Semester –VI
Professional Elective – IV

CS3239-Data Warehousing and Mining

CIE-3 Mini Project
15/04/2025

Visual Data Mining for Health Analytics: Heart Disease Prediction and COVID-19 Trends

Team Number: 05
Team Members:
Aditya Sinha – 1RVU22BSC005
Mohammed Ikram – 1RVU22BSC054
Niyanthri R Sridhar – 1RVU22BSC065

Introduction

Visual data mining is transforming healthcare by making complex medical data more interpretable and actionable.

This project integrates two major components using orange:

- Heart Disease Prediction
- COVID-19 Trend Analysis

Objective:

To use visual tools to analyze large, complex health datasets for predictive insights and public health decision-making.

Relevance / Importance of the Chosen Topic

Heart disease remains the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually.

COVID-19 has had a global impact, causing widespread disruption to health systems and societies.

The combination of these topics highlights the value of data-driven tools in:

- Early disease detection
- Monitoring health trends
- Supporting evidence-based public health decisions

Description of the Project & tool

Heart Disease Module

- Tool: Orange 3.36
- Dataset: UCI Heart Disease Dataset
- Techniques: Data cleaning, visualization, classification
- Algorithms: Random Forest

COVID-19 Module

- Tool: Orange 3.36
- Dataset: Global COVID-19 data (WHO)
- Techniques: Time-series analysis, trend visualization, comparison

Implementation

Heart Disease Prediction (Orange)

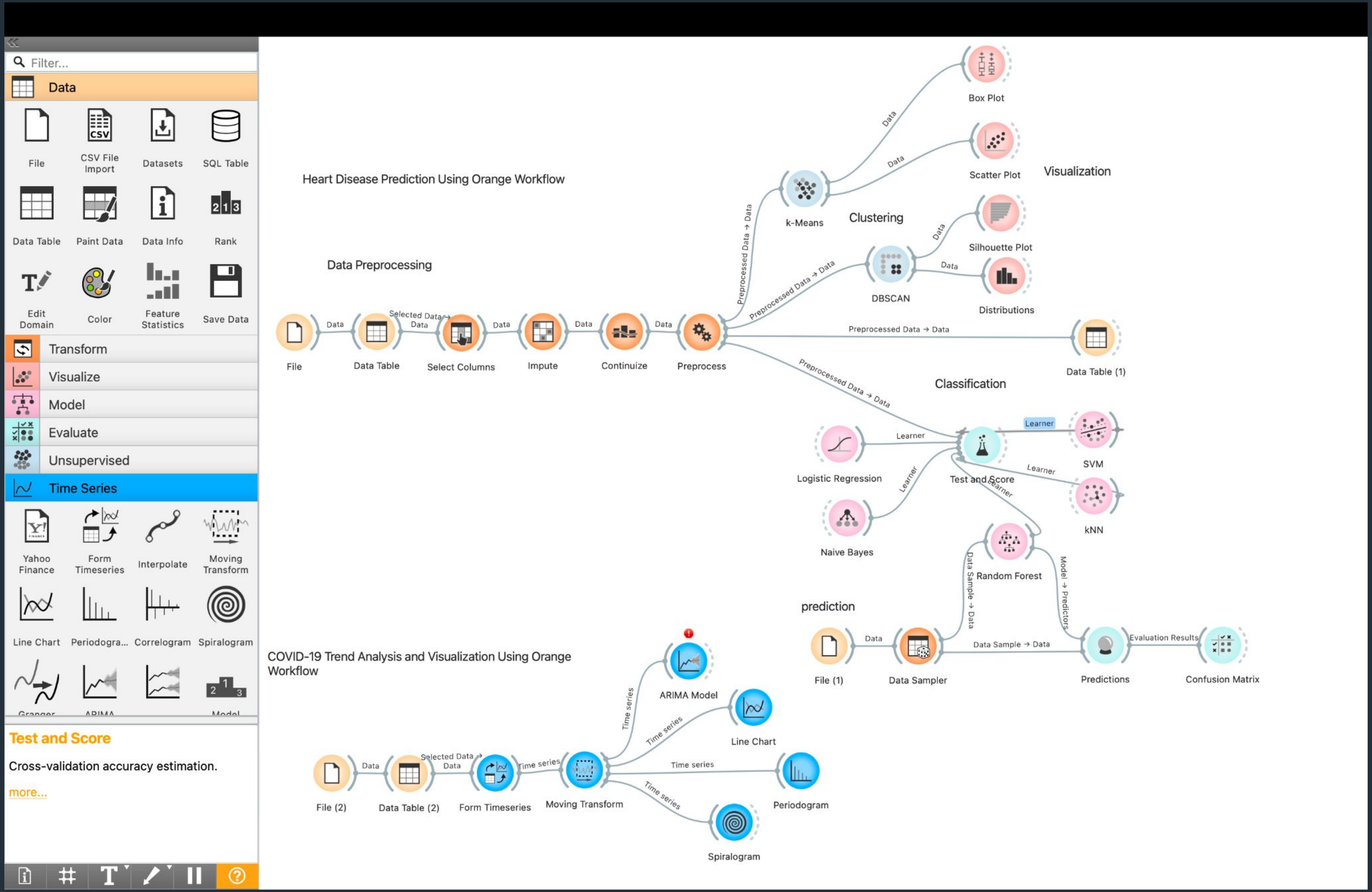
- Load dataset
- Preprocess: handle missing values, normalize attributes
- Visualize features
- Apply classification models
- Evaluate model using Test & Score

