Data Science for Medicine Applications

Graduation Project

Egyptian E-Learning University- South Valley (Hurghada)

Faculty of Computers and Information Technology

2023-2024

Supervisors

Dr. Ahmed Ezz

Eng. Sara Gamal

Prepared by

Alaa Ibrahim Amin	20-01771
Alaa Abdelhady Zaky	20-01773
Aya Ali Sayed	20-01775
Amira Mohamed Ahmed	20-01740
Abdelmeseh Talaat Fekry	20-01943
Haidy Mohamed Abu El-Safa	20-01669
Kirolos Elkomos Rofieal	20-00741

Acknowledgments

We would like to thank God Almighty for giving us the confidence and ability to follow up on our project. We are glad that our educational journey has been an amazing experience with a lot of learning curves.

I would like to thank Dr. Ahmed Ezz for being a great doctor and a strong mentor with great experience. He helped us choose the courses and we thank him for his support and continuous follow-up during the process of implementing this project.

How much we extend our sincere thanks and appreciation to our supervisor, Eng. Sara Gamal, for her tremendous efforts and exceptional support that she provided to us during our work on the project and for providing means of assistance at any time during the implementation of our graduation project.

We owe it to you since it helped us shape the problem and provide the solution. We wouldn’t have done it in time without your support, We thank all the staff of EELU University who contributed in many ways to the successful completion of my degree, We repeat our thanks to Dr. Ahmed and his Eng. Sara, who were supportive throughout the entire program. They always responded to private mails and meetings from time to time to follow us up and support us. We are lucky because we got our graduation project and we have professors like Dr. Ahmed, and to his Eng. Sara All thanks for your support.

Abstract

Breast cancer is the most common invasive cancer and the second leading cause of cancer death in women. and regrettably, this rate is increasing every year. One of the aspects of all cancers, including breast cancer, is the recurrence of the disease, which causes painful consequences to the patients.

Moreover, the practical application of data mining in the field of breast cancer can help to provide some necessary information and knowledge required by physicians for accurate prediction of breast cancer recurrence and better decision-making.

The main objective of this study is to compare different data mining algorithms to select the most accurate model for predicting breast cancer recurrence.

This study is cross-sectional and data gathering of this research performed from June 2018 to June 2022 from the official statistics of Ministry of Health and Medical Education and the Egypt Cancer

Research Center for patients with breast cancer who had been followed for a minimum of 8 years from February 2014 to April 2022, including 5471 independent records. After initial pre-processing in dataset and variables, seven new and conventional data mining algorithms have been applied that each one represents one kind of data mining approach.

Results show that the 5 algorithms, one of them possibly could be a helpful tool for the prediction of breast cancer recurrence at the stage of distant recurrence and nonrecurrence, especially in the first to third years.

Also, Deg Malig involvement rate, Her3 values, Tumor size, free or closed tumor class were found to be the most important features in our dataset to predict breast cancer recurrence.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Introduction: -

Breast cancer is one of the most common cancers affecting women worldwide. Despite significant advances in treatment, many patients experience a recurrence of their disease, which can occur locally, regionally, or at distant sites. Understanding the factors that contribute to breast cancer recurrence is essential for improving patient outcomes. This project aims to investigate the predictors, mechanisms, and preventive strategies for breast cancer recurrence. By identifying clinical and genetic factors associated with recurrence, as well as developing effective interventions for early detection and prevention, this project seeks to enhance the quality of life and survival rates for breast cancer survivors.

Patients who were suffering from breast cancer in the past and received treatment have also attacked by recurrence cancer. Those patients need more improvement in the time of treatment. We are motivated to detect the whole recurrence and secondary cancer detection that has been detected in early stages.

In the early stages, the analysis of breast cancer can significantly affect the cure rate. Breast cancer can be diagnosed through diverse techniques, such as blood tests, biopsy, breast test uptake, examination, and such tests. Therefore, early diagnosis of breast cancer is strongly considered by the doctors and the researchers to increase the cure rate and to reduce the mortality rate.

The main objective of this study is to explore or analyze the risk level of breast cancer recurrence, or to predict the probability of 5-year recurrence using data mining models to aid recurrence.

Problem Statement: -

Breast cancer is one of the most common cancers affecting women worldwide. Despite advances in treatment, a significant number of patients experience cancer recurrence, which poses challenges for long-term survival and quality of life. Early and accurate detection of recurrence is critical for timely intervention and improved patient outcomes. However, predicting recurrence is complex due to the heterogeneous nature of the disease and the multitude of factors influencing recurrence risk.

Current methods for predicting breast cancer recurrence often rely on standard clinical and pathological parameters, which may not capture the nuanced interactions of various risk factors. This results in suboptimal prediction accuracy, leading to either over-treatment or under-treatment of patients.

Problem solution: -

To develop a robust, data-driven predictive model using advanced data mining techniques that accurately identifies breast cancer recurrence risk, facilitating early intervention and personalized treatment planning.

Project Objectives: -

The aim of the project is to detect breast cancer recurrence using data mining techniques.

Once treated, breast cancer risks coming back, so identifying the causes and preventing that is essential.

After analyzing several previous articles, we identify tumor recurrence using several machine learning algorithms at an accuracy of more than 95%.

By harnessing the power of advanced algorithms and computational models, healthcare professionals can now efficiently analyze vast amounts of patient data to identify early warning signs of cancer recurrence, ultimately leading to more timely interventions and personalized treatment strategies. This innovative approach not only enhances the accuracy and reliability of recurrence prediction but also enables the development of predictive models that can be tailored to individual patient profiles. Moving forward, continued research and collaboration between data scientists, clinicians, and researchers will be essential to further refine these methodologies and incorporate them into routine clinical practice, ultimately benefiting patients worldwide.

Important Statistics About Breast Cancer: -

The second leading cause of death among women is breast cancer, as it comes directly after lung cancer. Breast cancer considered the most invasive cancer in women, with more than million cases and nearly 600,000 deaths occurring worldwide annually.

Here are some statistics and corresponding graphs related to mental health disorders and their prevalence in different countries: 1. Prevalence of Mental Health Disorders Worldwide: According to the World Health Organization (WHO), approximately round 1 in 8 women in the World will get breast cancer in her lifetime.

In United States: It is diagnosed **every 2 minutes** in the United States, which translates to roughly **292,130 new cases** diagnosed every year.

An estimated **86.4% of** people will survive **5** or more years after being diagnosed with breast cancer. There are estimated to be more than **2.8 million** breast cancer survivors in the **U.S.**

In United Kingdom UK: Around 11,400 women die from breast cancer every year. This is equivalent to 30 deaths every day.

- 48% of deaths from breast cancer are in those aged 75 and over
- Breast cancer is the most common cause of death for women between 35-49 years of age.

In other Countries:

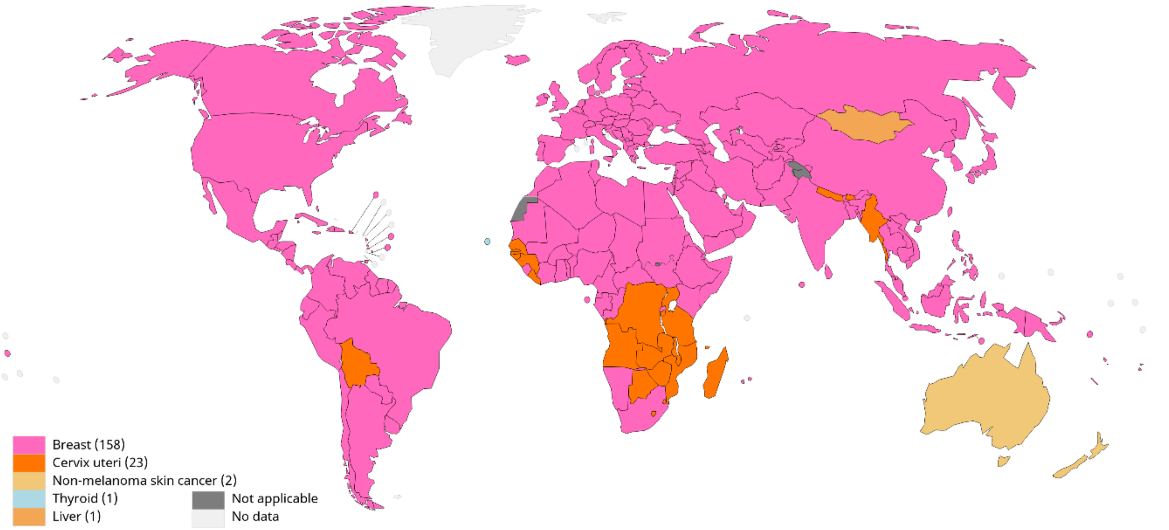
-EU-28 countries, a total of 404,920 new female breast cancer cases was estimated to occur in 2018, corresponding to an age-adjusted standardized rate (ASR) of 144.9/100,000

- In Egypt, breast cancer is the most common malignancy in women, accounting for 38.8% of cancers in this population, with the estimated number of breast cancer cases nearly **22,700** in 2020 and forecasted.

Breast cancer comes in the top of cancer list in Egypt by **40** cases per 100,000 of the population. But there are a great range of deaths in Egypt.

Important Graphs about Breast Cancer: -

Top cancer per country, estimated age-standardized incidence rates (World) in 2020, females, all ages

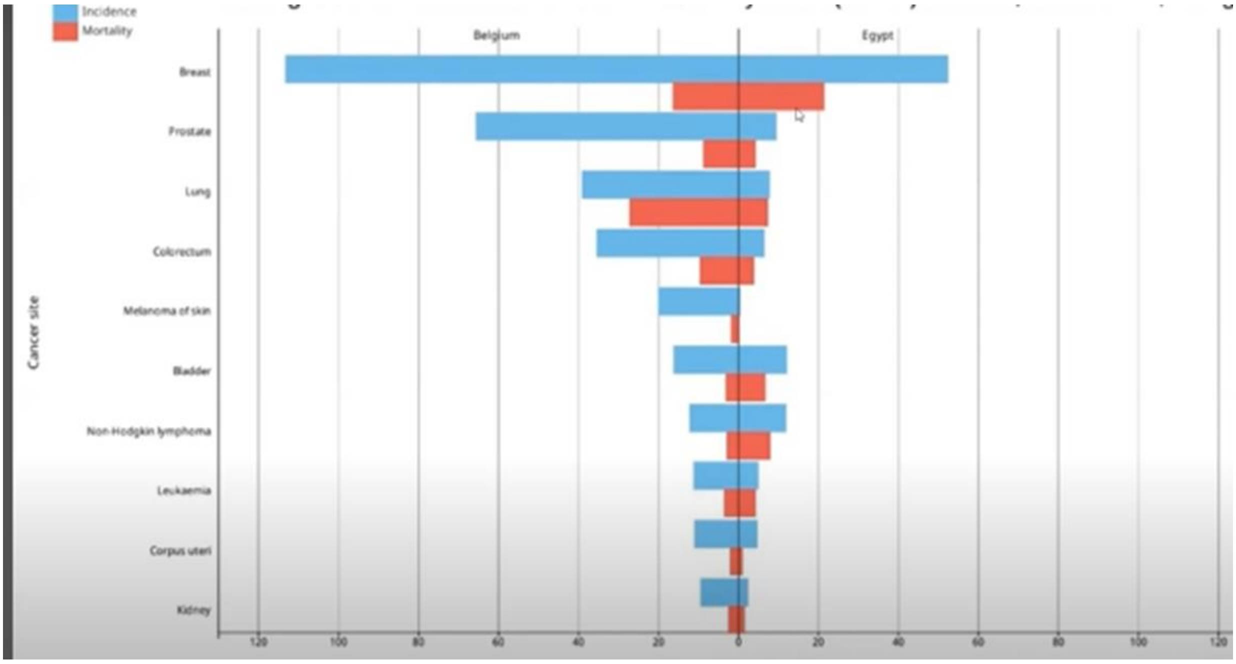


All rights reserved. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization / International Agency for Research on Cancer concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate borderlines for which there may not yet be full agreement.

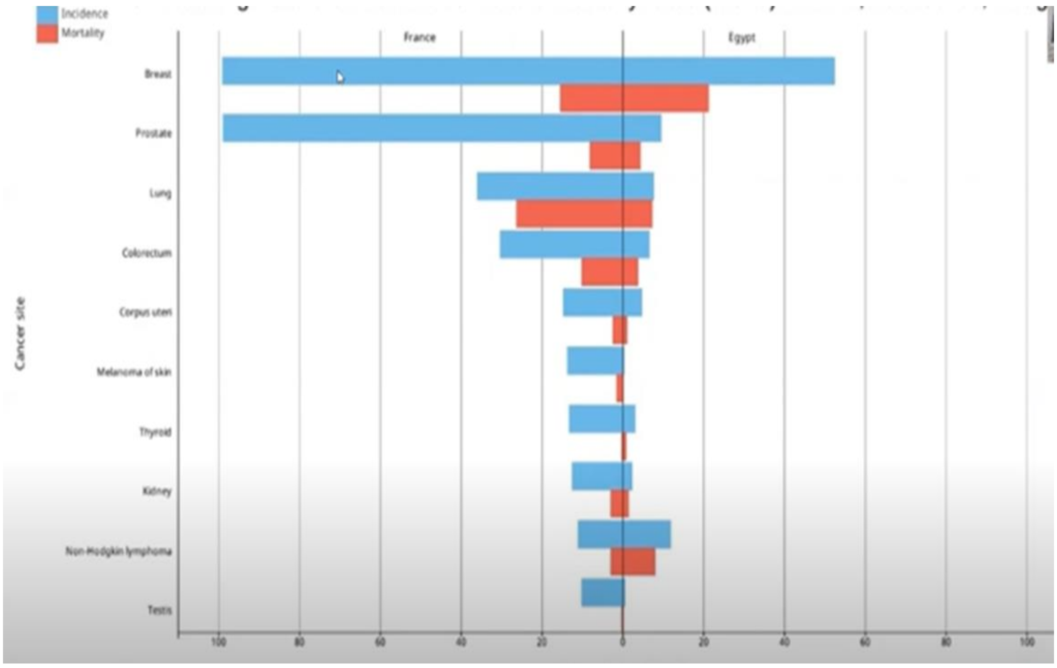
Data source: GLOBOCAN 2020
Graph production: IARC
(http://gco.iarc.fr/today)
World Health Organization

World Health Organization
© International Agency for Research on Cancer 2021

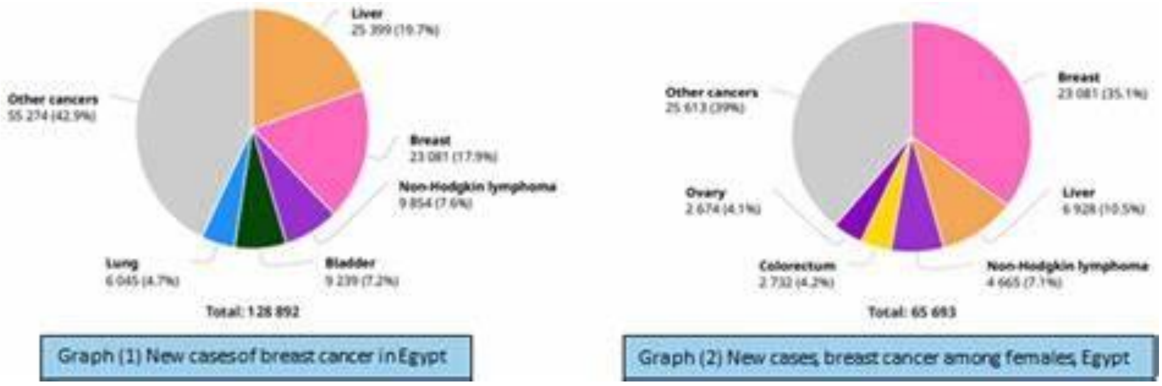
According to World Health Organization 2022, Breast cancer is the most common malignancy diagnosed in women worldwide. Fig (1)



According to World Health Organization 2022, ASR (rate of 100,000) in the world (Belgium 1st country in breast cancer and Egypt) Fig (2)



According to World Health Organization 2022, ASR (rate of 100,000) in the world (France 3rd country in breast cancer and Egypt) **Fig(3)**



Graph (1) New casesof breast cancer in Egypt

Graph (2) New cases, breast cancer among females, Egypt

According to the statistics of cancer in females in Egypt, the breast cancer is the most invasive disease percentage in women. **Fig (4)**

Chapter 2

Background

Current breast cancer (BC) recurrence models do not account for treatment modalities, one of the strongest prognostic factors. This analysis was conducted to apply machine learning (ML) algorithm to identify BC patients at a higher recurrence risk.

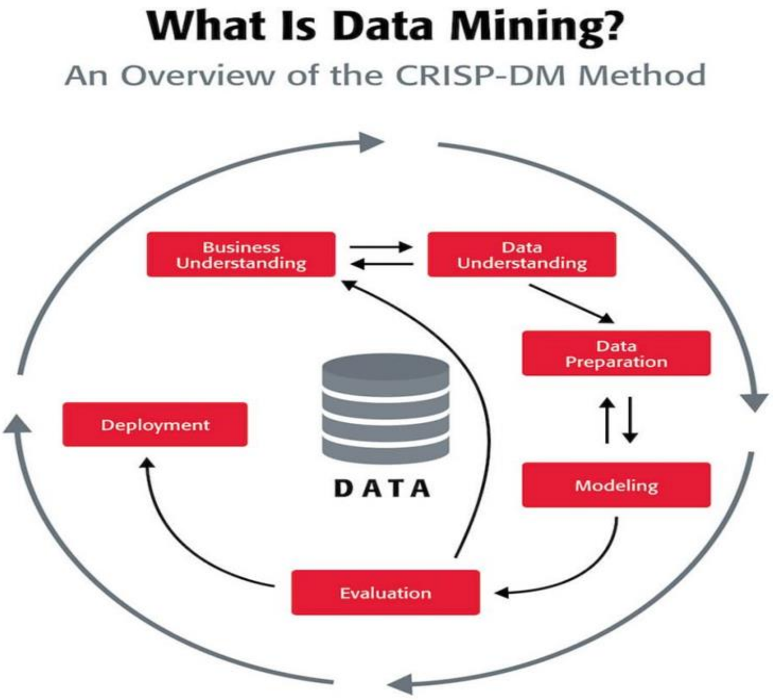
methods: It is based on a downloadable dataset, containing 9 independent (socio-demographic, tumor and treatment-related) and a dependent (recurrence) variable(s) using different types of machine learning models.

