RV UNIVERSITY

Go, change the world

an initiative of RV EDUCATIONAL INSTITUTIONS

CS3239 – Data Warehousing and Mining

Mini Project Report

Visual Data Mining for health analytics: Heart Disease Prediction and COVID-19 Trends

---

By

Mohammed Ikram - 1RVU22BSC054

School of Computer Science and Engineering

RV University, Bangalore

School of Computer Science and Engineering

CERTIFICATE

Certified that the CS3239 Data Warehousing and Mining Mini Project Report titled Visual Data Mining for health analytics: Heart Disease Prediction and COVID-19 Trends is carried out by Mohammed Ikram (1RVU22BSC054) who is a bonafide student of the School of Computer Science and Engineering, RV University, Bengaluru, during the year 2024–25. It is certified that all corrections/ suggestions from all the continuous internal evaluations have been incorporated in the project and in this report.

Faculty Guide

Program Director

Index Page with Tabulation

Introduction

The exponential growth of healthcare data has created a pressing need for effective tools that can derive actionable insights from complex, multidimensional datasets. Visual data mining addresses this challenge by enabling intuitive exploration, pattern discovery, and interpretability—key factors in clinical and public health decision-making.

This mini-project applies visual data mining techniques to two critical global health issues: heart disease and COVID-19. By integrating these domains, the project aims to demonstrate the versatility of data analytics in both predictive and trend-based analysis.

The project is divided into two parts:

- Heart Disease Prediction using the Orange Data Mining platform to apply supervised learning models like Random Forest.
- COVID-19 Trend Analysis using Orange to visualize time-series data and uncover infection waves, vaccination progress, and recovery patterns.

Together, these approaches highlight how visual analytics can enhance healthcare diagnosis, monitoring, and policy formulation.

Relevance / Importance of chosen topic

Heart disease remains the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually (WHO, 2023). COVID-19 has had a global impact, causing widespread disruption to health systems and societies.

The convergence of these two issues offers an opportunity to demonstrate how data-driven techniques can transform healthcare. Early diagnosis of heart disease through predictive modeling and visual analytics of COVID-19 trends enhances preparedness, resource allocation, and long-term public health strategies.

By leveraging intuitive tool like Orange, this project underscores the significance of accessible, interpretable data mining in modern medical research and practice.

Description of the Mini Project and the tool used

This mini project is divided into two parts, each addressing a major health concern—Heart Disease and COVID-19—using Orange to extract meaningful insights.

Heart Disease Prediction:

- **Dataset:** UCI Heart Disease Dataset
- **Techniques:** Data cleaning, visualization, and classification using algorithms - Random Forest.
- **Tool:** Orange 3.36 – a visual programming platform for data mining and machine learning

In this part, we utilize Orange's drag-and-drop interface to preprocess and analyze heart disease data. The focus is on identifying patterns and risk factors that contribute to heart disease, enabling early prediction and supporting preventive healthcare strategies.

COVID-19 Trend Analysis

- **Dataset:** Global COVID-19 data from WHO
- **Tools:** Orange 3.36
- **Purpose:** To explore trends, identify pandemic waves, and understand the broader public health impact using visual analytic.

The goal is to track the spread of COVID-19 over time, visualize key metrics such as daily cases and death rates, and analyze the effectiveness of global response measures. Together, both parts of the project demonstrate the power of visual data mining in addressing real-world healthcare challenges. Both modules were implemented using Orange 3.36, which offers visual programming features for data mining and machine learning.

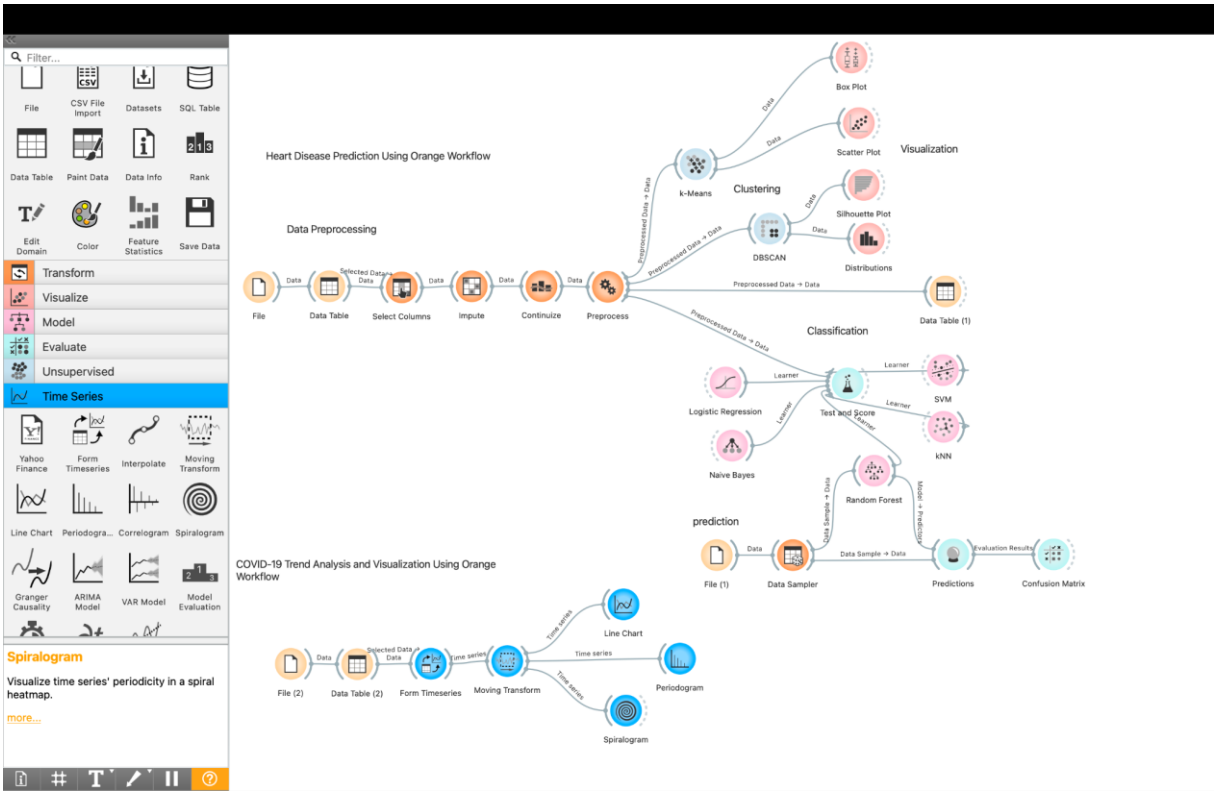
Implementation / Procedure / Steps to execute the task