COVID-19 Trend Analysis (Orange)

- Load dataset
- Clean and aggregate data
- Visualize trends using line plots and bar charts
- Identify waves and analyze patterns

Screen Shot

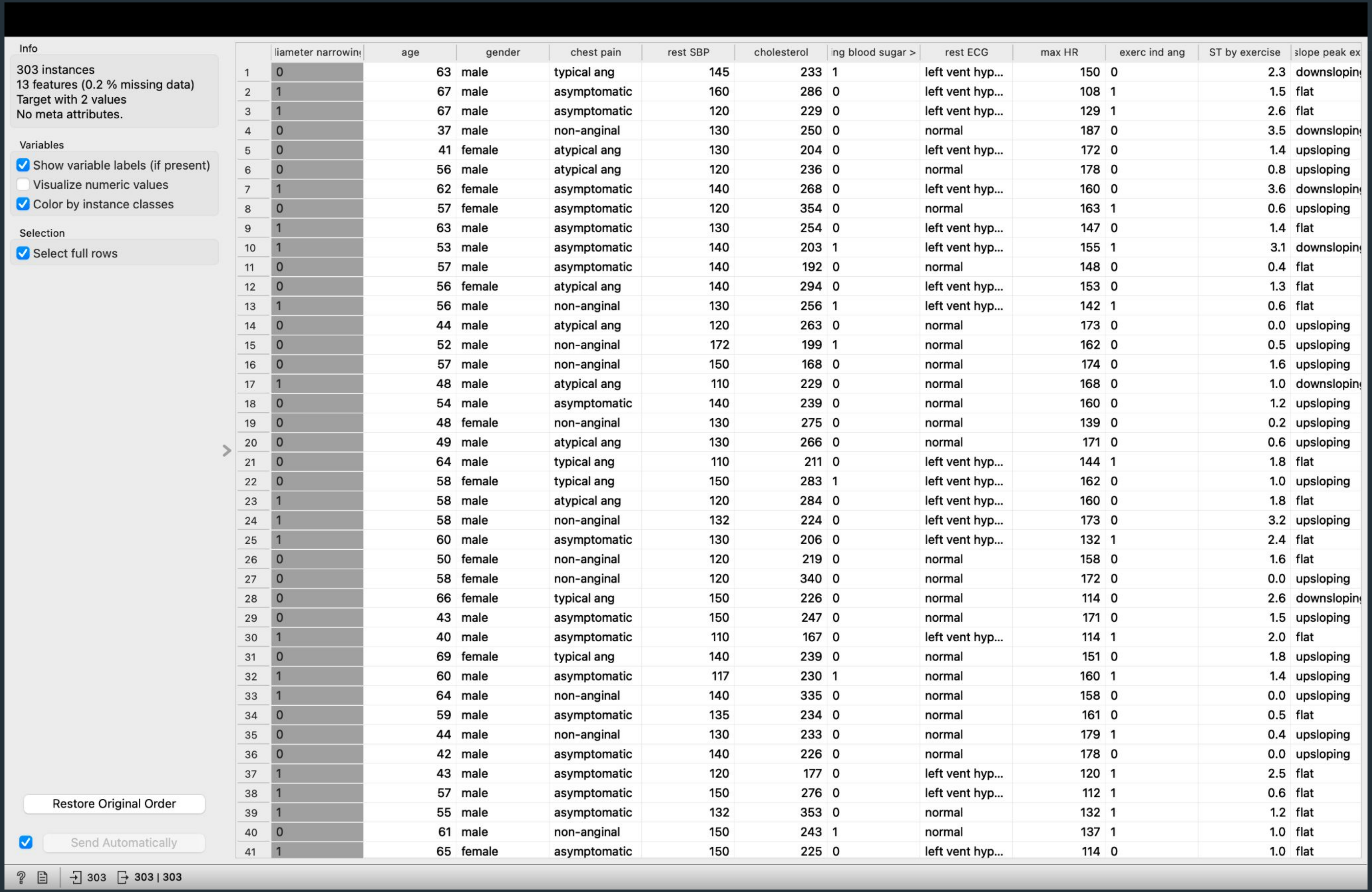
RV UNIVERSITY
Go, change the world
an initiative of RV EDUCATIONAL INSTITUTIONS