Results: 277 patients (recurrence (n=81)) were included. In univariate analysis, tumor size ($p=0.002$), invasive nodes number ($p<0.001$), node capsule ($p<0.05$), degree of malignancy ($p<0.001$), and irradiation ($p<0.001$), were associated with recurrence were significant in a modeling.

Data mining:

Data mining is the process of sorting through large data sets to identify patterns and relationships that can help solve business problems through data analysis. Data mining techniques and tools help enterprises to predict future trends and make more informed business decisions.

Data mining is a key part of data analytics and one of the core disciplines in data science, which uses advanced analytics techniques to find useful information in data sets. At a more granular level, data mining is a step in the knowledge discovery in databases (KDD) process, a data science methodology for gathering, processing and analyzing data. Data mining and KDD are sometimes referred to interchangeably, but they're more commonly seen as distinct things.



Fig(5)

Data cleaning

Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.[1] Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting or a data quality firewall.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleaning differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

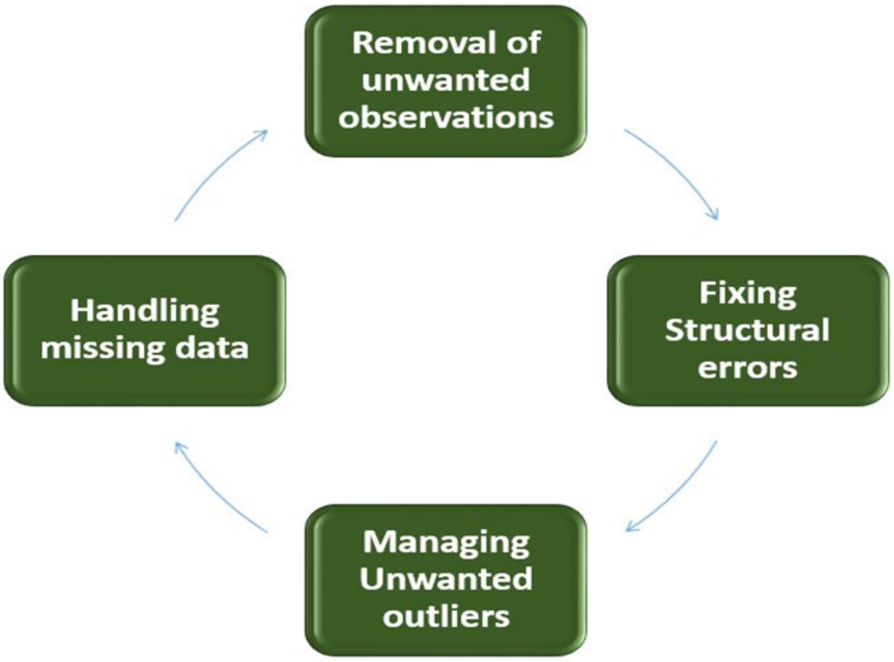


Fig (6)

Data Preprocessing

Data processing plays a critical role in breast cancer recurrence detection for several reasons:

1. Improving Data Quality

Raw data collected from medical records, imaging, and other sources often contain noise, missing values, and inconsistencies. Data processing techniques such as data cleaning, normalization, and imputation are essential to enhance the quality of data. High-quality data is crucial for building accurate and reliable predictive models.

2. Feature Extraction and Selection

Data processing helps in identifying and extracting relevant features that are critical for predicting breast cancer recurrence. Feature extraction involves transforming raw data into meaningful features, while feature selection involves choosing the most relevant features from the extracted ones. These steps are vital because they:

- Reduce dimensionality, which simplifies the model and reduces the risk of overfitting.
- Enhance the interpretability of the model by focusing on the most important factors influencing recurrence.
- Improve model performance by eliminating irrelevant or redundant features.

3. Handling Imbalanced Data

Breast cancer recurrence datasets are often imbalanced, with fewer cases of recurrence compared to non-recurrence. Data processing techniques such as resampling (oversampling the minority class or under sampling the majority class) and synthetic data generation (using techniques like SMOTE) are essential to address this imbalance. Properly balanced data ensures that the predictive model does not become biased towards the majority class.

4. Enhancing Model Performance

Data processing steps such as scaling, normalization, and transformation of features can significantly impact the performance of machine learning algorithms. For instance:

- **Normalization** ensures that features have a similar scale, which is particularly important for distance-based algorithms like k-nearest neighbors (KNN).
- **Scaling** helps in speeding up the convergence of gradient descent-based algorithms.

5. Reducing Overfitting

Overfitting occurs when a model learns the noise in the training data rather than the actual patterns. Data processing techniques such as feature selection, dimensionality reduction (e.g., PCA), and data augmentation help in reducing overfitting by simplifying the model and making it more generalizable to new, unseen data.

6. Improving Interpretability

Processed data, with clearly defined and relevant features, makes it easier to interpret the model's predictions. Clinicians and researchers can better understand the factors contributing to breast cancer recurrence, which aids in clinical decision-making and patient management.

7. Facilitating Data Integration

Breast cancer recurrence detection often requires integrating data from multiple sources, such as genomic data, imaging data, and clinical records. Data processing techniques enable the harmonization and integration of these heterogeneous data sources, providing a comprehensive dataset for model training and evaluation.

8. Enabling Real-time Analysis

Efficient data processing pipelines are essential for enabling real-time or near-real-time analysis, which is crucial for timely decision-making in clinical settings. Automated data processing workflows ensure that new data can be quickly and accurately incorporated into predictive models.

So, data processing is foundational in breast cancer recurrence detection as it ensures data quality, facilitates feature extraction and selection, handles data imbalances, enhances model performance, reduces overfitting, improves interpretability, enables data integration, and supports real-time analysis. Without proper data processing, the accuracy, reliability, and utility of predictive models for breast cancer recurrence would be significantly compromised.

Machine Learning

Machine learning (ML) is defined as a discipline of artificial intelligence (AI) that provides machines with the ability to automatically learn from data and past experiences to identify patterns and make predictions with minimal human intervention. This article explains the fundamentals of machine learning, its types, and the top five applications. It also shares the top 10 machine learning trends.

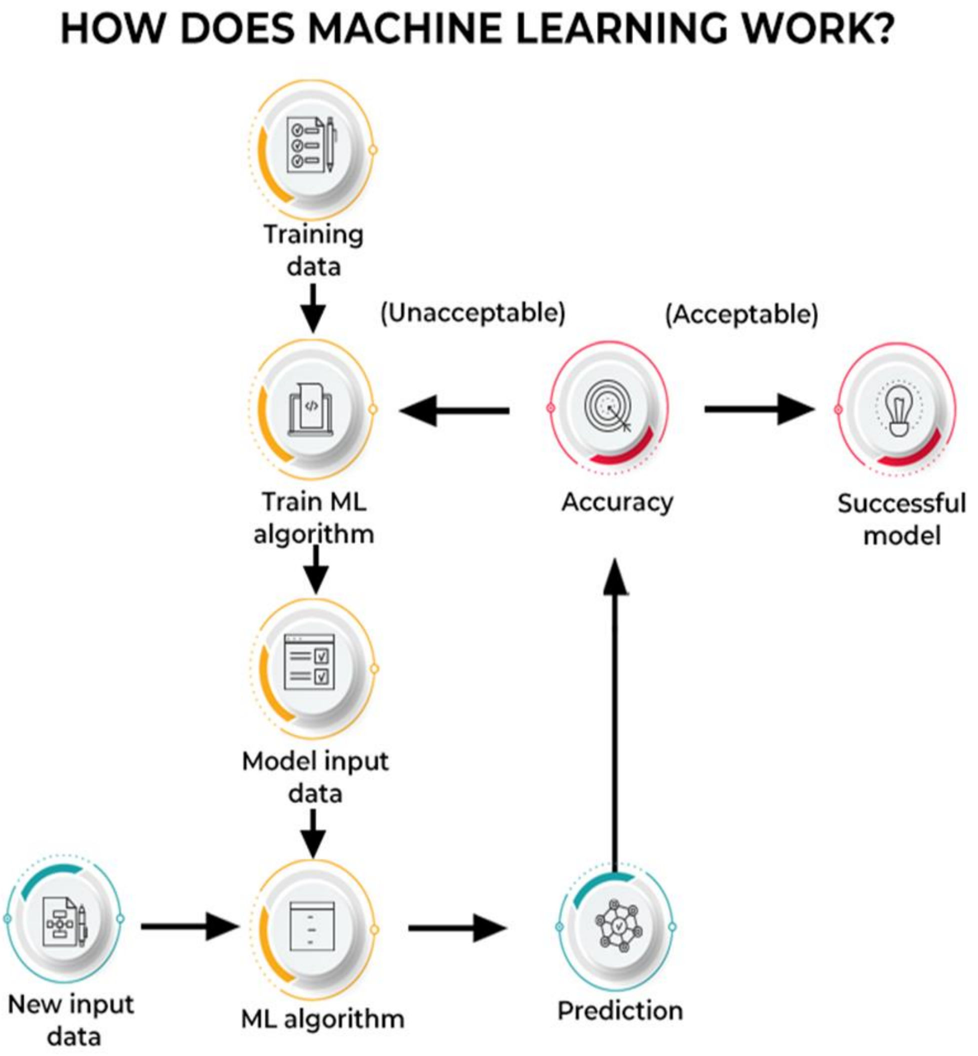


Fig (7)

Types of Machine Learning

Machine learning algorithms can be trained in many ways, with each method having its pros and cons. Based on these methods and ways of learning, machine learning is broadly categorized into four main types:

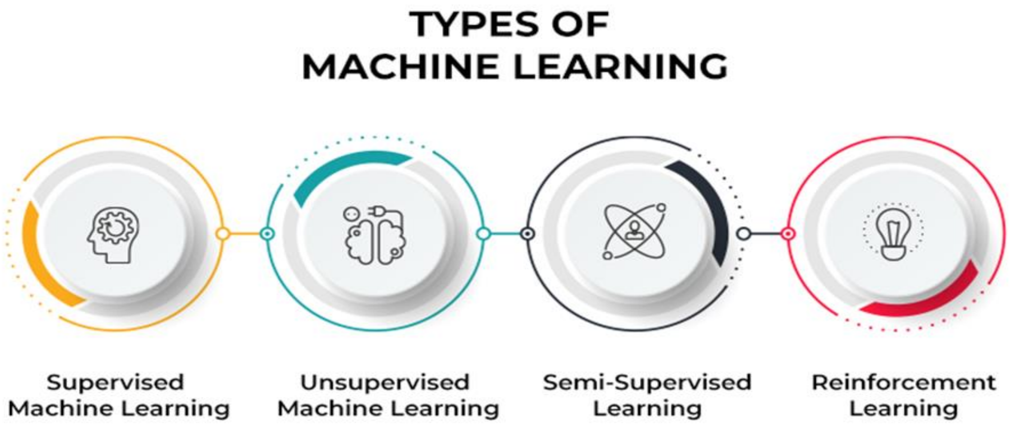


Fig (8)

1. Supervised machine learning

This type of ML involves supervision, where machines are trained on labelled datasets and enabled to predict outputs based on the provided training. The labelled dataset specifies that some input and output parameters are already mapped. Hence, the machine is trained with the input and corresponding output. A device is made to predict the outcome using the test dataset in subsequent phases.

The primary objective of the supervised learning technique is to map the input variable (a) with the output variable (b). Supervised machine learning is further classified into two broad categories:

Classification: These refer to algorithms that address classification problems where the output variable is categorical; for example, yes or no, true or false, male or female, etc. Real-world applications of this category are evident in spam detection and email filtering.

Some known classification algorithms include the Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, and Support Vector Machine Algorithm.

Regression: Regression algorithms handle regression problems where input and output variables have a linear relationship. These are known to predict continuous output variables. Examples include medical prediction, Diseases detection, etc.

2. Unsupervised machine learning

Unsupervised learning refers to a learning technique that's devoid of supervision. Here, the machine is trained using an unlabeled dataset and is enabled to predict the output without any supervision. An unsupervised learning algorithm aims to group the unsorted dataset based on the input's similarities, differences, and patterns.

unsupervised machine learning is further classified into two types:

Clustering: The clustering technique refers to grouping objects into clusters based on parameters such as similarities or differences between objects. For example, grouping customers by the products they purchase.

Association: Association learning refers to identifying typical relations between the variables of a large dataset. It determines the dependency of various data items and maps associated variables.

Typical applications include web usage mining and market data analysis.

3. Semi-supervised learning

Semi-supervised learning comprises characteristics of both supervised and unsupervised machine learning. It uses the combination of labeled and unlabeled datasets to train its algorithms. Using both types of datasets, semi-supervised learning overcomes the drawbacks of the options mentioned above.

4. Reinforcement learning

Reinforcement learning is a feedback-based process. Here, the AI component automatically takes stock of its surroundings by the hit & trial method, acts, learns from experiences, and improves performance. The component is rewarded for each good action and penalized for every wrong move. Thus, the reinforcement learning component aims to maximize the rewards by performing good actions.

Chapter 3

Project Development:

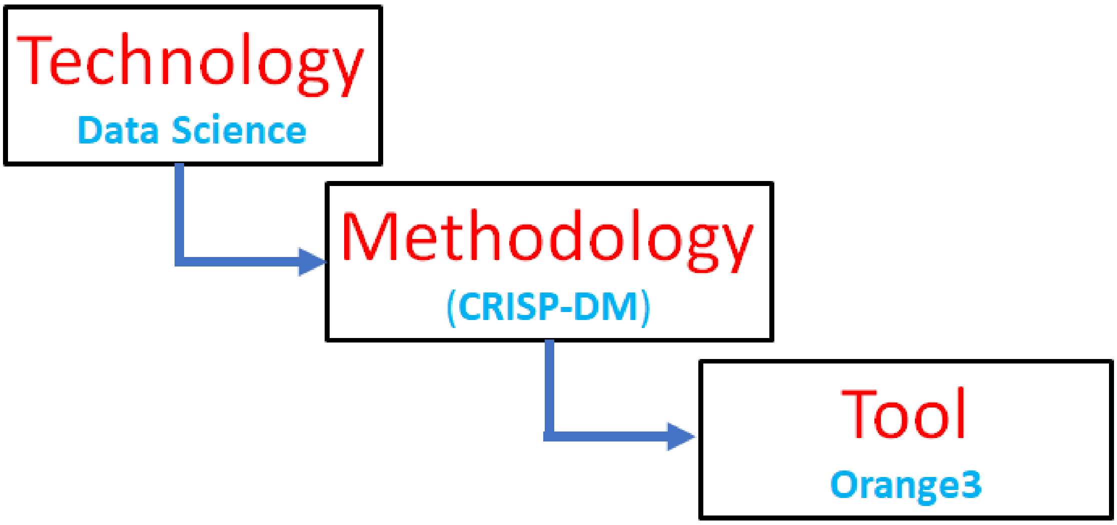
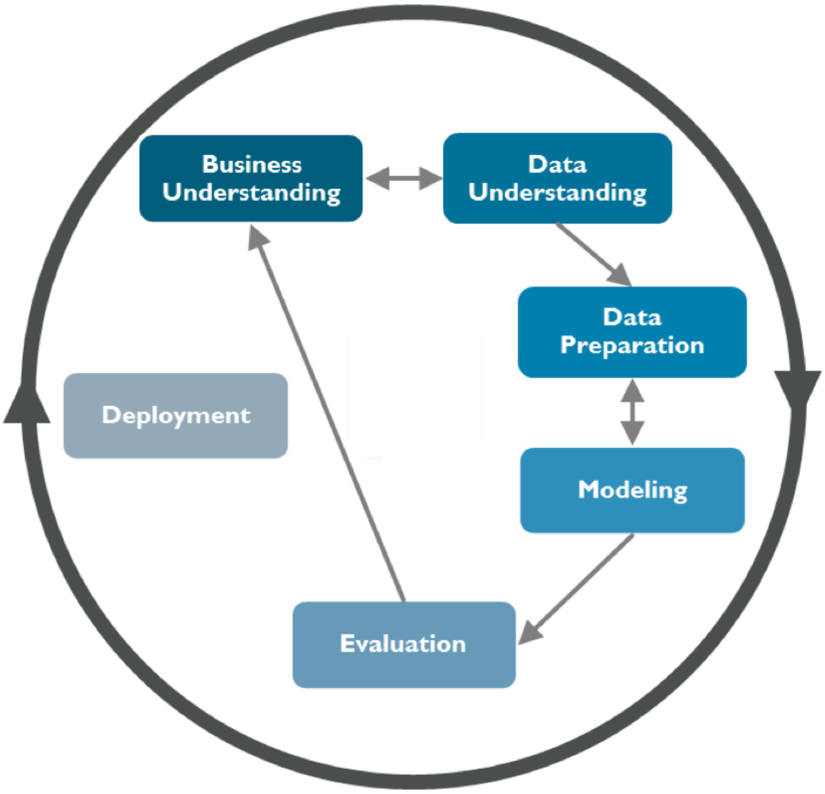


Fig (9)

We are referring to a combination of methodologies, tools, and technologies for project development in data science, with a focus on the CRISP-DM methodology and Orange3 tool.

Combining CRISP-DM with Orange3 can streamline your data science project development process. You can use Orange3 to perform tasks such as data exploration, preprocessing, modeling, and evaluation, while following the structured approach outlined in CRISP-DM.