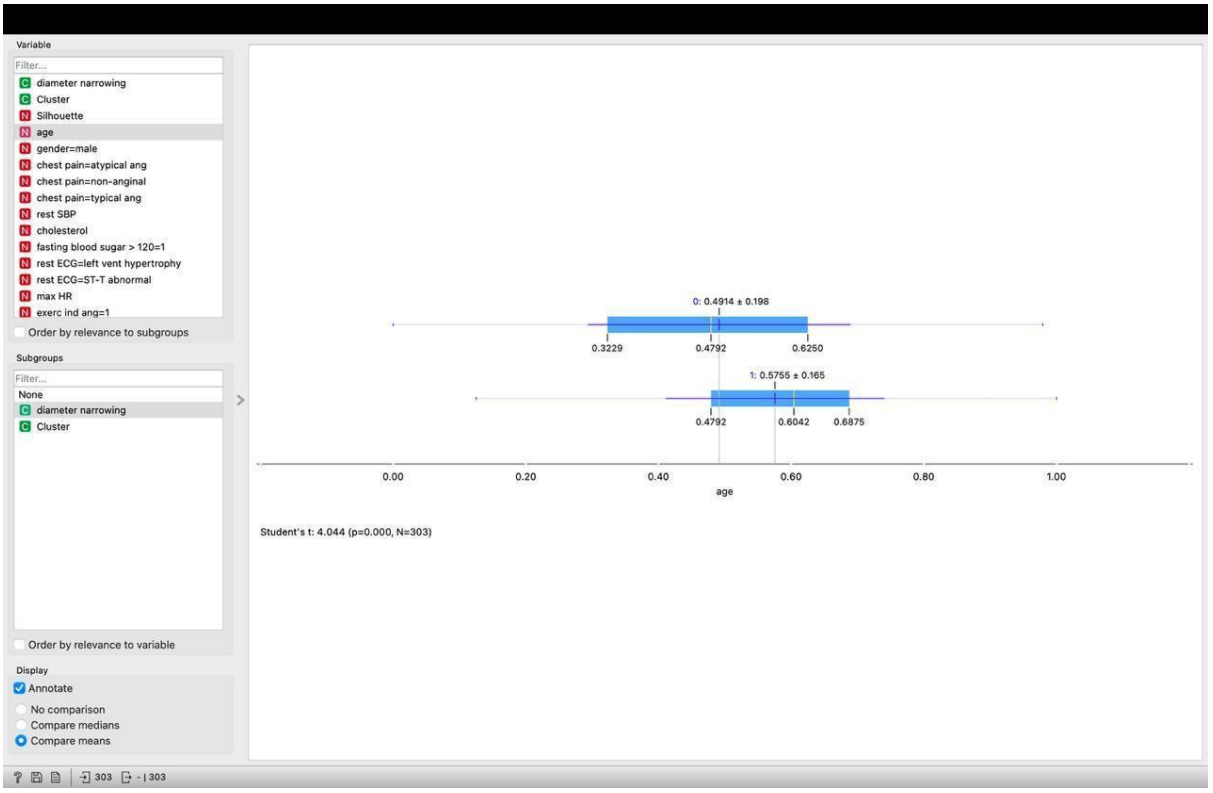
In the heart disease prediction module, the UCI Heart Disease dataset was imported into Orange, where preprocessing involved normalization, handling missing values, and encoding categorical attributes. Exploratory visualizations such as box plots and scatter plots helped identify influential features (e.g., age, cholesterol).