Test and Score

Cross-validation accuracy estimation.

more...



Info

303 instances
13 features (0.2 % missing data)
Target with 2 values
No meta attributes.

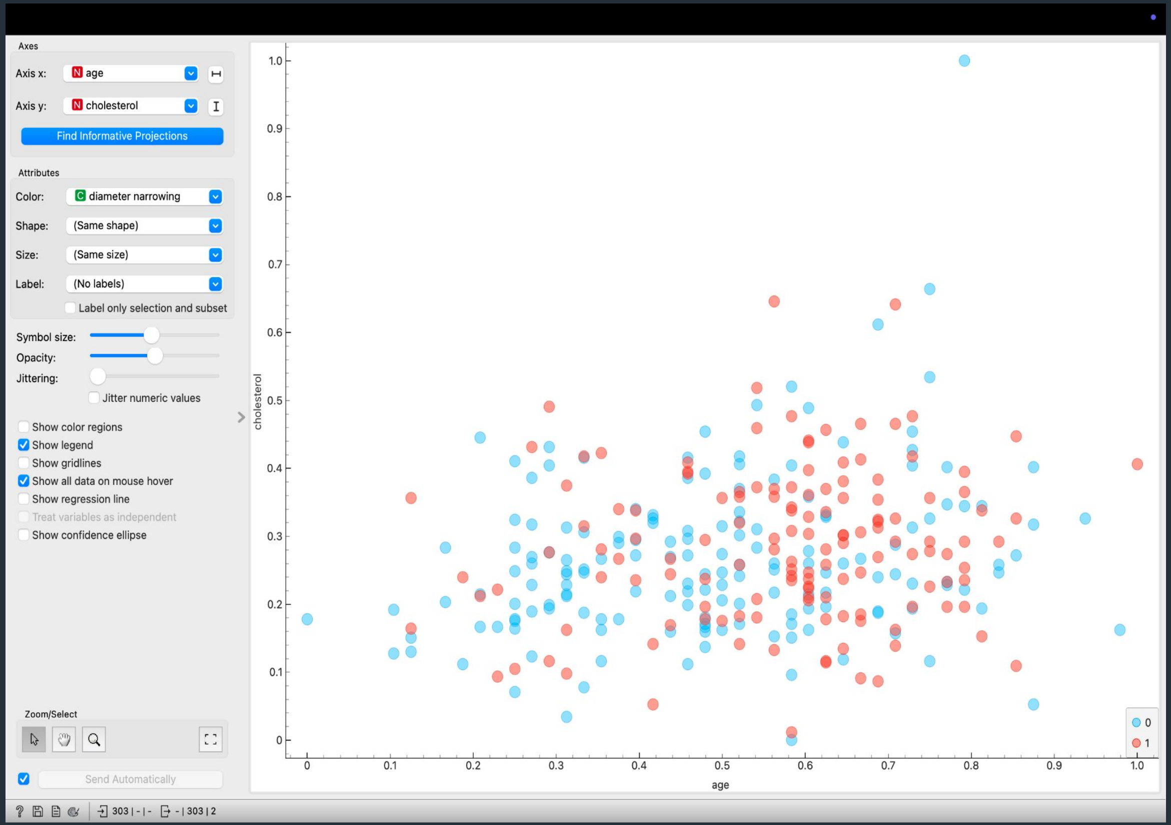
Variables

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

Selection

☑ Select full rows

	liameter narrowin	age	gender	chest pain	rest SBP	cholesterol	ing blood sugar >	rest ECG	max HR	exerc ind ang	ST by exercise	slope peak ex
1	0	63	male	typical ang	145	233	1	left vent hyp...	150	0	2.3	downslopin
2	1	67	male	asymptomatic	160	286	0	left vent hyp...	108	1	1.5	flat
3	1	67	male	asymptomatic	120	229	0	left vent hyp...	129	1	2.6	flat
4	0	37	male	non-anginal	130	250	0	normal	187	0	3.5	downslopin
5	0	41	female	atypical ang	130	204	0	left vent hyp...	172	0	1.4	upsloping
6	0	56	male	atypical ang	120	236	0	normal	178	0	0.8	upsloping
7	1	62	female	asymptomatic	140	268	0	left vent hyp...	160	0	3.6	downslopin
8	0	57	female	asymptomatic	120	354	0	normal	163	1	0.6	upsloping
9	1	63	male	asymptomatic	130	254	0	left vent hyp...	147	0	1.4	flat
10	1	53	male	asymptomatic	140	203	1	left vent hyp...	155	1	3.1	downslopin
11	0	57	male	asymptomatic	140	192	0	normal	148	0	0.4	flat
12	0	56	female	atypical ang	140	294	0	left vent hyp...	153	0	1.3	flat
13	1	56	male	non-anginal	130	256	1	left vent hyp...	142	1	0.6	flat
14	0	44	male	atypical ang	120	263	0	normal	173	0	0.0	upsloping
15	0	52	male	non-anginal	172	199	1	normal	162	0	0.5	upsloping
16	0	57	male	non-anginal	150	168	0	normal	174	0	1.6	upsloping
17	1	48	male	atypical ang	110	229	0	normal	168	0	1.0	downslopin
18	0	54	male	asymptomatic	140	239	0	normal	160	0	1.2	upsloping
19	0	48	female	non-anginal	130	275	0	normal	139	0	0.2	upsloping