CRISP-DM:



CRISP-DM, or Cross-Industry Standard Process for Data Mining, is a widely accepted framework that provides a structured approach to data mining and analytics projects. It offers a comprehensive methodology for guiding organizations through the entire lifecycle of a data science project, from understanding business objectives to deploying data-driven solutions. CRISP-DM emphasizes iterative development, allowing for flexibility and adaptation as insights are gained and requirements evolve. By following the CRISP-DM methodology, organizations can effectively manage data mining projects, maximize the value of their data assets, and achieve actionable insights to drive business decisions.

Business Understanding:

Breast cancer is the most common cancer affecting females worldwide. Breast cancer recurrence or nonrecurrence prediction is challenging and a complex research task. Existing approaches engage statistical methods or supervised machine learning to assess/predict the recurrence or non-recurrence of breast cancer.

Data Understanding:

Collect and explore relevant breast cancer data. Understand the characteristics of the dataset.

Data Field	**Description:**	**Data Values:**
age:	Represents the age of the patient. (numeric), This is a nominal attribute that	Categorized into age ranges, such as "20-29, 30-39, 40-49, 50-59, 60-69, 70-79”

	can be used to compare the incidence of breast cancer among different age groups.	
menopause:	Indicates the menopausal status of the patient. , This is also a nominal attribute that can be used to study the effect of menopause on breast cancer risk.	Includes either premeno (premenopausal), ge40 (greater than or equal to 40 years old at menopause), or lt40 (less than 40 years old at menopause).
tumour-size:	Describes the size of the tumor. This is an ordinal attribute that can be used to measure the severity of the tumor and its potential to spread.	tumor size range such as 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54

inv-nodes:	This column indicates the number of axillary lymph nodes that are involved by the tumor	Categorized into age ranges, such as 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 24-26
node-caps:	Indicates whether the tumor has affected the surrounding lymph nodes' capsules. This is a binary attribute that can be used to determine the invasiveness of the tumor and its likelihood of metastasis.	Binary, "yes" or "no."
deg-malig:	Denotes the degree of malignancy or aggressiveness of the tumor.	either 3 (high), 1 (low), or 2 (medium)

breast:	This column indicates which breast is affected.	either right or left.
breast-quad:	This column indicates which quadrant of the breast is affected by the tumor	either left_up (upper left), left_low (lower left), central (central), right_up (upper right), or right_low (lower right).
irradiat:	This column indicates whether the patient received radiation therapy as part of the treatment	either yes or no
Class:	This column indicates the recurrence status of the patient	either recurrence-events (the patient had a recurrence of breast cancer) or no-recurrence-events (the patient did not have a recurrence of breast cancer).

Load Data:

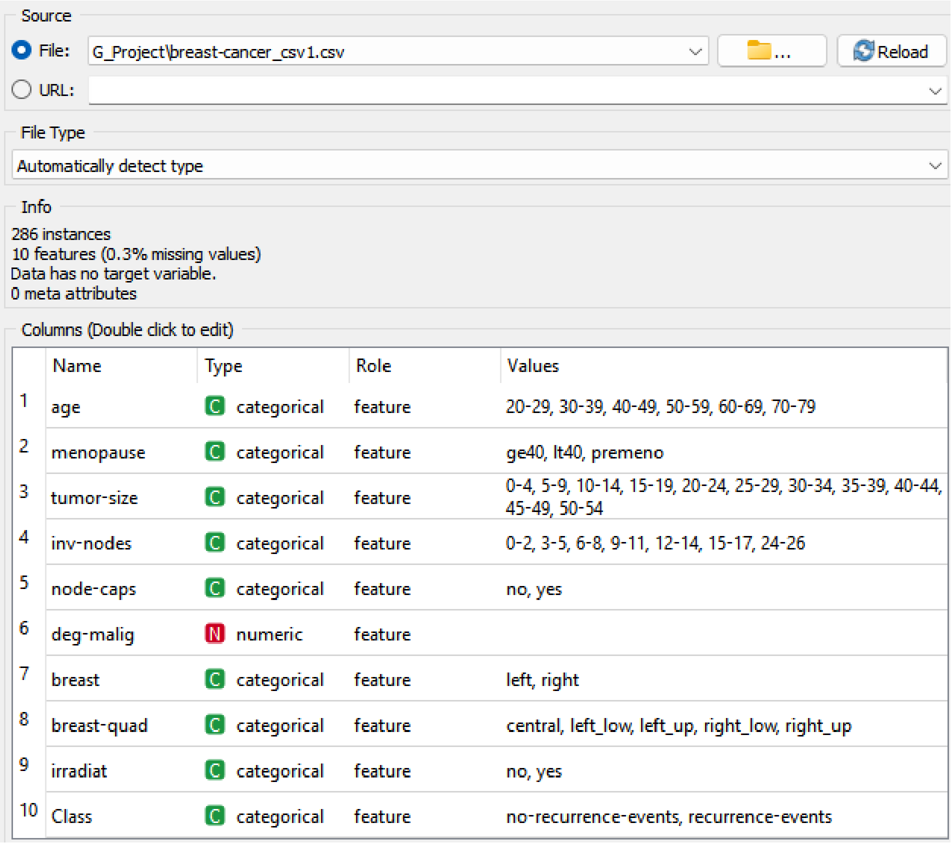
File Reads attribute-value data from an input file.

Outputs

Data: dataset for Breast cancer from the file

The **File** widget reads the input data file (data table with data instances) and sends the dataset to its output channel. The history of most recently opened files is maintained in the widget. The widget also includes a directory with sample datasets that come pre-installed with Orange.

The widget reads data from Excel (.xlsx), simple tab-delimited (.txt), comma-separated files (.csv) or URLs. For other formats see Other Formats section below.



Source

● File: G_Project\breast-cancer_csv1.csv

○ URL:

File Type

Automatically detect type

Info

286 instances
10 features (0.3% missing values)
Data has no target variable.
0 meta attributes

Columns (Double click to edit)

	Name	Type	Role	Values
1	age	C categorical	feature	20-29, 30-39, 40-49, 50-59, 60-69, 70-79
2	menopause	C categorical	feature	ge40, lt40, premeno
3	tumor-size	C categorical	feature	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54
4	inv-nodes	C categorical	feature	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 24-26
5	node-caps	C categorical	feature	no, yes
6	deg-malig	N numeric	feature	
7	breast	C categorical	feature	left, right
8	breast-quad	C categorical	feature	central, left_low, left_up, right_low, right_up
9	irradiat	C categorical	feature	no, yes
10	Class	C categorical	feature	no-recurrence-events, recurrence-events

Browse through previously opened data files, or load any of the

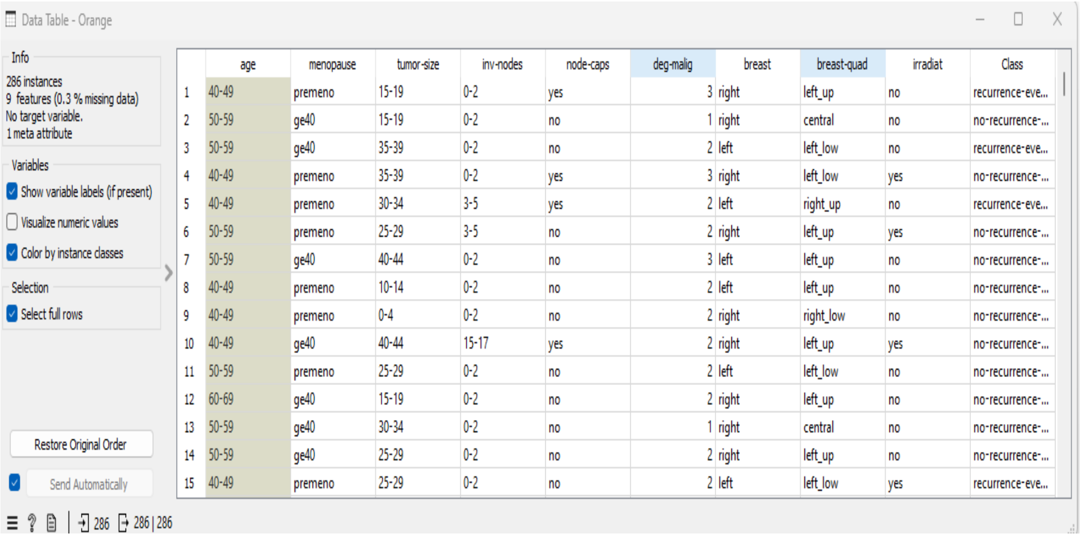
1. sample ones.
2. Browse for a data file.
3. Reloads currently selected data file.
4. Insert data from URL addresses, including data from Google Sheets.
5. Information on the loaded dataset: dataset size, number and types of data features.

6. Additional information on the features in the dataset. Features can be edited by double-clicking on them. The user can change the attribute names, select the type of variable per each attribute (Continuous, Nominal, String, Datetime), and choose how to further define the attributes (as Features, Targets or Meta). The user can also decide to ignore an attribute.
7. Browse documentation datasets.
8. Produce a report.

Data Exploration:

Data Tab

The Data Table widget receives one or more datasets in its input and presents them as a spreadsheet. Data instances may be sorted by attribute values. The widget also supports manual selection of data instances.



Data Table - Orange

Info

286 instances
9 features (0.3 % missing data)
No target variable.
1 meta attribute

Variables

☑ Show variable labels (if present)

□ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

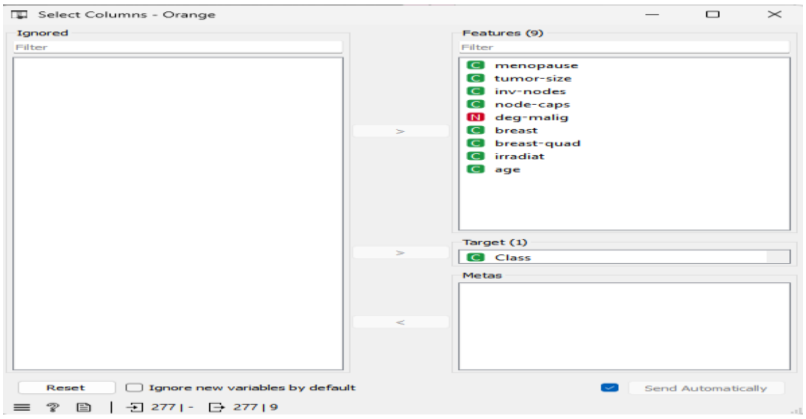
	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat	Class
1	40-49	premeno	15-19	0-2	yes	3	right	left_up	no	recurrence-eve...
2	50-59	ge40	15-19	0-2	no	1	right	central	no	no-recurrence-...
3	50-59	ge40	35-39	0-2	no	2	left	left_low	no	recurrence-eve...
4	40-49	premeno	35-39	0-2	yes	3	right	left_low	yes	no-recurrence-...
5	40-49	premeno	30-34	3-5	yes	2	left	right_up	no	recurrence-eve...
6	50-59	premeno	25-29	3-5	no	2	right	left_up	yes	no-recurrence-...
7	50-59	ge40	40-44	0-2	no	3	left	left_up	no	no-recurrence-...
8	40-49	premeno	10-14	0-2	no	2	left	left_up	no	no-recurrence-...
9	40-49	premeno	0-4	0-2	no	2	right	right_low	no	no-recurrence-...
10	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes	no-recurrence-...
11	50-59	premeno	25-29	0-2	no	2	left	left_low	no	no-recurrence-...
12	60-69	ge40	15-19	0-2	no	2	right	left_up	no	no-recurrence-...
13	50-59	ge40	30-34	0-2	no	1	right	central	no	no-recurrence-...
14	50-59	ge40	25-29	0-2	no	2	right	left_up	no	no-recurrence-...
15	40-49	premeno	25-29	0-2	no	2	left	left_low	yes	recurrence-eve...

1. The name of the dataset (usually the input data file). Data instances are in rows and their attribute values in columns. In this example, the dataset is sorted by the attribute "sepal length".
2. Info on current dataset size and number and types of attributes
3. Values of continuous attributes can be visualized with bars; colours can be attributed to different classes.
4. Data instances (rows) can be selected and sent to the widget's output channel.
5. Use the *Restore Original Order* button to reorder data instances after attribute-based sorting.
6. Produce a report.
7. While auto-send is on, all changes will be automatically communicated to other widgets. Otherwise, press *Send Selected Rows*.

Select column:

The Select Columns widget is used to manually compose your data domain. The user can decide which attributes will be used and how. Orange distinguishes between ordinary attributes, (optional) class attributes and meta-attributes. For instance, for building a classification model, the domain would be composed of a set of attributes and a discrete class attribute. Meta attributes are not used in modelling, but several widgets can use them as instance labels.

Orange attributes have a type and are either discrete, continuous or a character string. The attribute type is marked with a symbol appearing before the name of the attribute (D, C, S, respectively).



1. Left-out data attributes that will not be in the output data file
2. Data attributes in the new data file
3. Target variable. If none, the new dataset will be without a target variable.
4. Meta attributes of the new data file. These attributes are included in the dataset but are, for most methods, not considered in the analysis.
5. Produce a report.
6. Reset the domain composition to that of the input data file.
7. Tick if you wish to auto-apply changes of the data domain.
8. Apply changes of the data domain and send the new data file to the output channel of the widget.

Data Preparation:

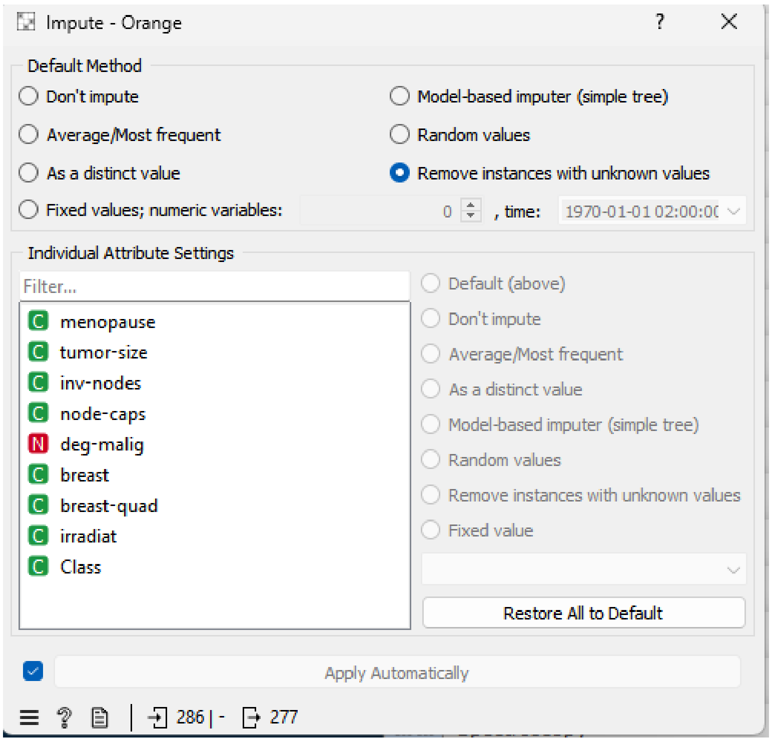
- Clean and preprocess the breast cancer data.
- Handle missing values and perform transformations as needed.
- Ensure the dataset is ready for analysis.

Impute Missing Values:

Impute

Replaces unknown values in the data.

Some of Orange's algorithms and visualizations cannot handle unknown values in the data. This widget does what statisticians call imputation: it substitutes missing values by values either computed from the data or set by the user. The default imputation is (1-NN).



Impute - Orange

Default Method

○ Don't impute
○ Model-based imputer (simple tree)
○ Average/Most frequent
○ Random values
○ As a distinct value
● Remove instances with unknown values
○ Fixed values; numeric variables: 0 , time: 1970-01-01 02:00:0(

Individual Attribute Settings

Filter...

C menopause
C tumor-size
C inv-nodes
C node-caps
N deg-malig
C breast
C breast-quad
C irradiat
C Class

○ 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

286 | - 277

1. In the top-most box, Default method, the user can specify a general imputation technique for all attributes.

Don't Impute does nothing with the missing values.

Average/Most frequent uses the average value (for continuous attributes) or the most common value (for discrete attributes).

As a distinct value creates new values to substitute the missing ones.

Remove examples with missing values removes the example containing missing values. This check also applies to the class attribute if Impute class values are checked.

2. It is possible to specify individual treatment for each attribute, which overrides the default treatment set. One can also specify a manually defined value used for

imputation. In the screenshot, we decided not to impute the values of "normalized-losses" and "make", the missing values of "aspiration" will be replaced by random values, while the missing values of "body-style" and "drive-wheels" are replaced by "hatchback" and "fwd.”, respectively. If the values of "length", "width" or "height" are missing, the example is discarded. Values of all other attributes use the default method set above (model-based imputer, in our case).