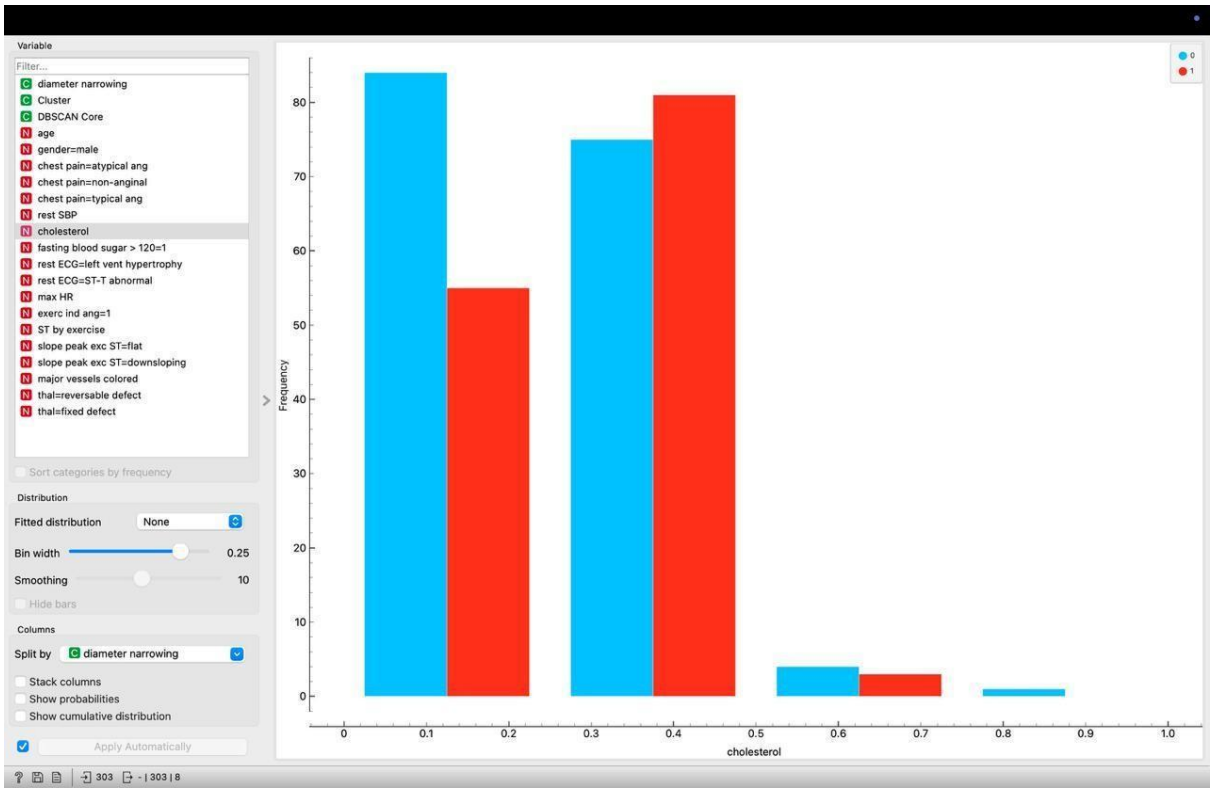
Heart Disease Prediction:
1. Load dataset into Orange
2. Preprocess: handle missing values, normalize attributes
3. Visualize data using scatter plots and box plots
4. Apply Random Forest classifier
5. Evaluate model performance using Test & Score widget

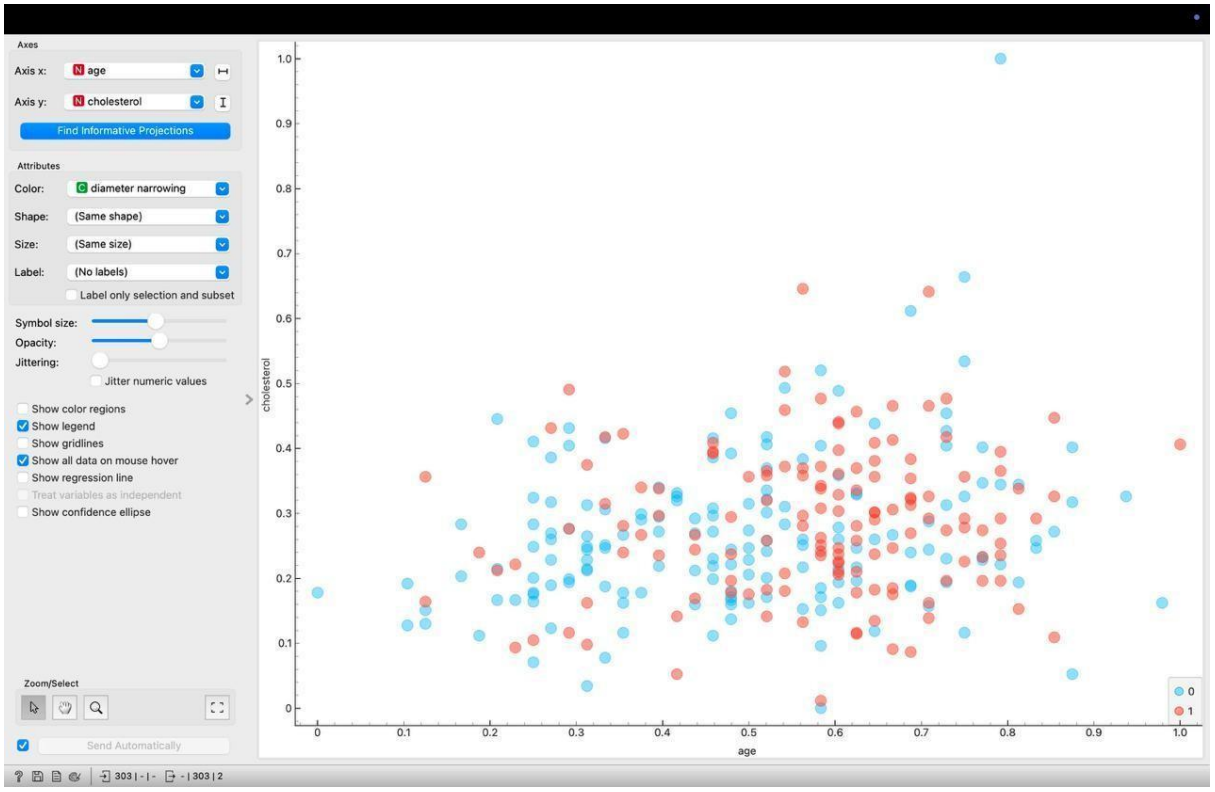
COVID-19 Trend Analysis:
1. Load COVID-19 dataset into Orange
2. Clean and aggregate data by date and region
3. Visualize trends using line plots and bar charts
4. Identify waves and analyze temporal patterns