20	0	49	male	atypical ang	130	266	0	normal	171	0	0.6	upsloping
21	0	64	male	typical ang	110	211	0	left vent hyp...	144	1	1.8	flat
22	0	58	female	typical ang	150	283	1	left vent hyp...	162	0	1.0	upsloping
23	1	58	male	atypical ang	120	284	0	left vent hyp...	160	0	1.8	flat
24	1	58	male	non-anginal	132	224	0	left vent hyp...	173	0	3.2	upsloping
25	1	60	male	asymptomatic	130	206	0	left vent hyp...	132	1	2.4	flat
26	0	50	female	non-anginal	120	219	0	normal	158	0	1.6	flat
27	0	58	female	non-anginal	120	340	0	normal	172	0	0.0	upsloping
28	0	66	female	typical ang	150	226	0	normal	114	0	2.6	downslopin
29	0	43	male	asymptomatic	150	247	0	normal	171	0	1.5	upsloping
30	1	40	male	asymptomatic	110	167	0	left vent hyp...	114	1	2.0	flat
31	0	69	female	typical ang	140	239	0	normal	151	0	1.8	upsloping
32	1	60	male	asymptomatic	117	230	1	normal	160	1	1.4	upsloping
33	1	64	male	non-anginal	140	335	0	normal	158	0	0.0	upsloping
34	0	59	male	asymptomatic	135	234	0	normal	161	0	0.5	flat
35	0	44	male	non-anginal	130	233	0	normal	179	1	0.4	upsloping
36	0	42	male	asymptomatic	140	226	0	normal	178	0	0.0	upsloping
37	1	43	male	asymptomatic	120	177	0	left vent hyp...	120	1	2.5	flat
38	1	57	male	asymptomatic	150	276	0	left vent hyp...	112	1	0.6	flat
39	1	55	male	asymptomatic	132	353	0	normal	132	1	1.2	flat
40	0	61	male	non-anginal	150	243	1	normal	137	1	1.0	flat
41	1	65	female	asymptomatic	150	225	0	left vent hyp...	114	0	1.0	flat



Axes

Axis x: N age

Axis y: N cholesterol

Attributes

Color: C diameter narrowing

Shape: (Same shape)

Size: (Same size)

Label: (No labels)

□ Label only selection and subset

Symbol size:

Opacity:

Jittering:

□ Jitter numeric values

□ Show color regions

☑ Show legend

□ Show gridlines

☑ Show all data on mouse hover

□ Show regression line

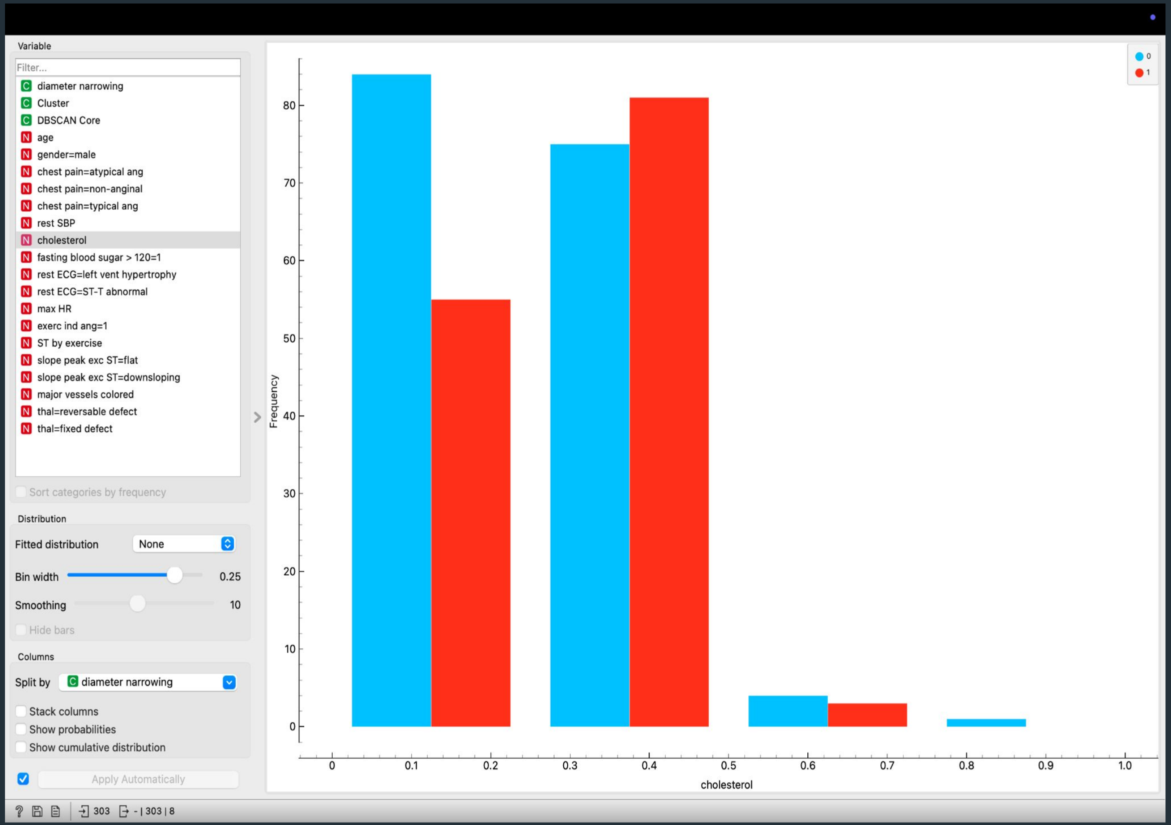
□ Treat variables as independent

□ Show confidence ellipse

Zoom/Select

☑ Send Automatically

303 | - | -   - | 303 | 2



Variable

Filter...

- C diameter narrowing
- C Cluster
- C DBSCAN Core
- N age
- N gender=male
- N chest pain=atypical ang
- N chest pain=non-anginal
- N chest pain=typical ang
- N rest SBP
- N cholesterol
- N fasting blood sugar > 120=1
- N rest ECG=left vent hypertrophy
- N rest ECG=ST-T abnormal
- N max HR
- N exerc ind ang=1
- N ST by exercise
- N slope peak exc ST=flat
- N slope peak exc ST=downsloping
- N major vessels colored
- N thal=reversable defect
- N thal=fixed defect