3. The imputation methods for individual attributes are the same as default methods.

4. Restore All to Default resets the individual attribute treatments to default.

5. Produce a report.

6. All changes are committed immediately if Apply automatically is checked. Otherwise, apply needs to be ticked to apply any new settings

Chapter 4

Univariable analysis: Table 1 showed the univariate analysis of different variables among 277 BC patients. The age (p=0.2169), menopausal status (p=0.2178), breast (p=0.5788) and breast quadrant (p=0.5073) were not significantly different among patients with and without recurrence. However, the tumor size (p=0.002), number of invasive nodes (p<0.001), node capsule (p<0.001), degree of malignancy (p<0.001 and irradiation (p<0.001), were significant factors influencing the recurrence of BC

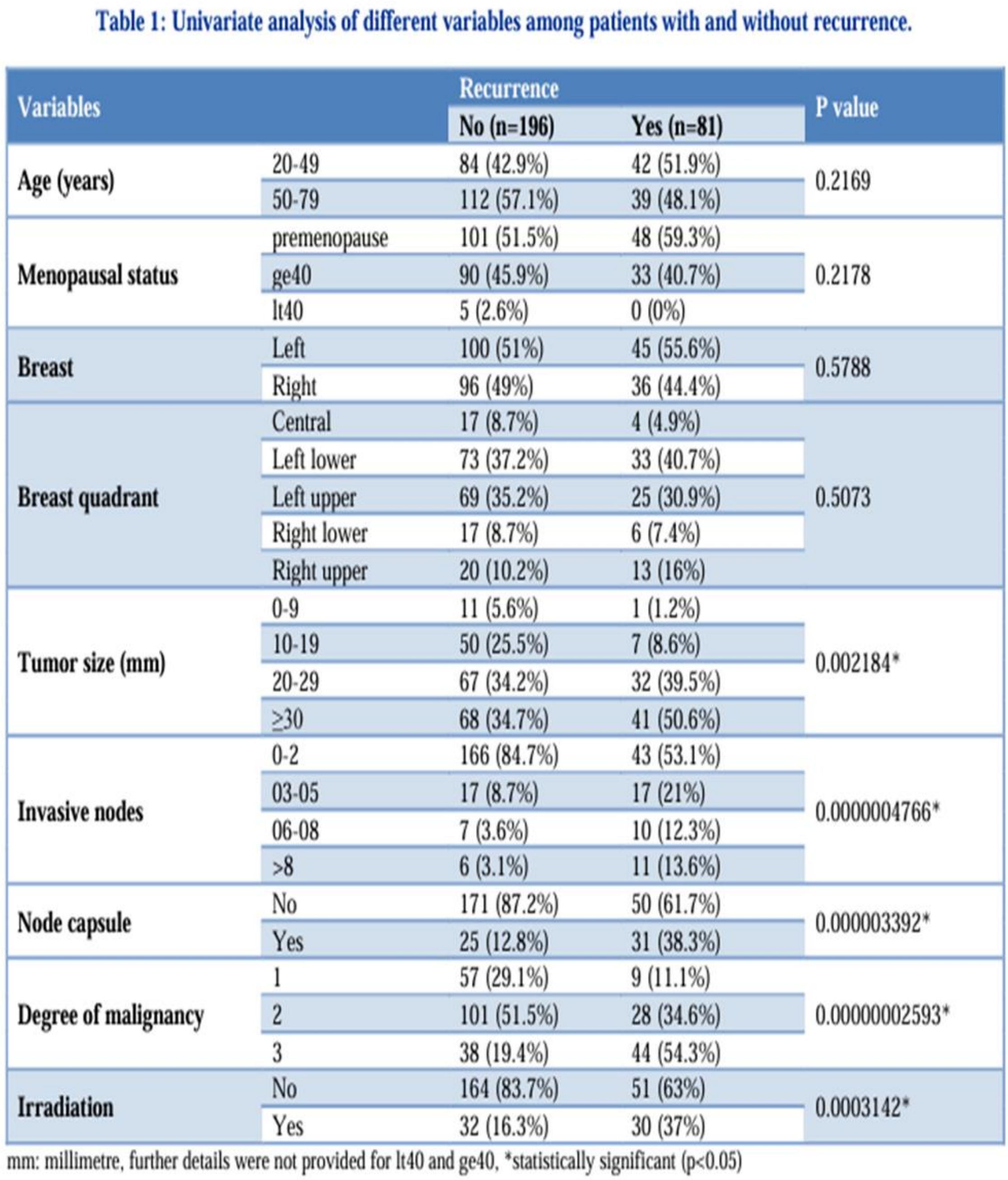


Table 1: Univariate analysis of different variables among patients with and without recurrence.

Variables		Recurrence		P value
		No (n=196)	Yes (n=81)	
Age (years)	20-49	84 (42.9%)	42 (51.9%)	0.2169
	50-79	112 (57.1%)	39 (48.1%)	
Menopausal status	premenopause	101 (51.5%)	48 (59.3%)	0.2178
	ge40	90 (45.9%)	33 (40.7%)	
	lt40	5 (2.6%)	0 (0%)	
Breast	Left	100 (51%)	45 (55.6%)	0.5788
	Right	96 (49%)	36 (44.4%)	
Breast quadrant	Central	17 (8.7%)	4 (4.9%)	0.5073
	Left lower	73 (37.2%)	33 (40.7%)	
	Left upper	69 (35.2%)	25 (30.9%)	
	Right lower	17 (8.7%)	6 (7.4%)	
	Right upper	20 (10.2%)	13 (16%)	
Tumor size (mm)	0-9	11 (5.6%)	1 (1.2%)	0.002184*
	10-19	50 (25.5%)	7 (8.6%)	
	20-29	67 (34.2%)	32 (39.5%)	
	≥30	68 (34.7%)	41 (50.6%)	
Invasive nodes	0-2	166 (84.7%)	43 (53.1%)	0.0000004766*
	03-05	17 (8.7%)	17 (21%)	
	06-08	7 (3.6%)	10 (12.3%)	
	>8	6 (3.1%)	11 (13.6%)	
Node capsule	No	171 (87.2%)	50 (61.7%)	0.000003392*
	Yes	25 (12.8%)	31 (38.3%)	
Degree of malignancy	1	57 (29.1%)	9 (11.1%)	0.00000002593*
	2	101 (51.5%)	28 (34.6%)	
	3	38 (19.4%)	44 (54.3%)	
Irradiation	No	164 (83.7%)	51 (63%)	0.0003142*
	Yes	32 (16.3%)	30 (37%)	

mm: millimetre, further details were not provided for lt40 and ge40, *statistically significant (p<0.05)

Fig (10)

Pivot Table

- Reshape data table based on column values.

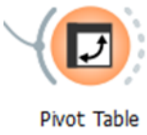
Inputs

- Data: input data set

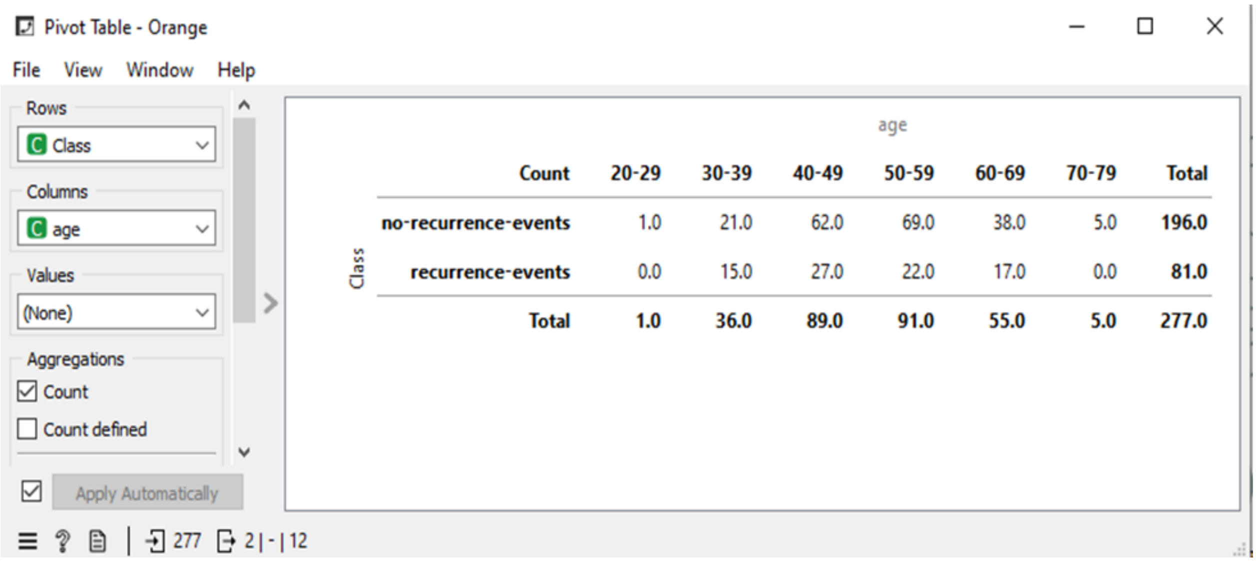
Outputs

- Pivot Table: contingency matrix as shown in the widget
- Filtered Data: subset selected from the plot
- Grouped Data: aggregates over groups defined by row values.

Pivot Table summarizes the data of a more extensive table into a table of statistics. Statistics can include sums, averages, counts, etc. The widget also allows selecting a subset from the table and grouping by row values, which must be a discrete variable. Data with only numeric variables cannot be displayed in the table.



Pivot table icon in orange.

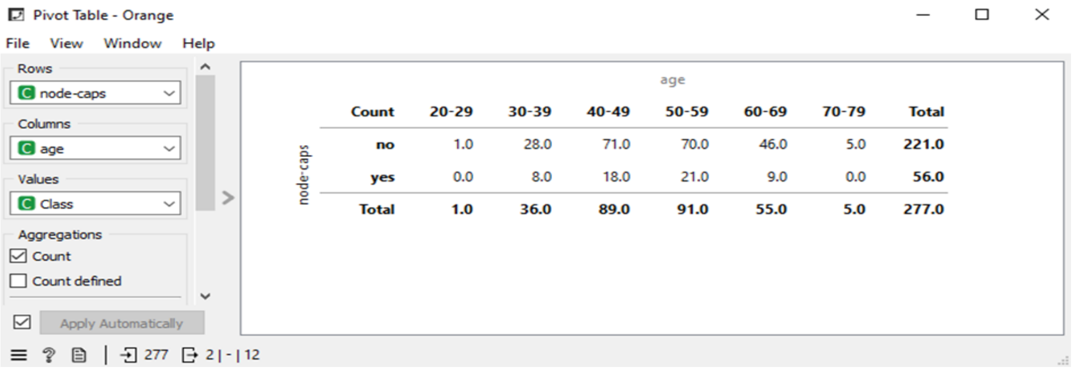


		age						
	Count	20-29	30-39	40-49	50-59	60-69	70-79	Total
Class	no-recurrence-events	1.0	21.0	62.0	69.0	38.0	5.0	196.0
	recurrence-events	0.0	15.0	27.0	22.0	17.0	0.0	81.0
	Total	1.0	36.0	89.0	91.0	55.0	5.0	277.0

1. Discrete or numeric variable used for row values. Numeric variables are considered as integers.
2. Discrete variable used for column values. Variable values will appear as columns in the table.
3. Values used for aggregation. Aggregated values will appear as cells in the table.
4. Aggregation methods:
   - For any variable type:
     - Count: number of instances with the given row and column value.
     - Count defined: number of instances where the aggregation value is defined.
   - For numeric variables:
     - Sum: sum of values.

  - Mean: average of values.
  - Mode: most frequent value of the subset.
  - Min: smallest value.
  - Max: highest value.
  - Median: middle value.
  - Var: variance of the subset.
- For discrete variables:
  - Majority: most frequent value of the subset.

5. Tick the box on the left to automatically output any changes. Alternatively, press Apply.



Rows: node-caps

Columns: age

Values: Class

Aggregations

☑ Count

☐ Count defined

☑ Apply Automatically

node-caps	Count	20-29	30-39	40-49	50-59	60-69	70-79	Total
	no	1.0	28.0	71.0	70.0	46.0	5.0	221.0
	yes	0.0	8.0	18.0	21.0	9.0	0.0	56.0
	Total	1.0	36.0	89.0	91.0	55.0	5.0	277.0

Feature Statistics

Show basic statistics for data features.

Inputs

- Data: input data

Outputs

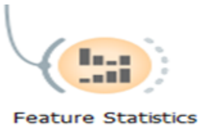
- Reduced data: table containing only selected features

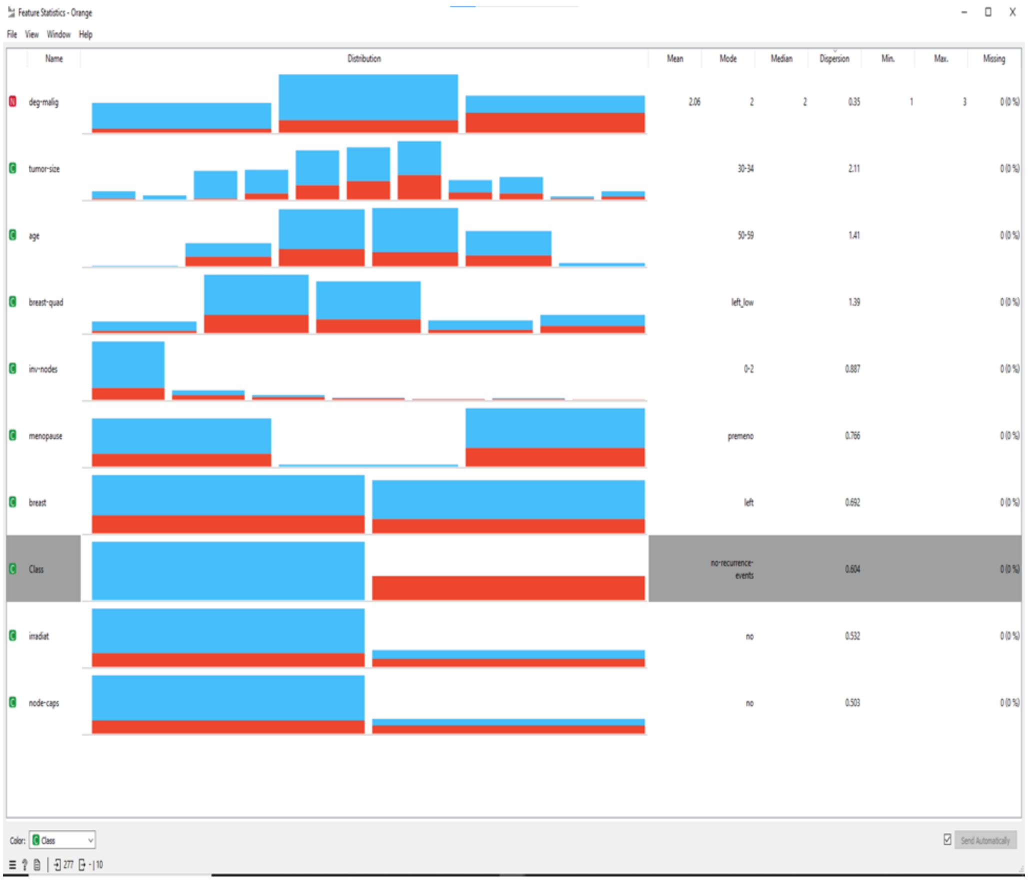
- Statistics: table containing statistics of the selected features

The **Feature Statistics** widget provides a quick way to inspect and find interesting features in each data set.

The Feature Statistics widget on the BC recurrence data set.

1. The feature type - can be categorical and numeric.
2. The name of the feature.
3. The histogram shows the distribution of feature's values. Values of numeric features are split into bins.
4. Further columns show different statistics. Mean, minimal and maximal values are computed only for numeric features. Mode shows the most common value for numeric or categorical feature. Dispersion shows coefficient of variation for numeric features, and entropy for categorical.
5. The bars in the histogram can be further split by value of another variable. The default choice is the target variable, but the user can change this to an arbitrary feature or none.





Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N deg-malig		2.06	2	2	0.35	1	3	0 (0 %)
C tumor-size			30-34		2.11			0 (0 %)
C age			50-59		1.41			0 (0 %)
C breast-quad			left_low		1.39			0 (0 %)
C inv-nodes			0-2		0.887			0 (0 %)
C menopause			premeno		0.766			0 (0 %)
C breast			left		0.692			0 (0 %)
C Class			no-recurrence-events		0.604			0 (0 %)
C irradiat			no		0.532			0 (0 %)
C node-caps			no		0.503			0 (0 %)

Color: C Class

☑ Send Automatically

Visualization

Visualization involves creating visual representations of data or information to help understand and communicate it more effectively. These can include graphs, charts, maps, infographics, and more. Here are some key aspects of visualization:

Visualization offers numerous benefits across various domains. Here are some key advantages:

Enhanced Understanding

1. **Simplifies Complex Data**: Visualization transforms complex data into easily understandable visual formats, making it easier to grasp intricate patterns and relationships.

2. **Quick Insights**: Visual representations enable faster comprehension and quicker insights compared to raw data tables or textual descriptions.

Improved Communication

1. **Effective Storytelling**: Visualizations help in telling compelling stories by highlighting key messages and insights, making the communication more engaging and persuasive.
2. **Universal Language**: Visual elements are often easier to understand than written text, making them accessible to a broader audience, regardless of language or technical proficiency.

Better Decision-Making

1. **Data-Driven Decisions**: Visualizations provide a clear view of data trends, anomalies, and patterns, facilitating more informed and data-driven decision-making.
2. **Identifying Trends and Outliers**: Visual tools make it easier to spot trends, correlations, and outliers that might be missed in traditional data analysis.

Increased Efficiency

1. **Time-Saving**: Visual tools can quickly summarize large datasets, saving time in analysis and reporting.
2. **Enhanced Productivity**: Interactive visualizations and dashboards enable users to explore data dynamically, leading to more efficient analysis and exploration.

Enhanced Memory Retention

1. **Visual Memory**: People tend to remember visual information better than text-based information. Effective visualizations can enhance memory retention of key insights and data points.

Greater Accessibility

1. **Broad Audience Reach**: Visualizations can make complex data accessible to non-experts, including stakeholders, clients, and the public, fostering better understanding and engagement

2. **Interactive Exploration**: Interactive visualizations and dashboards allow users to explore data on their own, providing a personalized and deeper understanding of the information.

Facilitates Pattern Recognition

1. **Identifying Relationships**: Visual tools help in recognizing relationships and correlations between different data variables, aiding in deeper analysis and discovery.
2. **Detecting Changes**: Visualizations can highlight changes and trends over time, making it easier to monitor and respond to evolving data.

Support for Storytelling and Persuasion

1. **Narrative Support**: Visualizations can complement and enhance storytelling by providing clear and compelling evidence to support narratives and arguments.
2. **Persuasive Communication**: Well-designed visualizations can be persuasive tools, helping to convince stakeholders and decision-makers of key insights and recommendations.

By leveraging these benefits, organizations and individuals can gain a deeper understanding of their data, communicate more effectively, and make better-informed decisions.

Distributions

Displays value distributions for a single attribute.