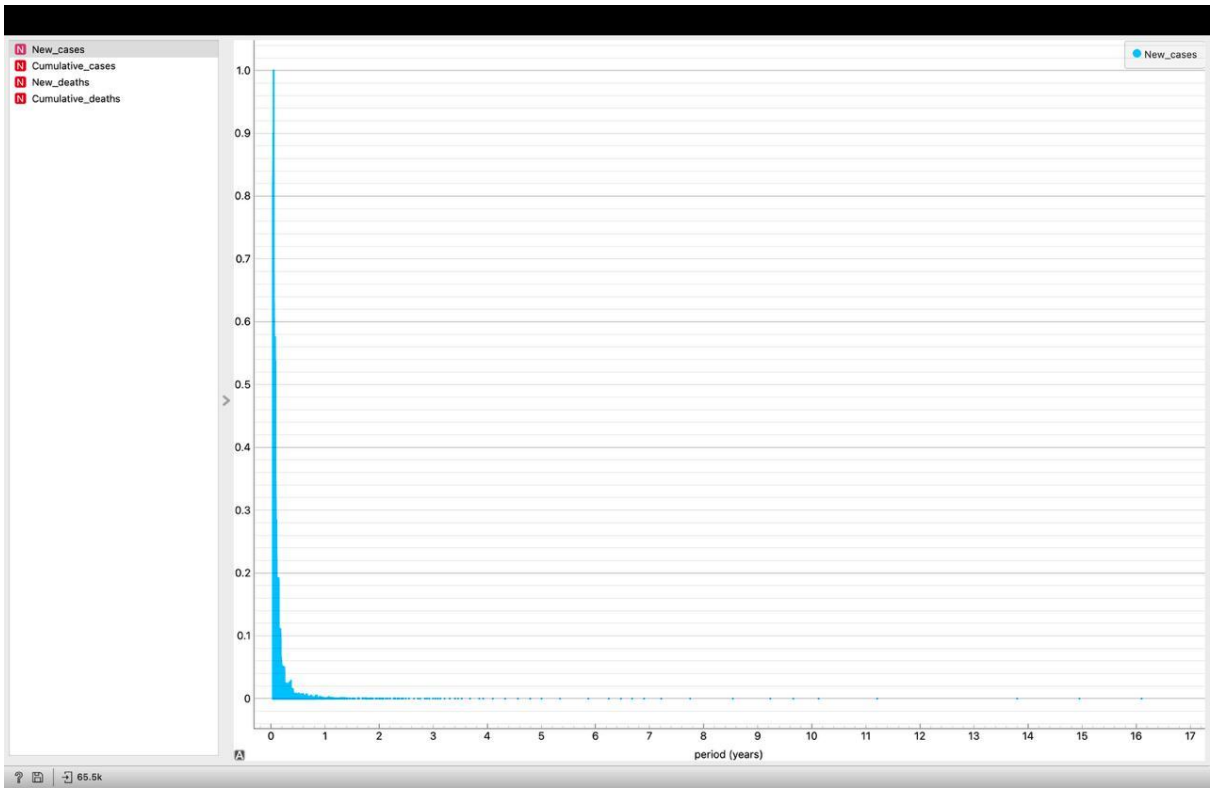
Screenshots of the task executed

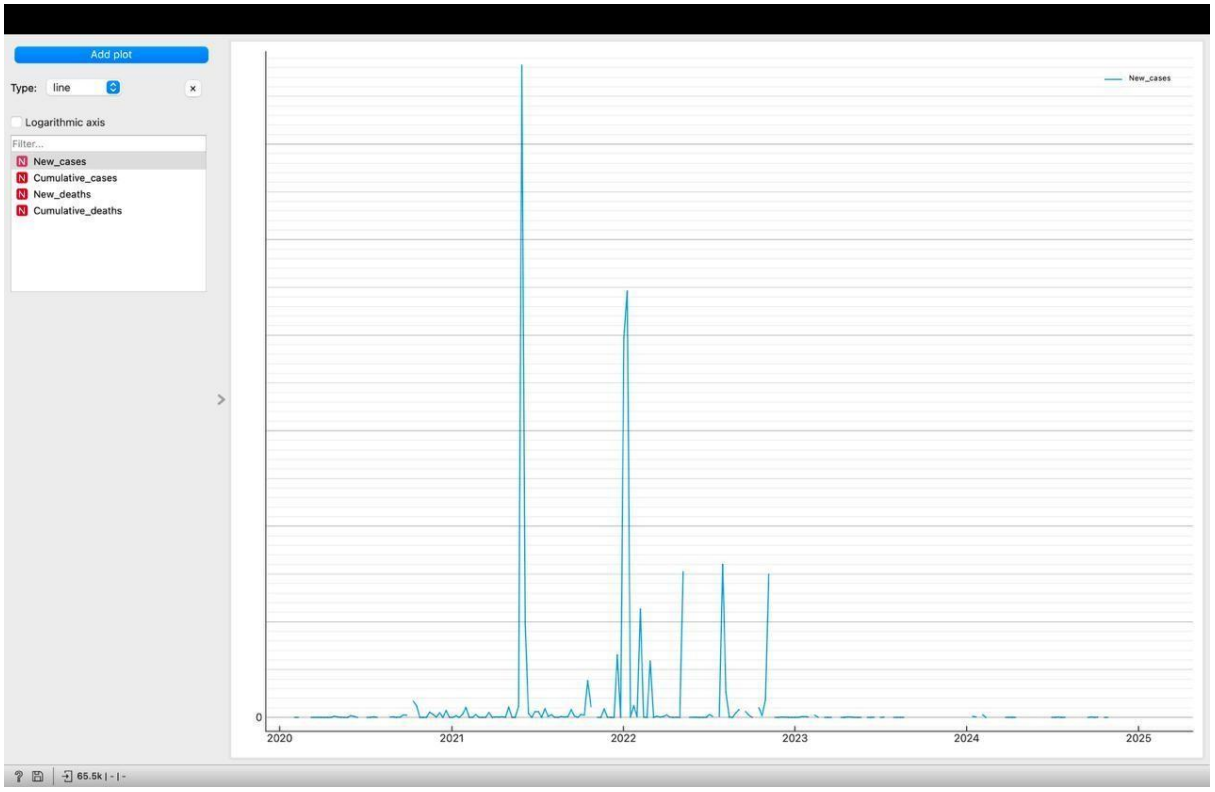


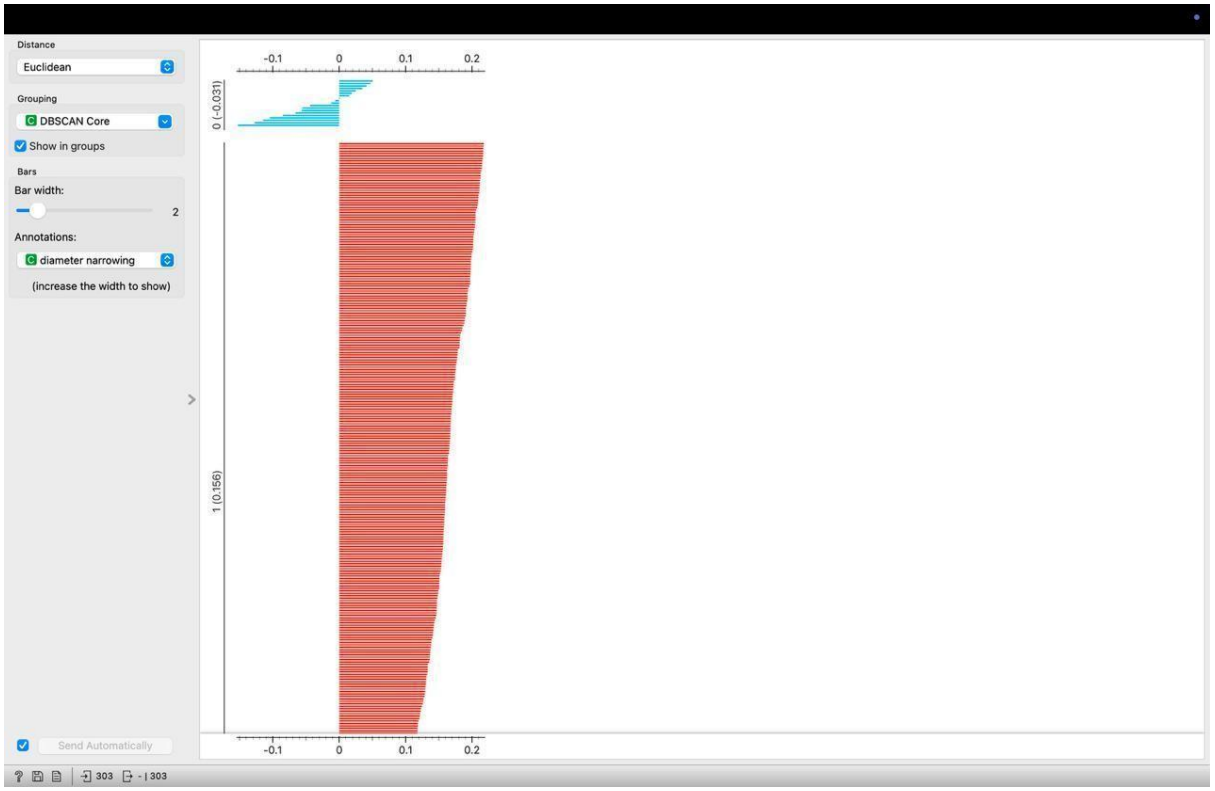

Variable
Filter...
diameter narrowing
Cluster
Silhouette
age
gender=male
chest pain=atypical ang
chest pain=non-anginal
chest pain=typical ang
rest SBP
cholesterol
fasting blood sugar > 120=1
rest ECG=left vent hypertrophy
rest ECG=ST-T abnormal
max HR
exerc ind ang=1
Order by relevance to subgroups
Subgroups
Filter...
None
diameter narrowing
Cluster
Order by relevance to variable
Display
Annotate
No comparison
Compare medians
Compare means
0: 0.4914 ± 0.198
0.3229
0.4792
0.6250
1: 0.5755 ± 0.165
0.4792
0.6042
0.6875
0.00
0.20
0.40
0.60
0.80
1.00
age
Student's t: 4.044 (p=0.000, N=303)
303
- | 303