□ Sort categories by frequency

Distribution

Fitted distribution: None

Bin width: 0.25

Smoothing: 10

□ Hide bars

Columns

Split by: C diameter narrowing

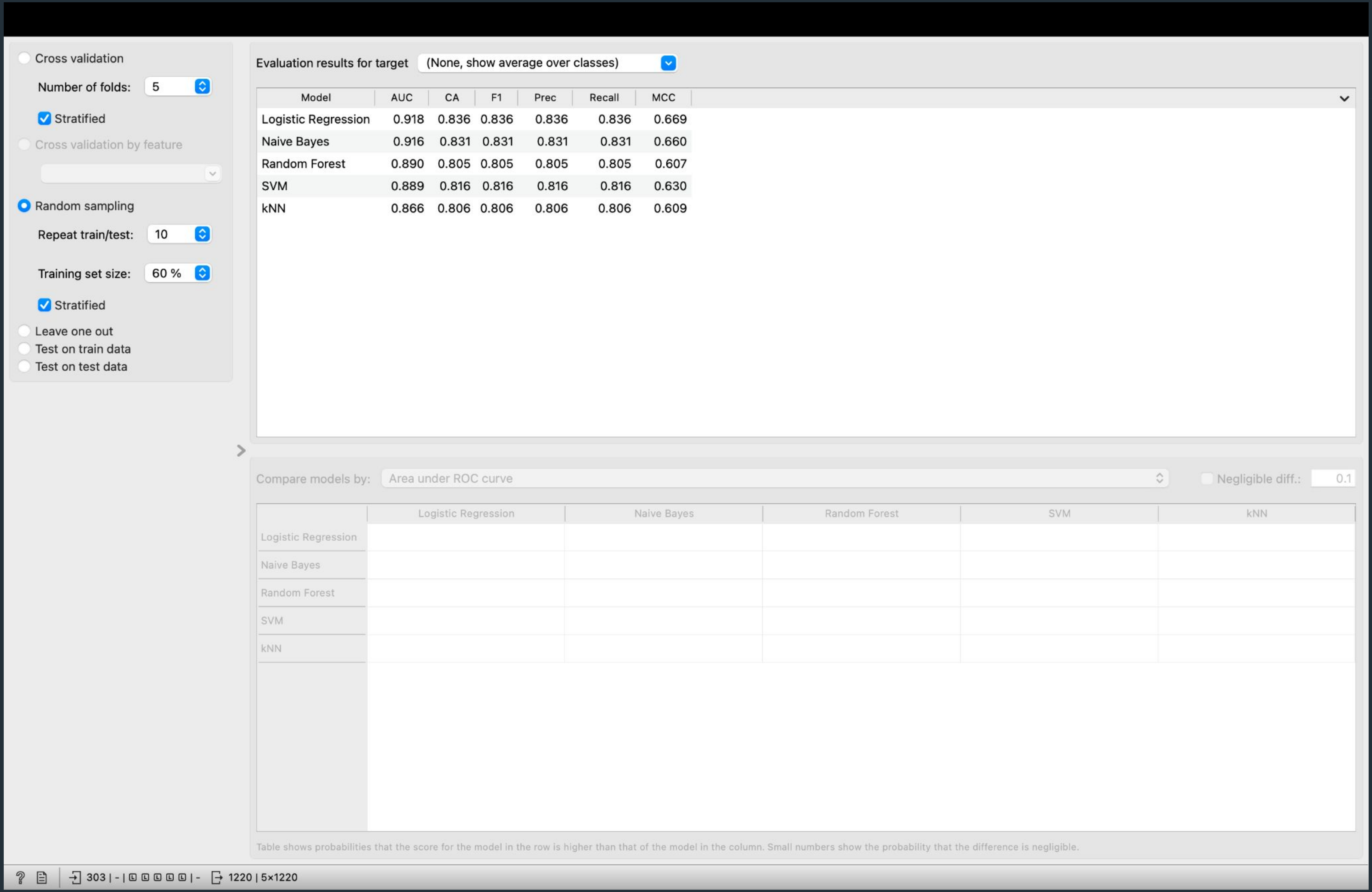
□ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