Inputs

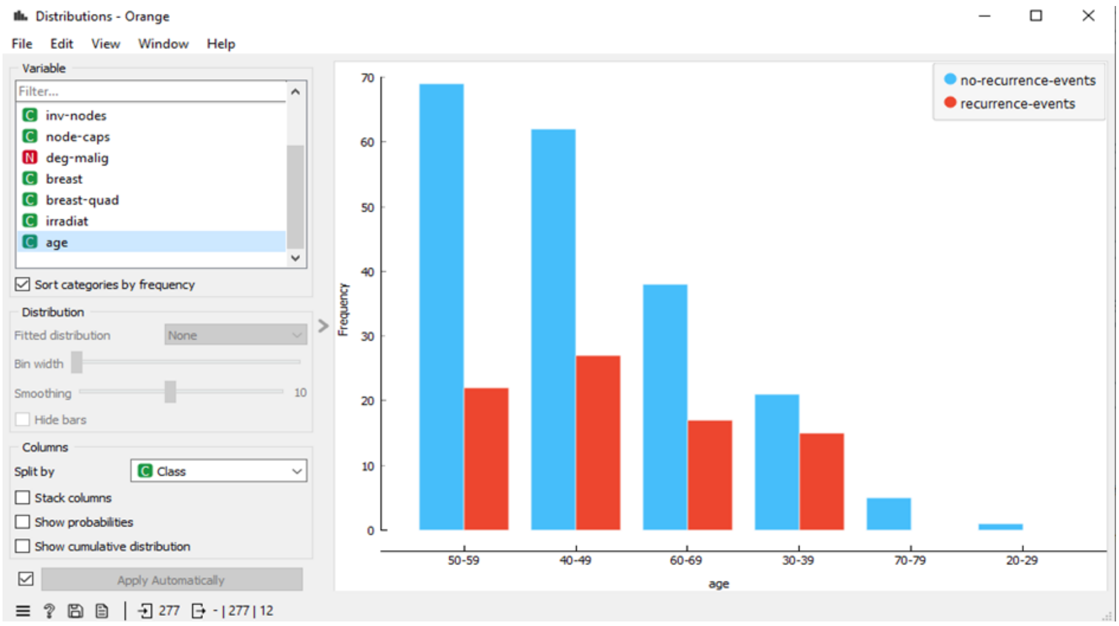
- Data: input dataset

Outputs

- Selected Data: instances selected from the plot
- Data: data with an additional column showing whether an instance is selected
- Histogram Data: bins and instance counts from the histogram

The **Distributions** widget displays the value distribution of discrete or continuous attributes. If the data contains a class variable, distributions may be conditioned on the class.

The graph shows how many times (e.g., in how many instances) each attribute value appears in the data. If the data contains a class variable, class distributions for each of the attribute values will be displayed (like in the snapshot below).



Distributions - Orange

Variable

- inv-nodes
- node-caps
- deg-malig
- breast
- breast-quad
- irradiat
- age

☑ Sort categories by frequency

Distribution

Fitted distribution: None

Bin width

Smoothing: 10

□ Hide bars

Columns

Split by: Class

□ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

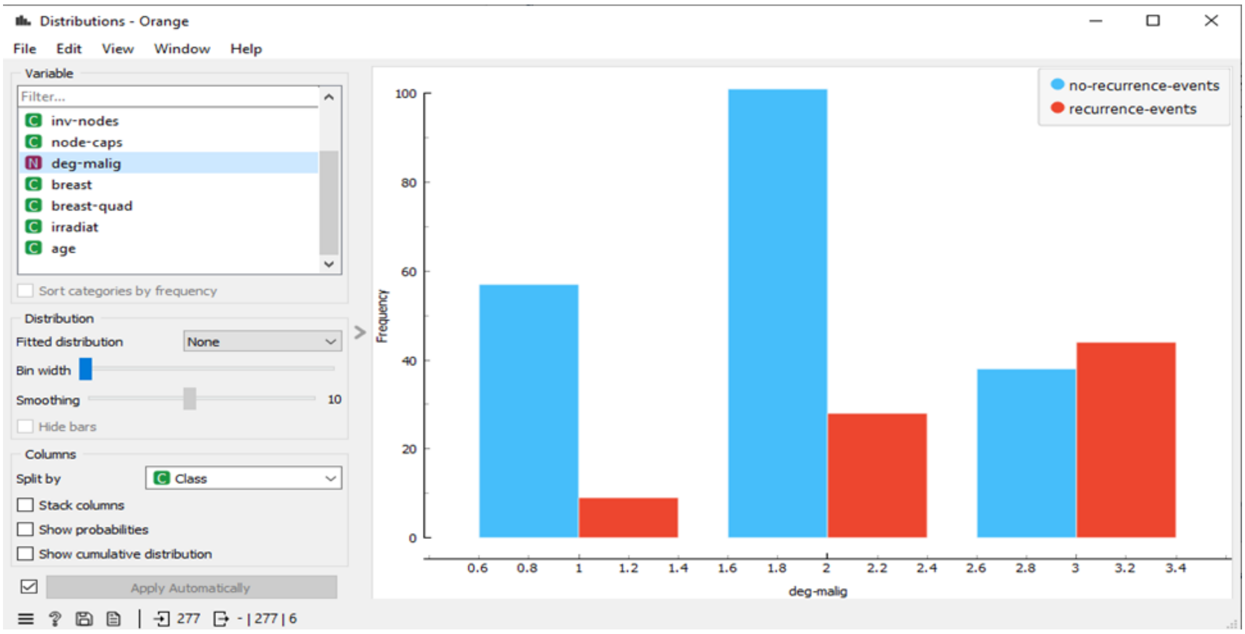
1. A list of variables for display. Sort categories by frequency orders displayed values by frequency.
2. Set Bin width with the slider. Precision scale is set to sensible intervals. Fitted distribution fits selected distribution to the plot. Options
   are Normal, Beta, Gamma, Rayleigh, Pareto, Exponential, Kernel density.
3. Columns:

- Split by displays value distributions for instances of a certain class.
- Stack columns display one column per bin, colored by proportions of class values.
- Show probabilities shows probabilities of class values at selected variable.
- Show cumulative distribution cumulatively stacks frequencies.

1. If Apply Automatically is ticked, changes are

communicated automatically. Alternatively, click Apply.

For continuous attributes, the attribute values are also displayed as a histogram. It is possible to fit various distributions to the data, for example, a Gaussian kernel density estimation. Hide bars hides histogram bars and shows only distribution (old behavior of Distributions).



Variable
- C inv-nodes
- C node-caps
- N deg-malig
- C breast
- C breast-quad
- C irradiat
- C age

□ Sort categories by frequency

Distribution

Fitted distribution: None

Bin width

Smoothing: 10

□ Hide bars

Columns

Split by: C Class

□ Stack columns

□ Show probabilities

□ Show cumulative distribution

☑ Apply Automatically

277 | - | 277 | 6

Mosaic Display

Display data in a mosaic plot.

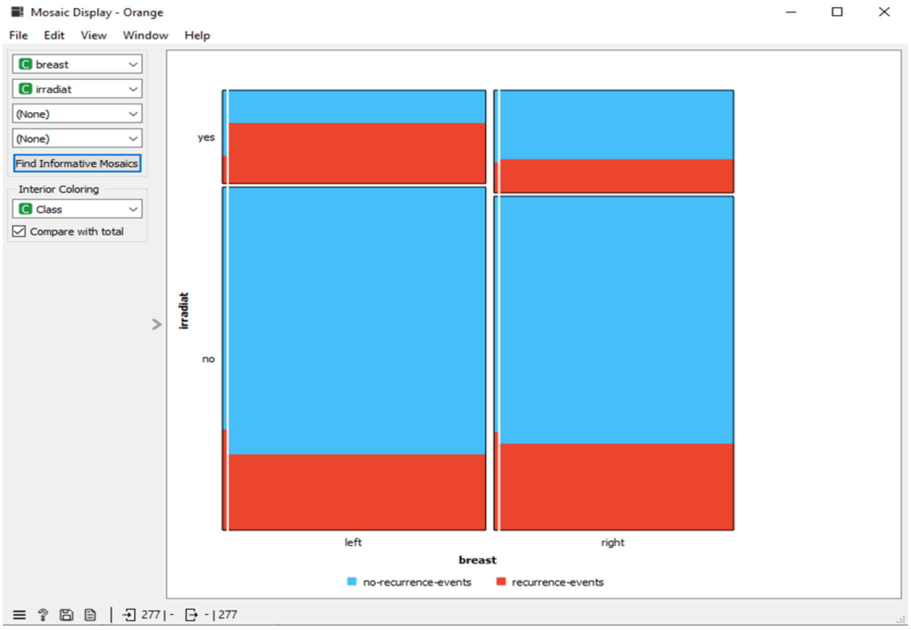
Inputs

- Data: input dataset
- Data subset: subset of instances

Outputs

- Selected data: instances selected from the plot

The **Mosaic plot** is a graphical representation of a two-way frequency table or a contingency table. It is used for visualizing data from two or more qualitative variables and was introduced in 1981 by Hartigan and Kleiner and expanded and refined by Friendly in 1994. It provides the user with the means to more efficiently recognize relationships between different variables



breast
irradiat
(None)
(None)
Interior Coloring
Class
☑ Compare with total

Diagnostic Analytics

Diagnostic Analytics is a type of advanced analytics that examines data to understand the causes behind certain events or outcomes. It goes beyond descriptive analytics, which focuses on what happened, by providing insights into why it happened. This deeper level of analysis is essential for identifying root causes and understanding complex patterns within the data.

Box Plot

Shows distribution of attribute values.

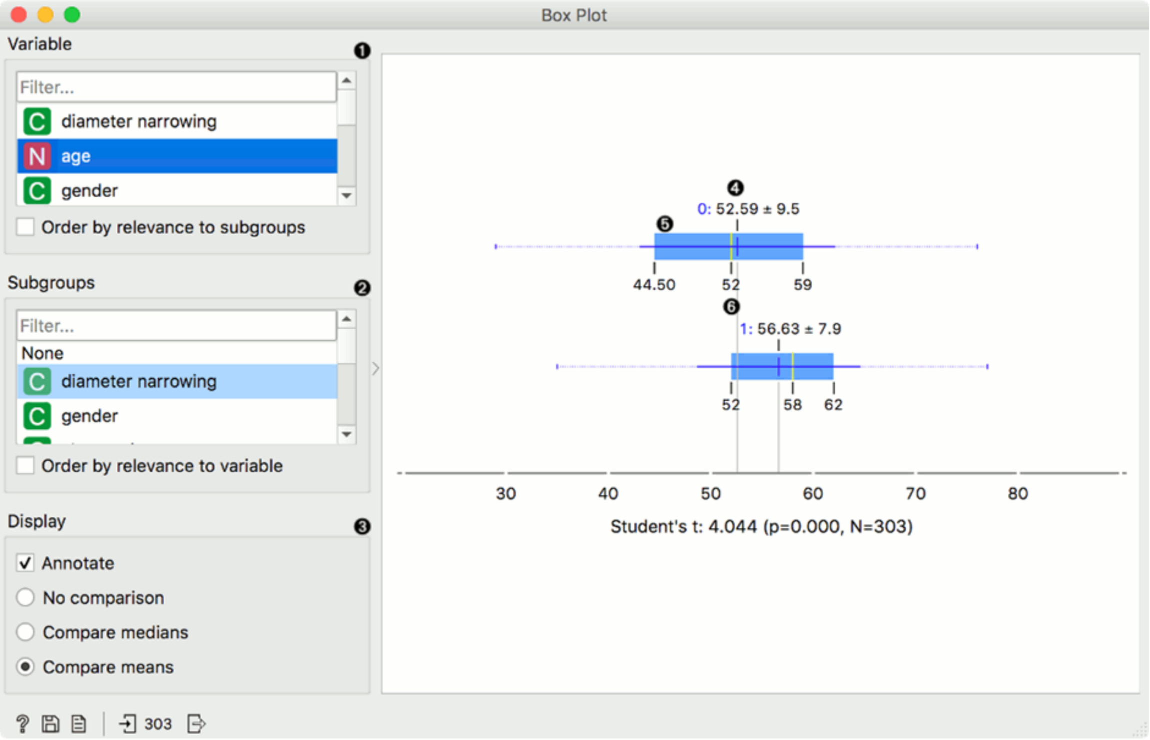
Inputs

- Data: input dataset

Outputs

- Selected Data: instances selected from the plot
- Data: data with an additional column showing whether a point is selected

The **Box Plot** widget shows the distributions of attribute values. It is a good practice to check any new data with this widget to quickly discover any anomalies, such as duplicated values (e.g., gray and grey), outliers, and alike. Bars can be selected - for example, values for categorical data or the quantile range for numeric data.


Box Plot

Variable ❶

Filter...
C diameter narrowing
N age
C gender

☐ Order by relevance to subgroups

Subgroups ❷

Filter...
None
C diameter narrowing
C gender

☐ Order by relevance to variable

Display ❸

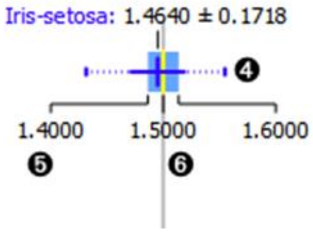
☑ Annotate
○ No comparison
○ Compare medians
● Compare means

303

Student's t: 4.044 (p=0.000, N=303)

1. Select the variable you want to plot. Tick *Order by relevance to subgroups* to order variables by Chi2 or ANOVA over the selected subgroup.
2. Choose *Subgroups* to see box plots displayed by a discrete subgroup. Tick *Order by relevance to variable* to order subgroups by Chi2 or ANOVA over the selected variable.

When instances are grouped by a subgroup, you can change the display mode. Annotated boxes will display the end values, the mean and the median, while comparing medians and compare means will, naturally, compare the selected value between subgroups



4.The mean (the dark blue vertical line). The thin blue line represents the standard deviation.

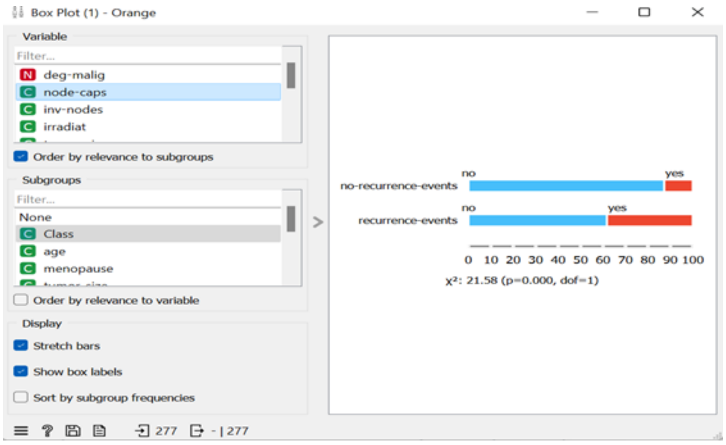
5.Values of the first (25%) and the third (75%) quantile. The blue highlighted area represents the values between the first and the third quartile.

6.The median (yellow vertical line).

Box plots are commonly used in various situations to visually represent the distribution and key statistical measures of a dataset

We use the chi-square test to determine the relationship between two categorical variables.

$P=0.000 < 0.05$, then there is a relationship between node-caps and class.



Box Plot (1) - Orange

Variable

Filter...

- N deg-malig
- C node-caps
- C inv-nodes
- C irradiat

☑ Order by relevance to subgroups

Subgroups

Filter...

- None
- C Class
- C age
- C menopause

☐ Order by relevance to variable

Display

☑ Stretch bars

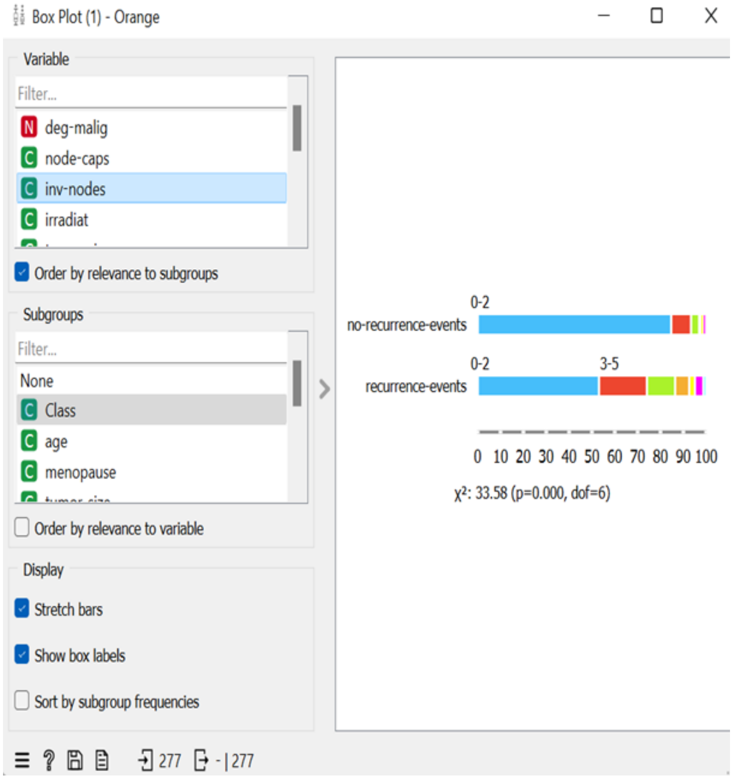
☑ Show box labels

☐ Sort by subgroup frequencies

277 | - | 277

χ²: 21.58 (p=0.000, dof=1)

- We use the chi-square test to determine the relationship between two categorical variables.
- P=0.000 < 0.05, then there is a relationship between **inv-nodes and class.**



Box Plot (1) - Orange

Variable

Filter...

- N deg-malig
- C node-caps
- C inv-nodes
- C irradiat

☑ Order by relevance to subgroups

Subgroups

Filter...