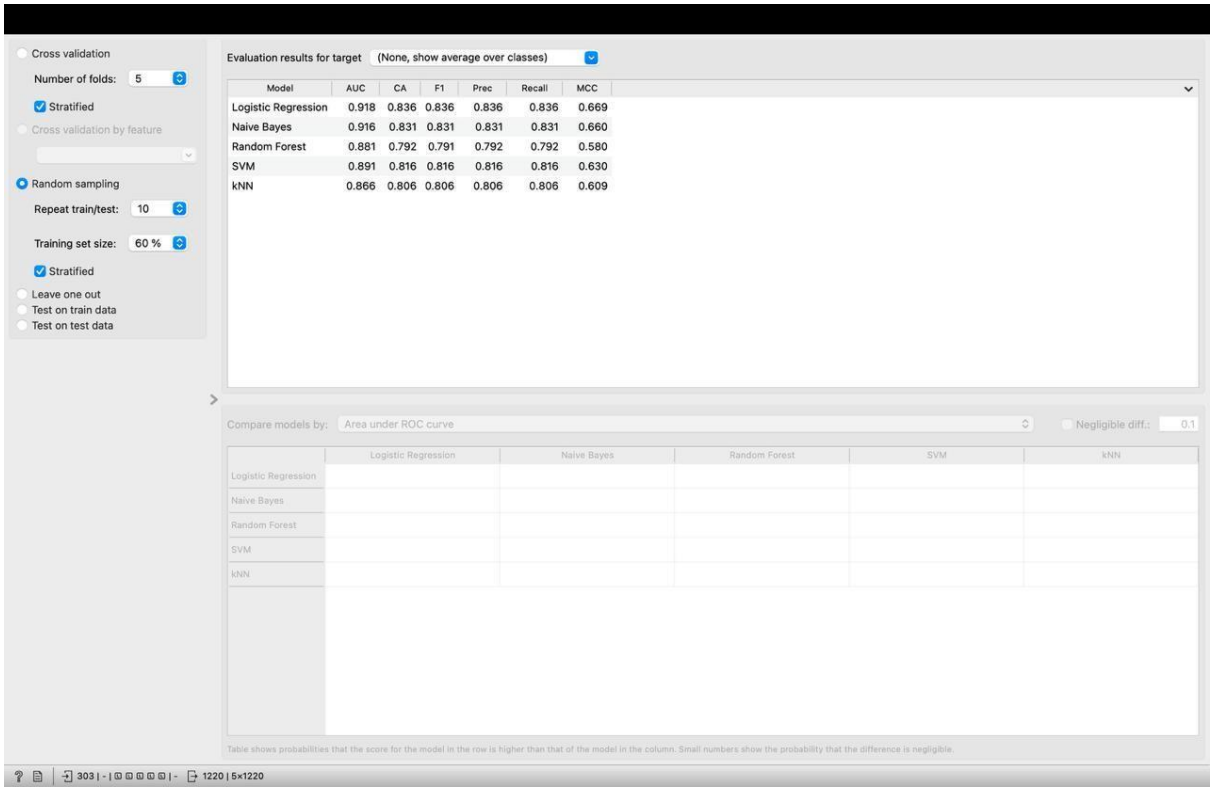

Variable
Filter...
diameter narrowing
Cluster
DBSCAN Core
age
gender=male
chest pain=atypical ang
chest pain=non-anginal
chest pain=typical ang
rest SBP
cholesterol
fasting blood sugar > 120=1
rest ECG=left vent hypertrophy
rest ECG=ST-T abnormal
max HR
exerc ind ang=1
ST by exercise
slope peak exc ST=flat
slope peak exc ST=downsloping
major vessels colored
thal=reversable defect
thal=fixed defect
Sort categories by frequency
Distribution
Fitted distribution
None
Bin width
0.25
Smoothing
10
Hide bars
Columns
Split by
diameter narrowing
Stack columns
Show probabilities
Show cumulative distribution
Apply Automatically
Frequency
0
10
20
30
40
50
60
70
80
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
cholesterol
0
1
303
- | 303 | 8

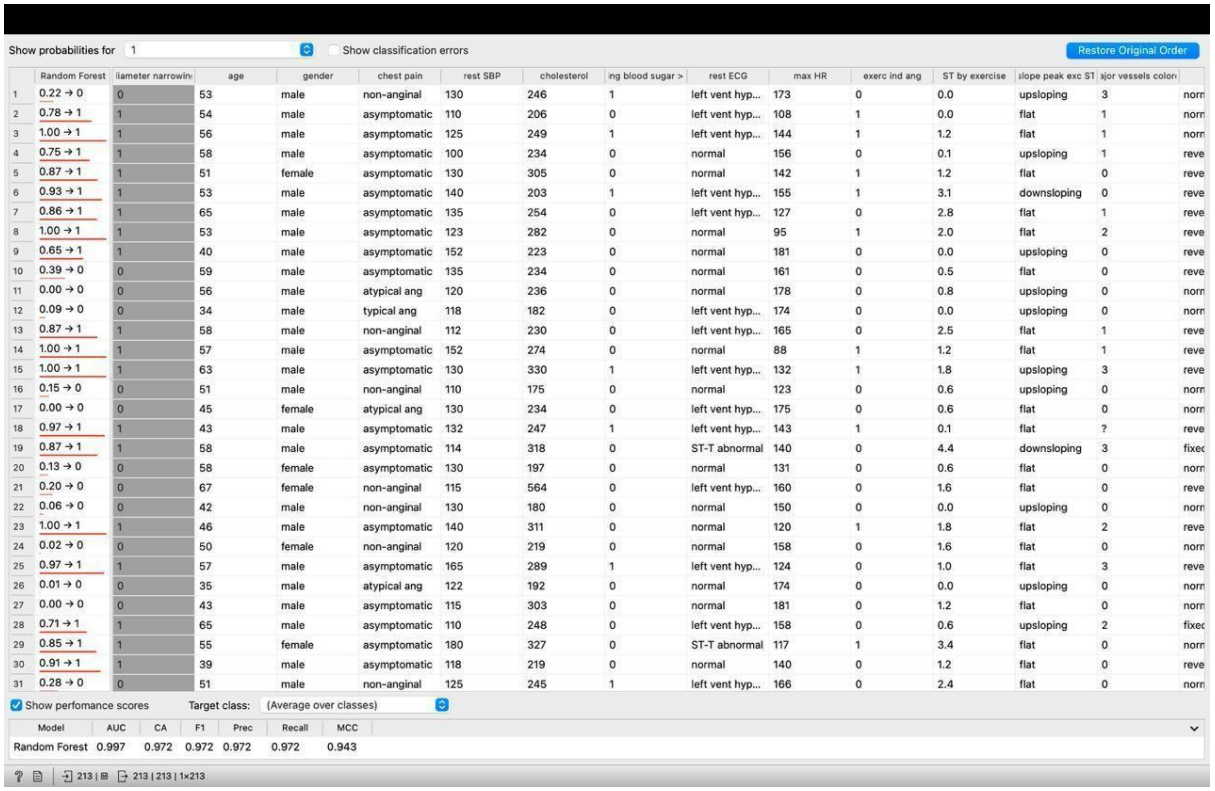

Axes
Axis x: age
Axis y: cholesterol
Find Informative Projections
Attributes
Color: diameter narrowing
Shape: (Same shape)
Size: (Same size)
Label: (No labels)
Label only selection and subset
Symbol size:
Opacity:
Jittering:
Jitter numeric values
Show color regions
Show legend
Show gridlines
Show all data on mouse hover
Show regression line
Treat variables as independent
Show confidence ellipse
Zoom/Select
Send Automatically
303 | - | - | 303 | 2
cholesterol
age
0
1