303   - | 303 | 8



○ Cross validation

Number of folds: 5

☑ Stratified

○ Cross validation by feature

● Random sampling

Repeat train/test: 10

Training set size: 60 %

☑ Stratified

○ Leave one out

○ Test on train data

○ Test on test data

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

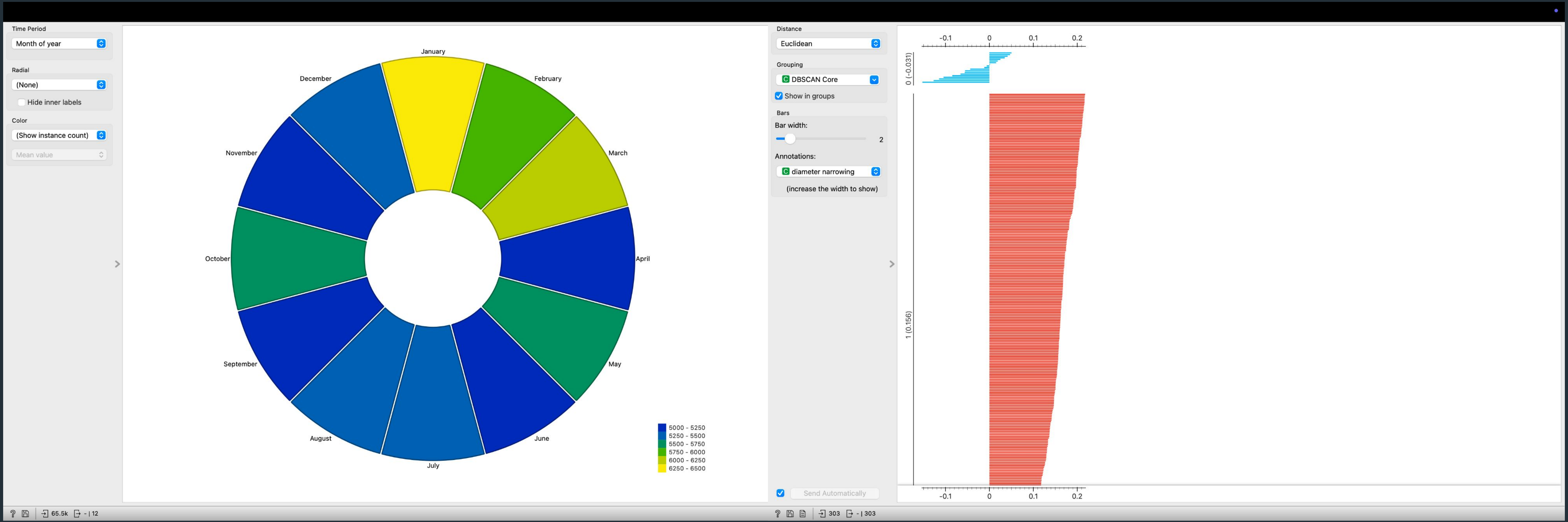
Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.918	0.836	0.836	0.836	0.836	0.669
Naive Bayes	0.916	0.831	0.831	0.831	0.831	0.660
Random Forest	0.890	0.805	0.805	0.805	0.805	0.607
SVM	0.889	0.816	0.816	0.816	0.816	0.630
kNN	0.866	0.806	0.806	0.806	0.806	0.609

Compare models by: Area under ROC curve

□ Negligible diff.: 0.1

	Logistic Regression	Naive Bayes	Random Forest	SVM	kNN
Logistic Regression					
Naive Bayes					
Random Forest					
SVM					
kNN					

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.



Time Period: Month of year

Radial: (None)

☐ Hide inner labels

Color: (Show instance count)

Mean value

65.5k - | 12

Distance: Euclidean

Grouping: DBSCAN Core

☑ Show in groups

Bars

Bar width: 2

Annotations: diameter narrowing

(increase the width to show)

☑ Send Automatically

303 - | 303

Applications

- Supports early diagnosis and risk prediction of heart disease
- Helps monitor COVID-19 infection patterns
- Assists public health decision-making and planning
- Useful for educational purposes in data science and epidemiology
- Demonstrates practical application of visual data mining techniques

Limitations / Challenges

- Datasets used are static and not updated in real-time
- Orange has limited capabilities for advanced modeling and parameter tuning
- Model performance depends heavily on data quality and completeness
- COVID-19 analysis focuses on visualization rather than prediction
- Generalizability of the models may be limited across diverse populations

Conclusion

This project demonstrates the utility of visual data mining in health analytics through practical applications in heart disease prediction and COVID-19 trend analysis. Orange provides an intuitive platform for building machine learning models.

Future enhancements could include:

- Real-time data integration
- Predictive modeling for COVID-19
- Deployment of insights in interactive dashboards or mobile apps

References

- UCI ML Repository – Heart Disease Dataset
- WHO COVID-19 Dashboard
- Orange Data Mining – https://orangedatamining.com

Thank you