- None
- C Class
- C age
- C menopause

☐ Order by relevance to variable

Display

☑ Stretch bars

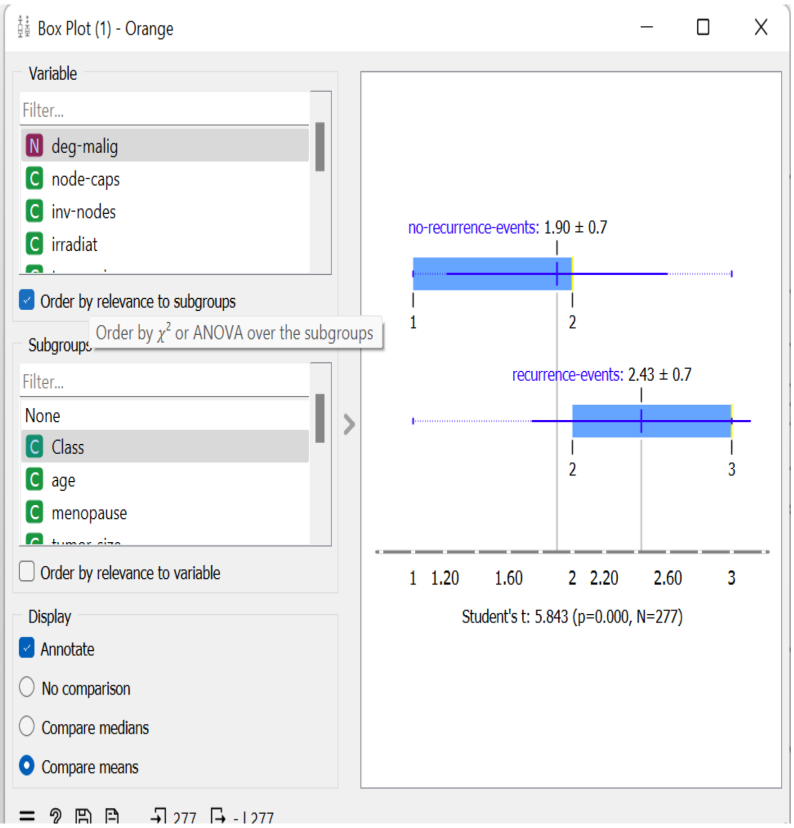
☑ Show box labels

☐ Sort by subgroup frequencies

277 | - | 277

χ²: 33.58 (p=0.000, dof=6)

- We use The Student's t-test to determine the relationship the means of two groups. (no-recurrence and recurrence events).
- P=0.000 < 0.05, then there is a relationship between **deg-malig and class**.



Box Plot (1) - Orange

Variable

Filter...

- N deg-malig
- C node-caps
- C inv-nodes
- C irradiat

☑ Order by relevance to subgroups

Order by χ^2 or ANOVA over the subgroups

Subgroups

Filter...

- None
- C Class
- C age
- C menopause

☐ Order by relevance to variable

Display

☑ Annotate

○ No comparison

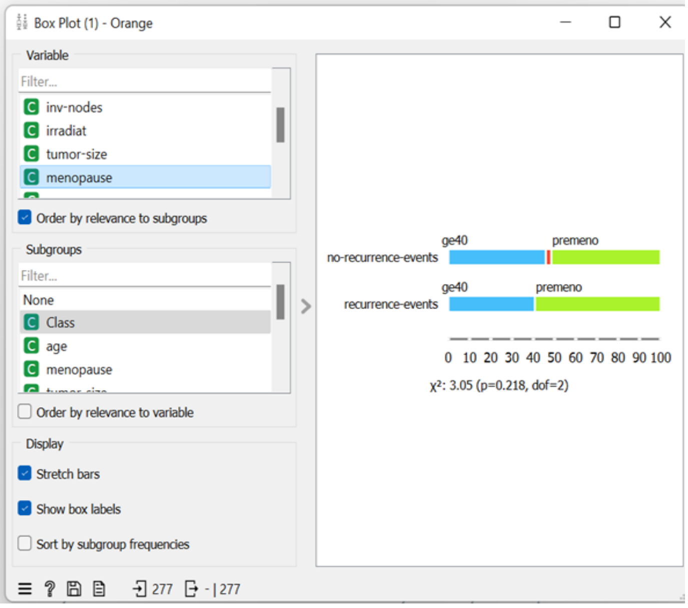
○ Compare medians

● Compare means

Student's t: 5.843 (p=0.000, N=277)

no-recurrence-events: 1.90 ± 0.7
1
2
recurrence-events: 2.43 ± 0.7
2
3
1 1.20 1.60 2 2.20 2.60 3
277
- | 277

- We use the chi-square test to determine the relationship between two categorical variables.
- P=0.218 > 0.05, then there is no a relationship between **menopause and class**.



Box Plot (1) - Orange

Variable

Filter...

- C inv-nodes
- C irradiat
- C tumor-size
- C menopause

☑ Order by relevance to subgroups

Subgroups

Filter...

- None
- C Class
- C age
- C menopause

☐ Order by relevance to variable

Display

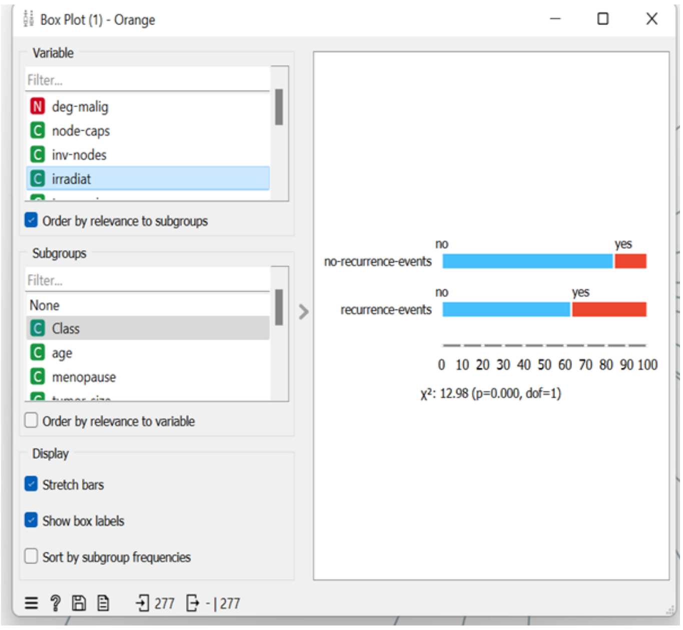
☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 3.05 (p=0.218, dof=2)

- We use the chi-square test to determine the relationship between two categorical variables.
- P=0.000 < 0.05, then there is a relationship between **irradiat and class**..



Box Plot (1) - Orange

Variable

Filter...

N deg-malig
C node-caps
C inv-nodes
C irradiat

☑ Order by relevance to subgroups

Subgroups

Filter...

None
C Class
C age
C menopause

☐ Order by relevance to variable

Display

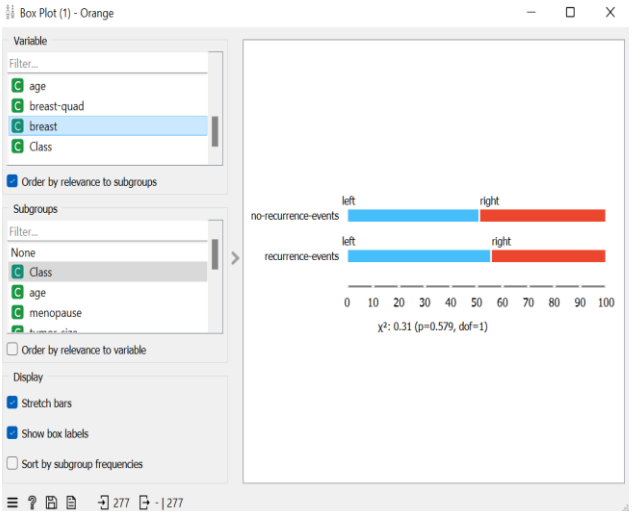
☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 12.98 (p=0.000, dof=1)

- We use the chi-square test to determine the relationship between two categorical variables.
- P=0.579 > 0.05, then there is no a relationship between **breast and class.**



Box Plot (1) - Orange

Variable

Filter...

C age
C breast-quad
C breast
C Class

☑ Order by relevance to subgroups

Subgroups

Filter...

None
C Class
C age
C menopause

☐ Order by relevance to variable

Display

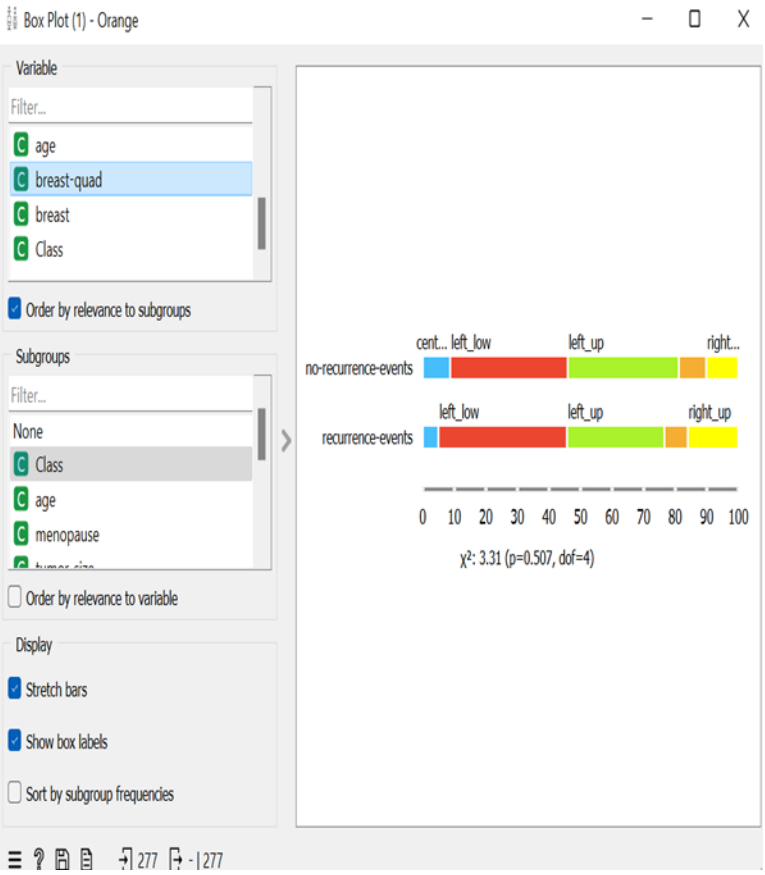
☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 0.31 (p=0.579, dof=1)

- We use the chi-square test to determine the relationship between two categorical variables.
- P=0.507 > 0.05, then there is no a relation between **breast-quad and class.**


Box Plot (1) - Orange

Variable

- age
- breast-quad
- breast
- Class

☑ Order by relevance to subgroups

Subgroups

- None
- Class
- age
- menopause
- tumor-size

☐ Order by relevance to variable

Display

☑ Stretch bars

☑ Show box labels

☐ Sort by subgroup frequencies

χ²: 3.31 (p=0.507, dof=4)

Then the order by relevance to class in terms of importance.

- deg-malig
- node-caps
- Inv-nodes
- Irradiat
- tumor-size
- menopause
- age

- breast-quad
- breast

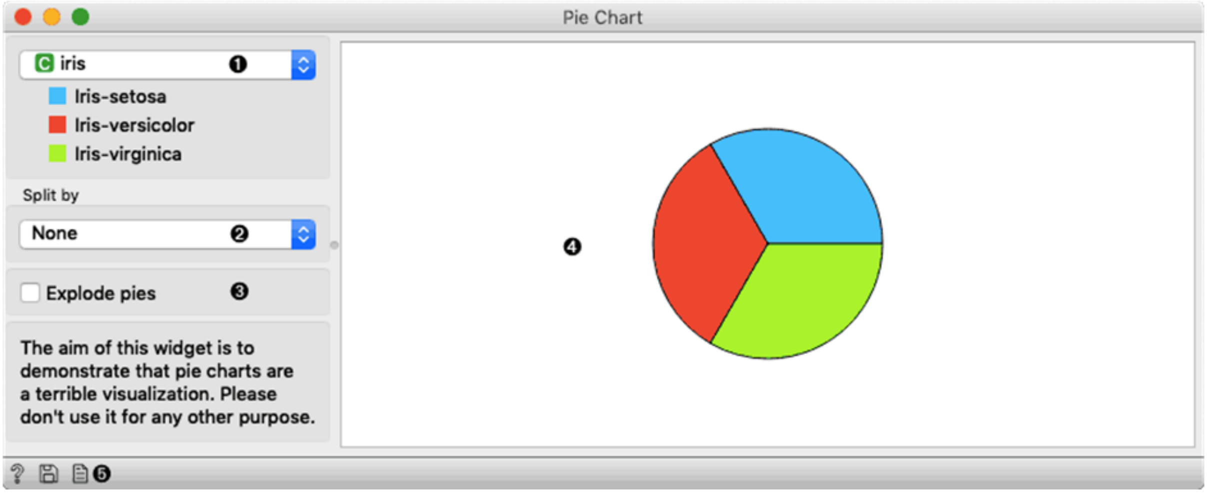
Pie Chart

The widget for visualizing discrete attributes in the pie chart.

Inputs

- Data: input data set

The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.



Pie Chart

iris
🟦 Iris-setosa
🟥 Iris-versicolor
🟩 Iris-virginica

Split by: None

□ Explode pies

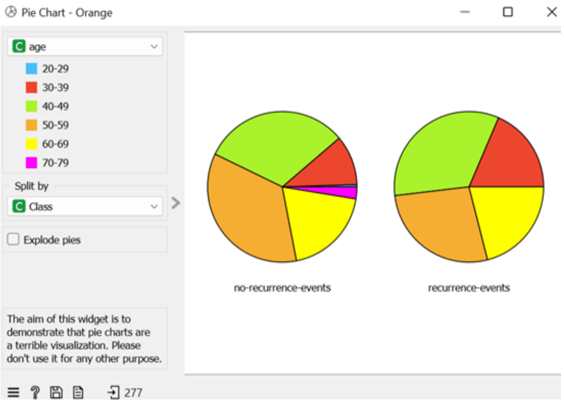
The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.

1. Select the attribute you want to visualize.
2. Select the attribute which is used to split data in more charts.
3. Check if you want pies to explode (parts of the pie will have space in between).
4. You will see your data visualized here.
5. With those buttons, you can either get help, save the plot, or include plots in the report.

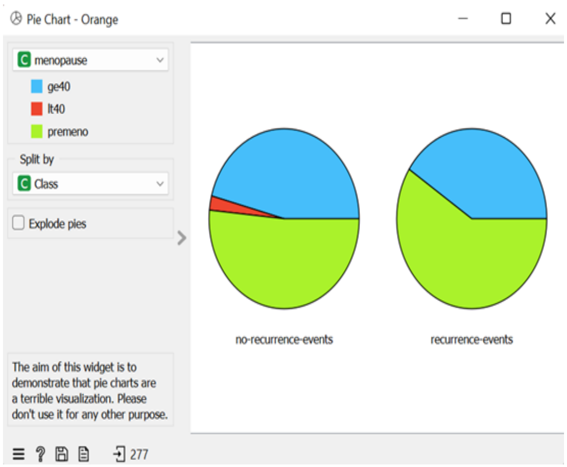
-A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions.

- Each slice represents a proportionate part of the whole, and the total of all slices forms the complete circle.

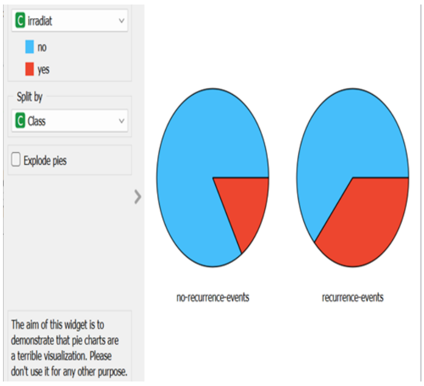
Pie charts are commonly used to show the distribution of categorical data or the relative sizes of different categories within a dataset.



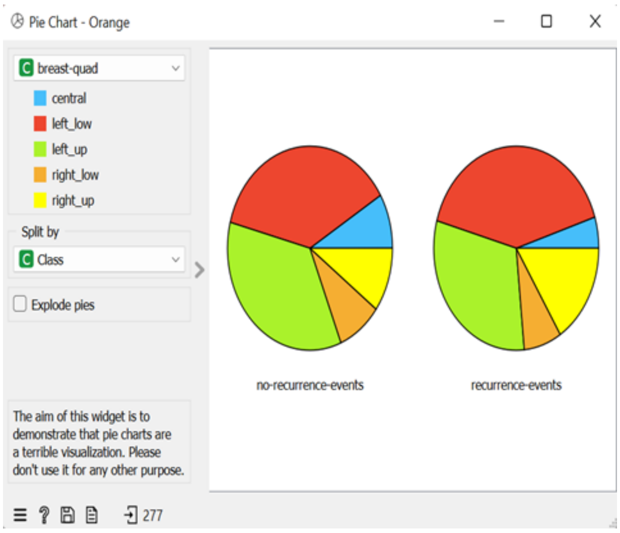
The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.



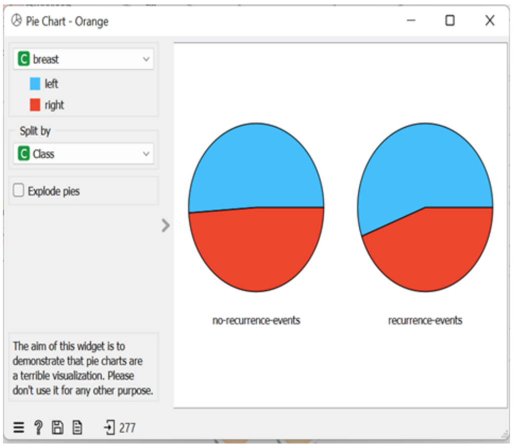
The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.



The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.



The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.



