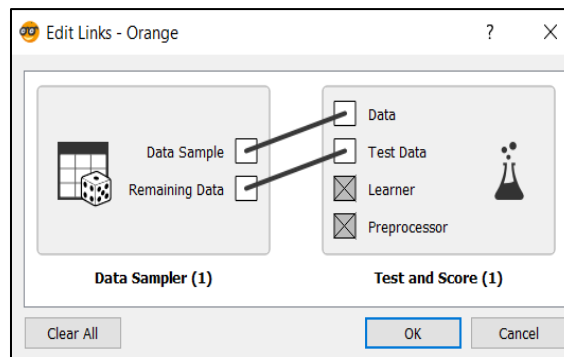
- For some of the features, each figure shows that the training data, which is 1108, and the testing data, which is 276 have been transformed into numerical form. The data can then be saved in the Save Data widget and used for prediction.

b) Choose the model to evaluation



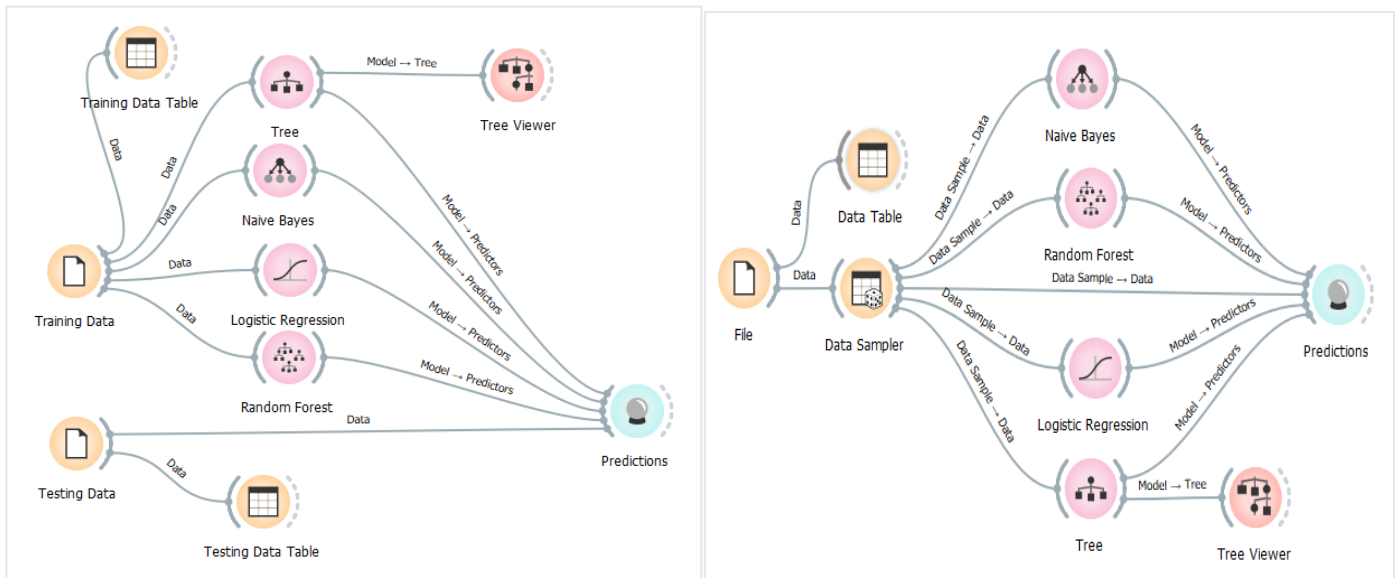
- Alternatively, the Data Sampler widget performs the same function as sklearn's train test split.
- In the Confusion Matrix widget, four classifications were connected to the Test and Score widget to analyse the results and compare the classification prediction.

- Cross-validation has been implemented in the Orange Test & Score widget, which is used for evaluating model performance, preventing overfitting and providing a better indicator of how well the model would perform on unseen data.
- Cross-validation is a widely used statistical method for measuring machine learning model performance (accuracy) and preventing overfitting in a prediction model, especially when data is limited.



- In the Data Sampler widget, Bootstrap was utilised as a Random Forest for categorization. Data Sample -> Data (Train Data: 80) and Data Sample ->Test Data were used in this project study (Test Data: 20).
- After sending the supervised machine learning models to the Test & Score widget with the Train and Test samples, the results of the models may be viewed in the table within the Test & Score widget. However, Test on test data and Test on train data must be clicked before watching the evaluation outcome because there are additional possibilities for evaluation. Then, in Edit Links, must link the relevant Data Sampler with the appropriate Test and Score.

5. Prediction



- There are two types of predictions that can be made: separating Training Data and Testing Data widgets or utilising the Data Sampler widget. Both have the same outcome.

<https://www.analyticsvidhya.com/blog/2020/11/predicting-employee-attrition-using-orange-ows-visual-programming-software/>

<https://towardsdatascience.com/data-science-made-easy-data-modeling-and-prediction-using-orange-f451f17061fa>

<http://docs.biolab.si/orange/2/widgets/rst/evaluate/predictions.html>

<https://phoenixnap.com/kb/handling-missing-data-in-python>

<https://www.analyticsvidhya.com/blog/2021/05/dealing-with-missing-values-in-python-a-complete-guide/>

<https://github.com/chhayac/Machine-Learning-Notebooks>

<https://jakevdp.github.io/PythonDataScienceHandbook/03.04-missing-values.html>

<https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=89096912>

[**https://www.meta.org/papers/using-different-machine-learning-models-to/34542195**](https://www.meta.org/papers/using-different-machine-learning-models-to/34542195)