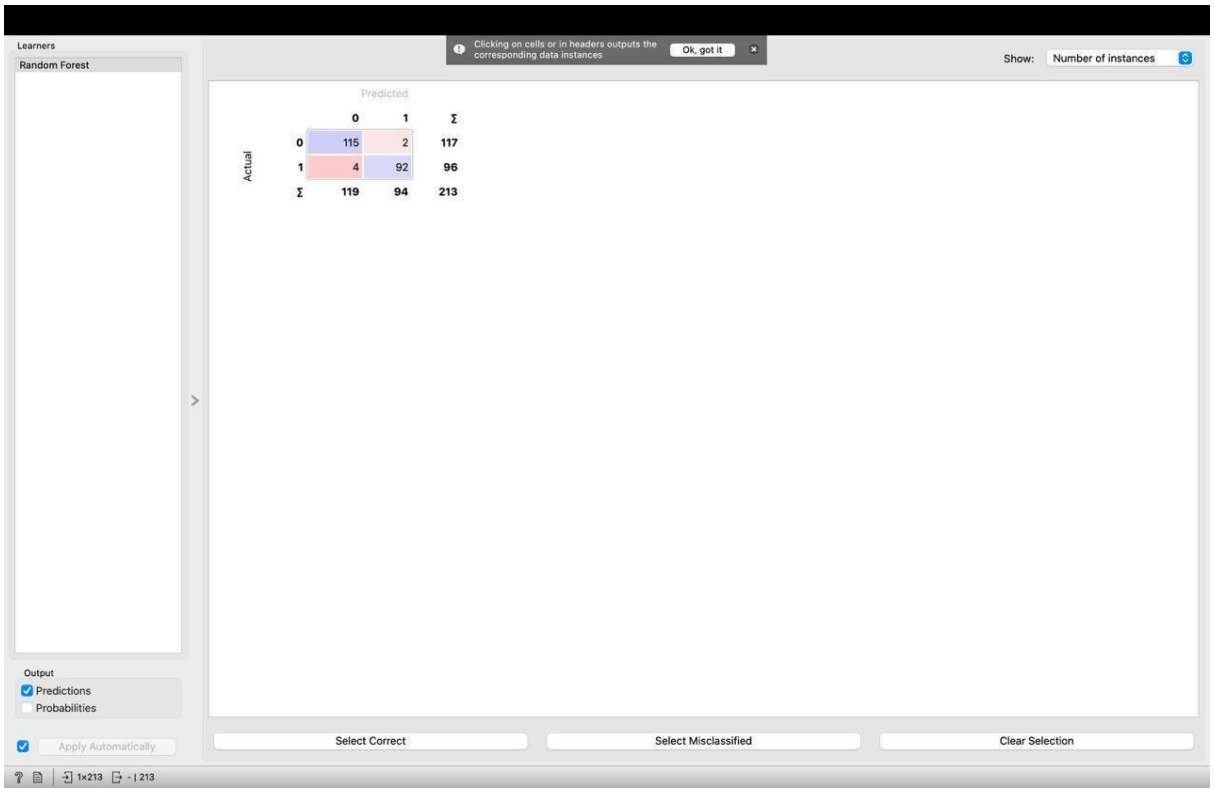

New_cases
Cumulative_cases
New_deaths
Cumulative_deaths
New_cases
period (years)
65.5k


Add plot
Type: line
Logarithmic axis
Filter...
New_cases
Cumulative_cases
New_deaths
Cumulative_deaths
New_cases
0
2020
2021
2022
2023
2024
2025
65.5k | - | -


Distance
Euclidean
Grouping
DBSCAN Core
Show in groups
Bars
Bar width:
2
Annotations:
diameter narrowing
(increase the width to show)
Send Automatically
-0.1
0
0.1
0.2
0 (-0.031)
1 (0.156)
303
- | 303


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 0.881 0.792 0.791 0.792 0.792 0.580
SVM 0.891 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
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.
303 | - | 1220 | 5×1220