Pie Chart - Orange

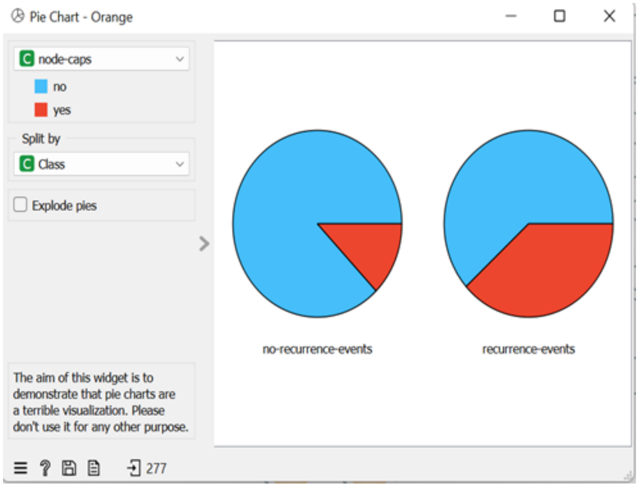
breast
🟦 left
🟥 right

Split by
Class

☐ Explode pies

The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.

277



Pie Chart - Orange

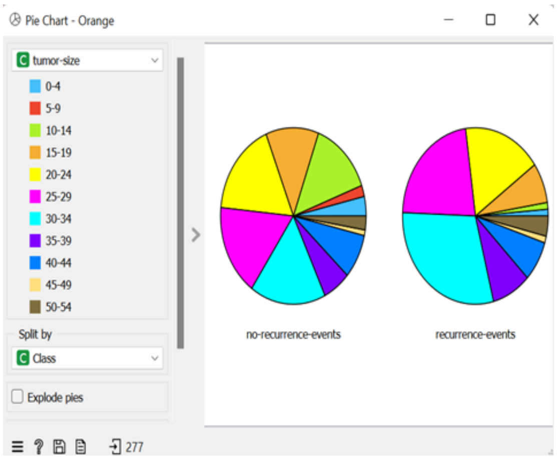
node-caps
🟦 no
🟥 yes

Split by
Class

☐ Explode pies

The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.

277



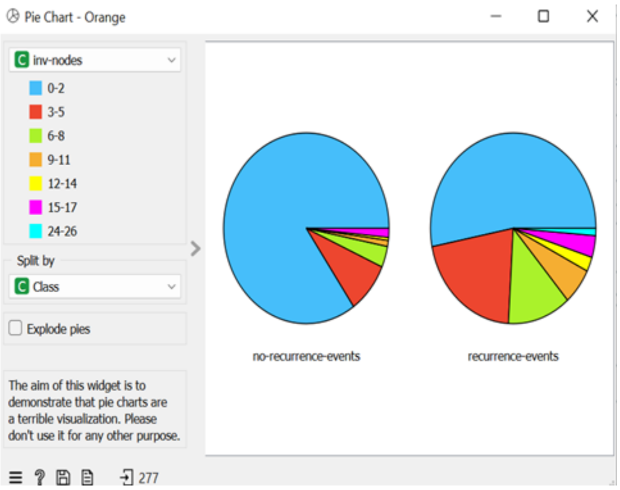
Pie Chart - Orange

tumor-size
🟦 0-4
🟥 5-9
🟩 10-14
🟧 15-19
🟨 20-24
🟪 25-29
🟦 30-34
🟪 35-39
🟦 40-44
🟧 45-49
🟫 50-54

Split by
Class

☐ Explode pies

277



Pie Chart - Orange

inv-nodes
🟦 0-2
🟥 3-5
🟩 6-8
🟧 9-11
🟨 12-14
🟪 15-17
🟦 24-26

Split by
Class

☐ Explode pies

The aim of this widget is to demonstrate that pie charts are a terrible visualization. Please don't use it for any other purpose.

277

Chapter 5

Predictive Analytics:

Predictive analytics involves using historical data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on past data. It aims to forecast future events, trends, and behaviors, helping organizations make informed decisions.

Benefits

- **Informed Decision-Making:** Provides data-driven insights to support strategic planning and operational decisions.
- **Cost Reduction:** Identifies inefficiencies and areas for cost and health savings.
- **Improved Efficiency:** Enhances productivity and performance by optimizing processes (Early Detection) .
- **Risk Management:** Anticipates and mitigates risks before they materialize.

Here are the key aspects of predictive analytics:

Key Components

1. **Data Collection and Preparation:**
    - **Data Gathering:** Collecting relevant data from various sources such as databases, sensors, or external data providers.
    - **Data Cleaning:** Removing or correcting errors, handling missing values, and ensuring data quality.
    - **Data Transformation:** Normalizing, scaling, and encoding data to make it suitable for analysis.
2. **Feature Engineering:**
    - **Feature Extraction:** Identifying and creating new features from raw data that better capture the underlying patterns.
    - **Feature Selection:** Choosing the most relevant features to improve model performance and reduce complexity like deg-malig, Inv-nodes and age.

3. **Model Building:**

    - **Choosing Algorithms:** Selecting appropriate statistical, machine learning, or deep learning algorithms based on the detection of disease as it reoccurred or not.

    - **Training Models:** Using historical data to train models to learn patterns and relationships within the data.

    - **Validation:** Evaluating model performance using techniques like cross-validation to ensure robustness and prevent overfitting.

4. **Model Deployment:**

    - **Implementation:** Integrating the best predictive model into business processes or systems to provide real-time predictions.

    - **Monitoring and Maintenance:** Continuously monitoring model performance and updating it with new data to maintain accuracy over time.

Model Building

Model building for breast cancer recurrence prediction involves several steps, from data preparation and feature engineering to model training and deployment. Key features such as degree of malignancy, tumor size, involved lymph nodes, and age play crucial roles in the model's ability to accurately predict recurrence. A well-built model can provide valuable insights for clinicians, aiding in early intervention and personalized treatment plans for patients at risk of recurrence.

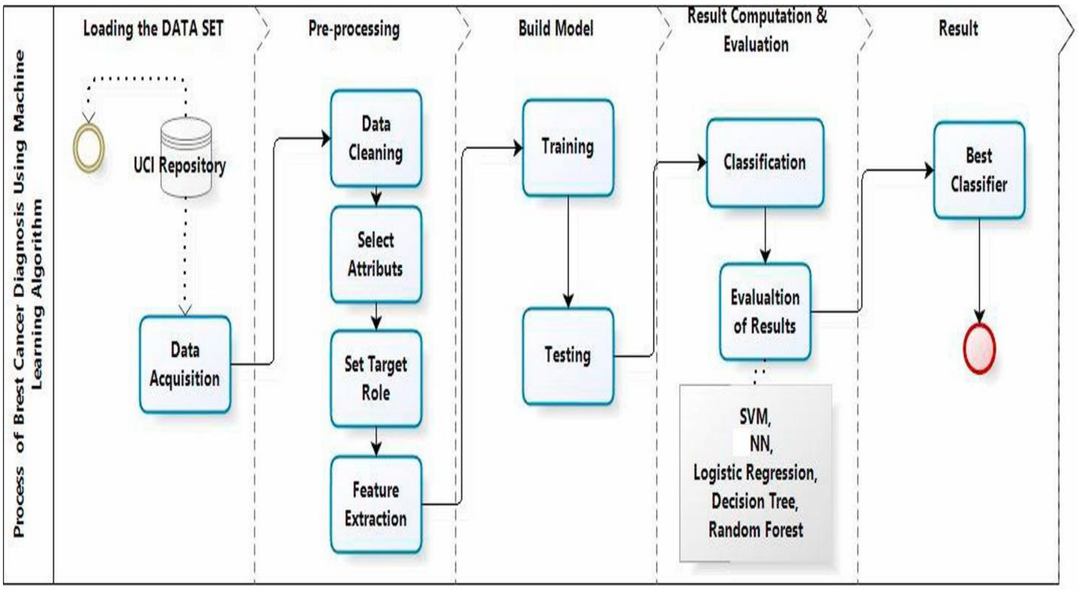


Fig (11)

- □ **choosing Algorithms:** Select appropriate algorithms for the prediction task. Common choices include logistic regression, decision trees, random forests, support vector machines, and neural networks.

- □ **Training Models:** Use historical data to train the predictive model. The model will learn the relationship between features (deg malig, tumor size, inv nodes, age, etc.) and the outcome (recurrence or not).

- □ **Validation:** Use techniques such as cross-validation to assess the model's performance and ensure it generalizes well to unseen data.

Model Training

Training a predictive model for breast cancer recurrence involves careful preparation of data, selection of relevant features, choice of appropriate algorithms, and rigorous evaluation and tuning. By focusing on key features like deg malig, tumor size, inv nodes, and age, the model can provide accurate predictions, aiding clinicians in identifying patients at higher risk of recurrence and tailoring treatment plans accordingly.

Splitting the Dataset

- **Training and Testing Split:** Divide the dataset into a training set and a testing set, typically using an 80/20 split to evaluate the model’s performance on unseen data.
- **Cross-Validation:** Optionally, use k-fold cross-validation to further assess the model’s robustness and ensure it generalizes well.

Choosing and Training the Model

- **Algorithm Selection:** Choose suitable machine learning algorithms for the prediction task. Common choices include:
  - **Logistic Regression:** For binary classification tasks.
  - **Decision Trees:** For interpretable models.
  - **Random Forests:** For robust, ensemble learning.
  - **Support Vector Machines (SVM):** For high-dimensional spaces.
  - **Neural Networks:** For complex patterns and interactions.

- **Model Training:** Train the selected algorithm on the training set. The model learns the relationships between the input features (deg malig, tumor size, inv nodes, age,etc) and the target variable (recurrence).

1st Model: Logistic Regression

Logistic regression is used for binary classification where we use sigmoid function, that takes input as independent variables and produces a probability value between 0 and 1.

For example, we have two classes Class 0 and Class 1 if the value of the logistic function for an input is greater than 0.5 (threshold value) then it belongs to Class 1 otherwise it belongs to Class 0. It's referred to as regression because it is the extension of linear regression but is mainly used for classification problems. in this project, we utilize Logistic Regression to predict whether breast cancer will recurrence or not recurrence based on a set of patient data

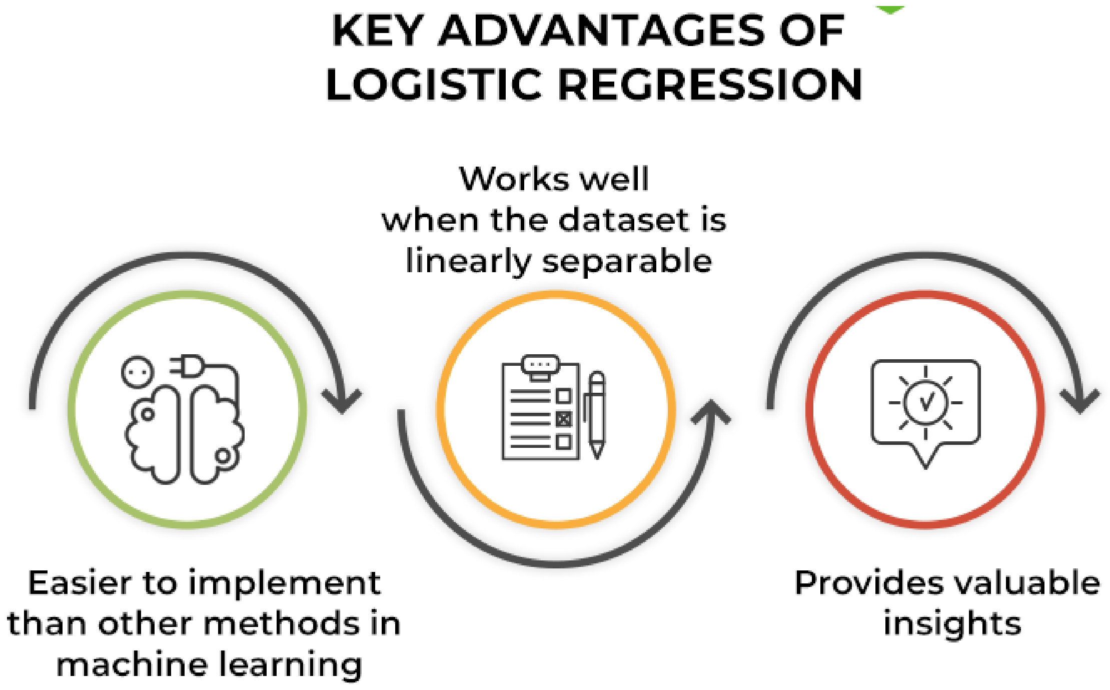
Logistic Regression is particularly suitable for this application due to its interpretability, allowing us to understand which patient characteristics are most influential in predicting recurrence.

Key Points:

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be of a categorical or discrete value.

It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).


KEY ADVANTAGES OF LOGISTIC REGRESSION
Works well when the dataset is linearly separable
Easier to implement than other methods in machine learning
Provides valuable insights

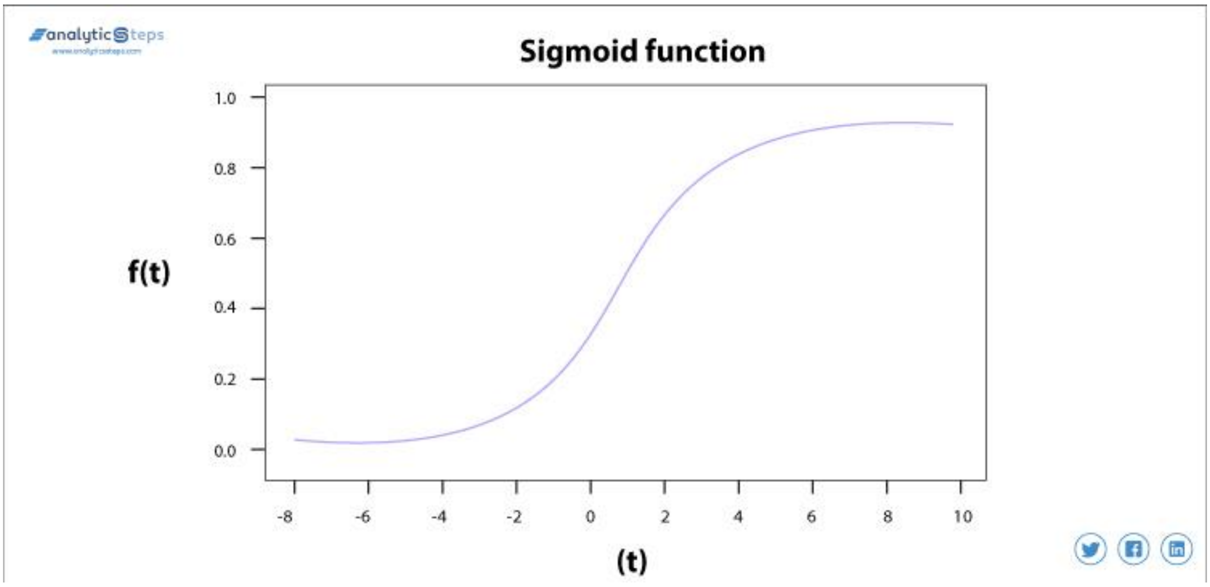
Logistic Function – Sigmoid Function

The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1. The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.



Types of Logistic Regression:

Based on the categories, Logistic Regression can be classified into three types:

-**Binomial**: In binomial Logistic regression, there can be only two possible types of dependent variables, such as 0 or 1, Pass or Fail, etc.

-**Multinomial**: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as in age “40-45”, “50-60”, or “60 to old”

-**Ordinal**: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as “low”, “Medium”, or “High”.

Terminologies involved in Logistic Regression:

-**Independent variables**: The input characteristics or predictor factors applied to the dependent variable's predictions.

-**Dependent variable**: The target variable in a logistic regression model, which we are trying to predict.

-**Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.

-**Odds:** It is the ratio of something occurring to something not occurring. It is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.

-Log-odds: The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.

-**Coefficient**: The logistic regression model's estimated parameters show how the independent and dependent variables relate to one another.

-**Intercept**: A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.

-**Maximum likelihood estimation**: The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

Advantages and disadvantages of logistic regression

advantages:

- **Easy to set up.** The main advantage of logistic regression is that it is a simple model which is much easier to set up and train
- **Efficient algorithms.** It is one of the most efficient algorithms when the different outcomes or distinctions represented by the data are linearly separable. This means that you can draw a straight line separating the results of a logistic regression calculation
- **Reveals interrelationships between variables.** it can help reveal the interrelationships between different variables and their effect on outcomes. This could quickly determine when two variables are positively or negatively correlated, such as the finding cited above that more studying tends to be correlated with higher test outcomes.
- **Transforms complex calculations into simple math problems.** It transforms complex calculations around probability into a straightforward arithmetic problem. The calculation itself is complex, but modern statistical methods and applications automate much of the calculations. This dramatically simplifies analyzing the effect of multiple variables and minimizes the effect of confounding factors. As a result, statisticians can quickly model and explore the contribution of various factors to a given outcome. For example, a medical researcher might want to know the effect of a new drug on treatment outcomes across different age groups. This involves a lot of nested multiplication and division for comparing the outcomes of young and older people who never received a treatment, younger people who received the treatment, older

people who received the treatment and then the whole spontaneous healing rate of the entire group. Logistic regression converts the relative probability of any subgroup into a logarithmic number, called a regression coefficient, that can be added or subtracted to arrive at the desired result. These more straightforward regression coefficients can also simplify other data science and machine learning algorithms.

- **Baseline for performance management.** Logistic regression is often used as a baseline to measure performance due to its quick and easy setup.

Logistic regression also comes with various

disadvantages:

- Assumption of linearity**.** Since logistic regression assumes a linear relationship between one dependent variable and the independent variables, its applicability in certain scenarios may be limited.
- **Overfitting and sensitivity to outliers.** Logistic regression is sensitive to outliers. If the number of observations is lesser than the number of features, logistic regression should not be used; otherwise it might lead to overfitting. L1 and L2 regularization techniques can be applied to help reduce overfitting.
- **Limited to binary outcomes.** Logistic regression is limited to modeling binary classification and outcomes and might not be suitable for scenarios with non-binary outcomes without modifications such as ordinal logistic regression.
- **Can only predict discrete functions.** Logistic regression is exclusively designed for predicting discrete functions, which constrains its use to dependent variables within a discrete

number set. This limitation poses challenges for predicting continuous data. **Accuracy: 0.788**

2nd Model: Decision Tree

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It is a non-parametric supervised learning method that creates a model predicting the value of a target variable by learning simple decision rules inferred from the data features.