Show probabilities for 1
Show classification errors
Restore Original Order
Random Forest | liameter narrowin | age | gender | chest pain | rest SBP | cholesterol | ng blood sugar > | rest ECG | max HR | exerc ind ang | ST by exercise | slope peak exc ST | ajor vessels color |
1 0.22 → 0 | 0 | 53 | male | non-anginal | 130 | 246 | 1 | left vent hyp... | 173 | 0 | 0.0 | upsloping | 3 | norn
2 0.78 → 1 | 1 | 54 | male | asymptomatic | 110 | 206 | 0 | left vent hyp... | 108 | 1 | 0.0 | flat | 1 | norn
3 1.00 → 1 | 1 | 56 | male | asymptomatic | 125 | 249 | 1 | left vent hyp... | 144 | 1 | 1.2 | flat | 1 | norn
4 0.75 → 1 | 1 | 58 | male | asymptomatic | 100 | 234 | 0 | normal | 156 | 0 | 0.1 | upsloping | 1 | reve
5 0.87 → 1 | 1 | 51 | female | asymptomatic | 130 | 305 | 0 | normal | 142 | 1 | 1.2 | flat | 0 | reve
6 0.93 → 1 | 1 | 53 | male | asymptomatic | 140 | 203 | 1 | left vent hyp... | 155 | 1 | 3.1 | downsloping | 0 | reve
7 0.86 → 1 | 1 | 65 | male | asymptomatic | 135 | 254 | 0 | left vent hyp... | 127 | 0 | 2.8 | flat | 1 | reve
8 1.00 → 1 | 1 | 53 | male | asymptomatic | 123 | 282 | 0 | normal | 95 | 1 | 2.0 | flat | 2 | reve
9 0.65 → 1 | 1 | 40 | male | asymptomatic | 152 | 223 | 0 | normal | 181 | 0 | 0.0 | upsloping | 0 | reve
10 0.39 → 0 | 0 | 59 | male | asymptomatic | 135 | 234 | 0 | normal | 161 | 0 | 0.5 | flat | 0 | reve
11 0.00 → 0 | 0 | 56 | male | atypical ang | 120 | 236 | 0 | normal | 178 | 0 | 0.8 | upsloping | 0 | norn
12 0.09 → 0 | 0 | 34 | male | typical ang | 118 | 182 | 0 | left vent hyp... | 174 | 0 | 0.0 | upsloping | 0 | norn
13 0.87 → 1 | 1 | 58 | male | non-anginal | 112 | 230 | 0 | left vent hyp... | 165 | 0 | 2.5 | flat | 1 | reve
14 1.00 → 1 | 1 | 57 | male | asymptomatic | 152 | 274 | 0 | normal | 88 | 1 | 1.2 | flat | 1 | reve
15 1.00 → 1 | 1 | 63 | male | asymptomatic | 130 | 330 | 1 | left vent hyp... | 132 | 1 | 1.8 | upsloping | 3 | reve
16 0.15 → 0 | 0 | 51 | male | non-anginal | 110 | 175 | 0 | normal | 123 | 0 | 0.6 | upsloping | 0 | norn
17 0.00 → 0 | 0 | 45 | female | atypical ang | 130 | 234 | 0 | left vent hyp... | 175 | 0 | 0.6 | flat | 0 | norn
18 0.97 → 1 | 1 | 43 | male | asymptomatic | 132 | 247 | 1 | left vent hyp... | 143 | 1 | 0.1 | flat | ? | reve
19 0.87 → 1 | 1 | 58 | male | asymptomatic | 114 | 318 | 0 | ST-T abnormal | 140 | 0 | 4.4 | downsloping | 3 | fixed
20 0.13 → 0 | 0 | 58 | female | asymptomatic | 130 | 197 | 0 | normal | 131 | 0 | 0.6 | flat | 0 | norn
21 0.20 → 0 | 0 | 67 | female | non-anginal | 115 | 564 | 0 | left vent hyp... | 160 | 0 | 1.6 | flat | 0 | reve
22 0.06 → 0 | 0 | 42 | male | non-anginal | 130 | 180 | 0 | normal | 150 | 0 | 0.0 | upsloping | 0 | norn
23 1.00 → 1 | 1 | 46 | male | asymptomatic | 140 | 311 | 0 | normal | 120 | 1 | 1.8 | flat | 2 | reve
24 0.02 → 0 | 0 | 50 | female | non-anginal | 120 | 219 | 0 | normal | 158 | 0 | 1.6 | flat | 0 | norn
25 0.97 → 1 | 1 | 57 | male | asymptomatic | 165 | 289 | 1 | left vent hyp... | 124 | 0 | 1.0 | flat | 3 | reve
26 0.01 → 0 | 0 | 35 | male | atypical ang | 122 | 192 | 0 | normal | 174 | 0 | 0.0 | upsloping | 0 | norn
27 0.00 → 0 | 0 | 43 | male | asymptomatic | 115 | 303 | 0 | normal | 181 | 0 | 1.2 | flat | 0 | norn
28 0.71 → 1 | 1 | 65 | male | asymptomatic | 110 | 248 | 0 | left vent hyp... | 158 | 0 | 0.6 | upsloping | 2 | fixed
29 0.85 → 1 | 1 | 55 | female | asymptomatic | 180 | 327 | 0 | ST-T abnormal | 117 | 1 | 3.4 | flat | 0 | norn
30 0.91 → 1 | 1 | 39 | male | asymptomatic | 118 | 219 | 0 | normal | 140 | 0 | 1.2 | flat | 0 | reve
31 0.28 → 0 | 0 | 51 | male | non-anginal | 125 | 245 | 1 | left vent hyp... | 166 | 0 | 2.4 | flat | 0 | norn
Show perfomance scores
Target class: (Average over classes)
Model AUC CA F1 Prec Recall MCC
Random Forest 0.997 0.972 0.972 0.972 0.972 0.943
213 | 213 | 213 | 1×213


Learners
Random Forest
Clicking on cells or in headers outputs the corresponding data instances
Ok, got it
Show:
Number of instances
Predicted
0
1
Σ
Actual
0
115
2
117
1
4
92
96
Σ
119
94
213
Output
Predictions
Probabilities
Apply Automatically
Select Correct
Select Misclassified
Clear Selection
1×213
- | 213

Applications

- **Clinical Support**: Helps physicians identify high-risk individuals for heart disease.
- **Pandemic Tracking**: Visual tools monitor virus transmission and recovery waves.
- **Public Health Policy:** Guides decisions on interventions such as lockdowns and vaccination campaigns.
- **Educational Use**: Supports teaching of machine learning and epidemiological trends.
- **Research Catalyst:** Enables discovery of data-driven insights for future studies.

Limitations / Challenges

- Datasets used are static and do not reflect real-time updates.
- Orange is limited in terms of hyperparameter tuning and scalability.
- COVID-19 analysis focused on visualization, not forecasting or prediction.
- Model performance depends on data quality and feature selection.
- Generalizability may be limited across different populations without re-training.

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

1. R. Detrano et al., "UCI Heart Disease Dataset," UCI Machine Learning Repository. [Online].
   Available: https://archive.ics.uci.edu/ml/datasets/heart+disease
2. World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," 2024. [Online].
   Available: https://covid19.who.int/
3. Orange Data Mining, "Documentation," [Online].
   Available: https://orangedatamining.com