Structure of Decision Trees

A decision tree is structured as a flowchart-like tree, where:

- **Root Node**: Represents the entire dataset, which is split into two or more homogeneous sets.

- **Internal Nodes**: Each node represents a decision point based on a feature.

- **Branches**: Represent the outcome of a decision, connecting one node to another.

- **Leaf Nodes**: Represent the final output or decision (class label or continuous value).

Building a Decision Tree

1.**Feature Selection**: The process starts by selecting the feature that best splits the data. Common criteria for splitting include:

- **Gini Index**: Measures the impurity of a node, with lower values indicating a more homogeneous node.

- **Entropy**: Used in information gain, measures the randomness in the information being processed.

- **Variance Reduction**: Used in regression trees to minimize the variance within subsets of the data.

2.**Splitting**: Based on the chosen feature, the dataset is split into subsets. The goal is to maximize the homogeneity of the resulting subsets (minimizing impurity or variance).

3. **Stopping Criteria**: The tree grows until it meets a stopping condition, such as:

- Maximum tree depth.

- Minimum number of samples per leaf.

- Minimum impurity decrease.

4. **Pruning**: To avoid overfitting, trees may be pruned after they are fully grown. Pruning removes branches that have little importance, enhancing the model's generalizability.

Types of Decision Trees

Decision trees come in various forms tailored to different types of prediction tasks. Here, we explore the main types :

Classification Trees

Classification trees are used when the target variable is categorical. The goal is to assign input data into predefined categories or classes.

Key Features:

- **Splitting Criteria**: Common metrics include Gini Index and Entropy (Information Gain). These metrics aim to maximize the separation between the classes at each split.

- **Output**: A class label (e.g., "yes" or "no", "recurrence" or "non-recurrence ").

- **Leaf Nodes**: Represent class labels that the model predicts.

It is used in Medical Diagnosis: Classifying patients as having a disease or not based on symptoms (Features).

Regression Trees

Regression trees are used when the target variable is continuous. The goal is to predict a numeric value based on input features.

Key Features:

- **Splitting Criteria**: Common metrics include Mean Squared Error (MSE) and Mean Absolute Error (MAE). These metrics aim to minimize the prediction error at each split.

- **Leaf Nodes**: Represent the predicted value or the average value of the target variable in that node.

Advantages and Disadvantages of Decision Trees

Advantages:

1. **Easy to Understand and Interpret**:

- Decision trees are intuitive and straightforward to visualize, making them easy to understand even for non-experts.

2. **No Need for Data Normalization:**

- Unlike some algorithms that require data normalization (e.g., SVMs), decision trees do not need the data to be scaled or standardized.

3. **Handle Both Numerical and Categorical Data:**

- Decision trees can manage both types of data, providing flexibility in handling various datasets.

4. **Require Little Data Preparation:**

- They do not require extensive preprocessing of data, such as normalization or handling missing values separately.

5. **Non-parametric Nature:**

- Decision trees do not assume any underlying distribution of the data, making them versatile in modeling different types of data distributions.

6. **Feature Selection and Importance**:

- They automatically perform feature selection, ranking features by their importance which can be insightful for understanding the data.

7. **Versatility:**

- They can be used for both classification and regression tasks.

Disadvantages:

1. **Prone to Overfitting:**

- Decision trees can easily become overly complex, capturing noise in the training data, which leads to poor generalization to new data.

- This is particularly a problem with deep trees that have many levels.

2. **Unstable**

- Small changes in the data can result in very different trees being generated, leading to high variance in predictions.

- This instability can be mitigated by ensemble methods like Random Forests.

3. **Bias Towards Dominant Classes:**

- In datasets with imbalanced classes, decision trees can become biased towards the majority class, leading to suboptimal performance for minority classes.

4. **Greedy Nature of Splits:**

- The algorithm uses a greedy approach to select splits, which means it chooses the best split at each step without considering the global optimal split. This can s lead to suboptimal solutions.

5. **Complexity with Large Datasets:**

- For very large datasets, decision trees can become computationally expensive and require significant memory and processing time.

6**. Interpretability Declines with Complexity:**

- As trees grow deeper and more complex, their interpretability diminishes.

- Pruning techniques and limiting tree depth can help but at the cost of potentially losing valuable information.

7. **Limited Expressiveness**:

- Decision trees can struggle to model some relationships due to their piecewise constant nature.

- They may fail to capture smooth and continuous changes in the data.

Accuracy 0.90

Decision trees offer a clear and interpretable way to make decisions based on data, supporting both classification and regression tasks. However, they come with challenges such as overfitting, instability, and bias, especially in the presence of imbalanced data. Balancing their use involves careful consideration of tree complexity, pruning techniques, and potentially using ensemble methods to enhance performance and robustness.

3rd Model: Random Forest Model

Random forest, a popular machine learning algorithm that merges the outputs of numerous decision trees to produce a single outcome. Its popularity stems from its user-friendliness and versatility, making it suitable for both classification and regression tasks.

How it works?

Random forest algorithms have three main hyperparameters, which need to be set before training. These include node size, the number of trees, and the number of features sampled. From there, the random forest classifier can be used to solve regression or classification problems.

The random forest algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement, called the bootstrap sample. Of that training sample, one-third of it is set aside as test data, known as the out-of-bag (oob) sample, which we'll come back to later. Another instance of randomness is then injected through feature bagging, adding more diversity to the dataset and reducing the correlation among decision trees. Depending on the type of problem, the determination of the prediction will vary. For a regression task, the individual decision trees will be averaged, and for a classification task, a majority vote i.e. the most frequent categorical variable will yield the predicted class. Finally, the oob sample is then used for cross-validation, finalizing that prediction.

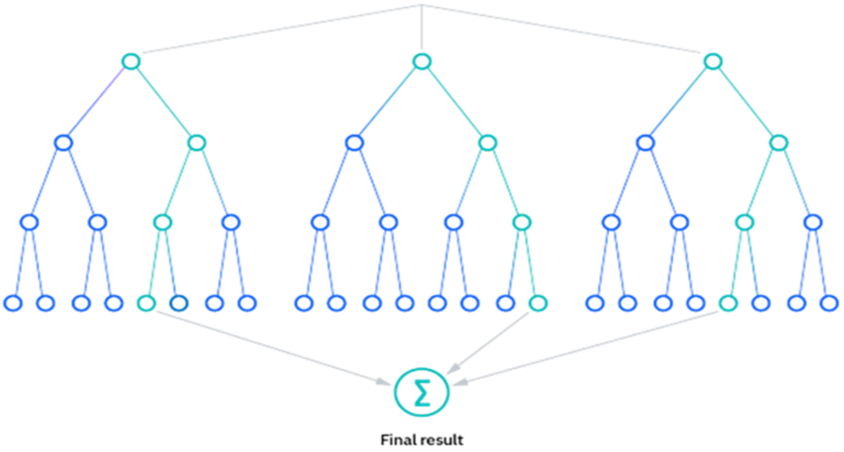


Fig (12)

Key Benefits

- **Reduced risk of overfitting**: Decision trees run the risk of overfitting as they tend to tightly fit all the samples within training data. However, when there's a robust number of decision trees in a random forest, the classifier won't overfit the model since the averaging of uncorrelated trees lowers the overall variance and prediction error.

- **Provides flexibility**: Since random forest can handle both regression and classification tasks with a high degree of accuracy, it is a popular method among data scientists. Feature bagging also makes the random forest classifier an effective tool for estimating missing values as it maintains accuracy when a portion of the data is missing.

- **Easy to determine feature importance**: Random Forest makes it easy to evaluate variable importance, or contribution, to the model. There are a few ways to evaluate feature importance. Gini importance and mean decrease in impurity (MDI) are usually used to measure how much the model's accuracy decreases when a given variable is excluded. However, permutation importance, also known as mean decrease accuracy (MDA), is another important measure. MDA identifies the average decrease in accuracy by randomly permutating the feature values in oob samples.

Key Challenges

- **Time-consuming process**: Since random forest algorithms can handle large data sets, they can provide more accurate predictions, but can be slow to process data as they are computing data for each individual decision tree.

- **Requires more resources**: Since random forests process larger data sets, they'll require more resources to store that data.

- **More complex**: The prediction of a single decision tree is easier to interpret when compared to a forest of them.

The random forest algorithm has been applied across several industries, allowing them to make better business decisions. We use it in Healthcare: The random forest algorithm has applications within computational biology (link resides outside ibm.com), allowing doctors to tackle problems such as cancer disease expression classification. As a result, doctors can make estimates around medicine responses to specific medications.

Random forest is a great choice if anyone wants to build the model fast and efficiently, as one of the best things about the random forest Classifier is it can handle missing values. It is one of the best techniques with high performance, widely used in various industries for its efficiency. It can handle binary, continuous, and categorical data. Overall, random forest is a fast, simple, flexible, and robust model with some limitations.

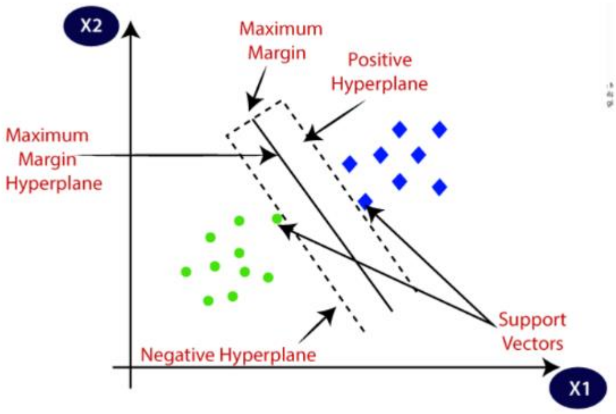
The Accuracy :0.928

4th Model: Support Vector Machines (SVM)

Support Vector Machine abbreviated as (SVM) is a powerful supervised algorithm that works best on smaller datasets and on complex ones. can be used for both regression, classification and outlier detection purposes tasks, but generally, they work best in classification problems.

Idea of SVM:

The idea of SVM by finding the hyperplane that best separates different classes in the feature space.

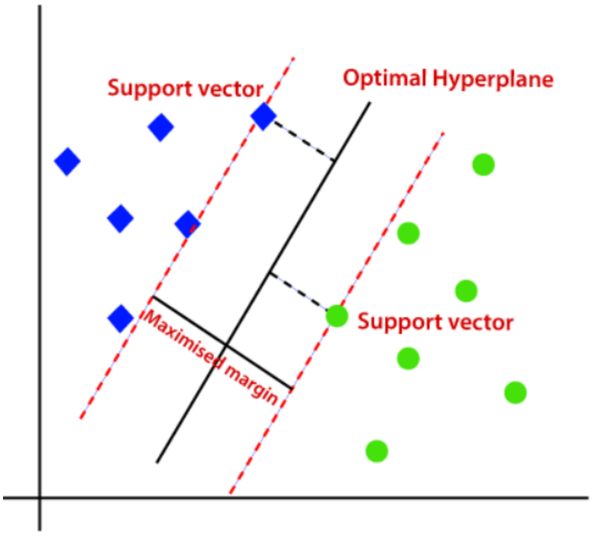


How an SVM works:

-SVM works is to have two classes separate them by Generate an optimal hyperplane that separates two classes of data points in the best possible way, .t is known as the **maximum margin hyperplane**.

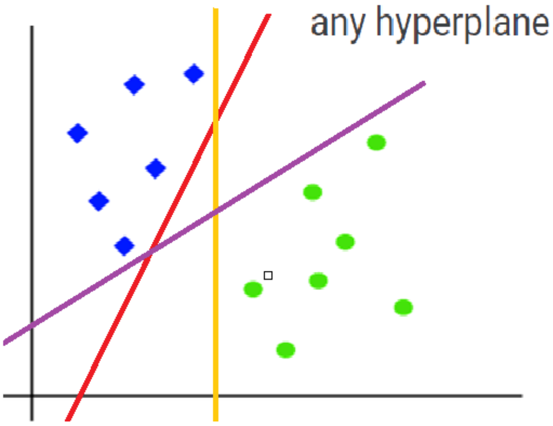
- There are many hyperplanes that might classify the data. We should look for the best hyperplane that represents the largest separation, or margin, between the two classes.

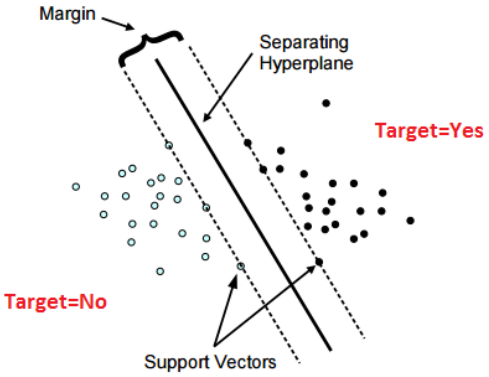
- The reason we aim to maximize the margin in SVM is to increase the "**robustness**" and "**generalizability**" of the model,



where "**robustness**": A larger margin provides a buffer zone where new data points are less likely to influence the hyperplane position. This makes the model more robust to noise and small changes in the data.

"**Generalizability**": A hyperplane that is farther away from the data points (i.e. has a large margin) is less likely to exceed the training data, this solves problems **Overfitting** occurs when a model fits the training data poorly and performs poorly on unseen data. By maximizing the margin, we ensure that the model generalizes well to new, unseen data, **we can solve this problem by using parameter (C).**





Support Vector Machine

-The **margin** is the distance between the hyperplane and the closest data points from each category, these closest data points are known as support vectors.

-**Support vectors** are the sample data points, which are closest to the hyperplane, called "support" because they support me in being able to distinguish between two classes from each other, meaning they help me in making decisions in the classification process. Through it, I can identify the best Hyperplane that separates two classes from each other.

Implement the Model:

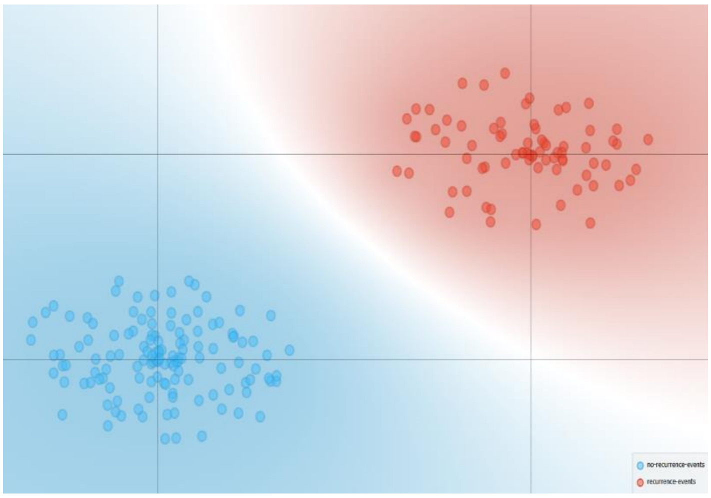
- Cost (C): 5.00

The cost parameter C controls the trade-off between maximizing the margin and minimizing the classification error. while a low C value aims for a larger margin but allows some classification errors (miss-classification), which can help in better generalization.

-Regression Loss Epsilon (ε): 0.010 (default value)

The epsilon (ε) parameter in (SVM) provides several benefits, primarily related to how the model handles errors in data and achieves generalization by: Making the model robust to minor fluctuations and noise, helping in better generalization and avoiding overfitting.

Kernel type (hyperplane) to separates different classes**:**



Radial Basis Function (RBF) Kernel (Y or gamma) =0.14

Formula: $K (x, y) =exp(-\gamma\|x-y\|^2)$, where: $K (x, y)$ is the kernel function measures the similarity between two data points x and y.

x and y are two input vectors, Gamma (γ or g): The parameter γ controls the width of the Gaussian kernel, $\|x-y\|2$ is the squared Euclidean distance between x and y, exp denotes the exponential function.

The exponential function is applied to the negative squared distance multiplied by γ.

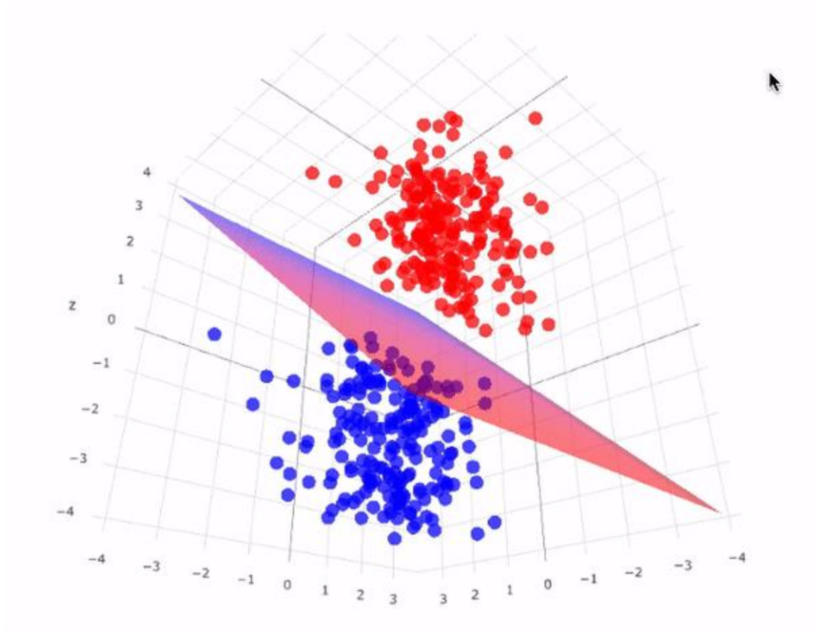
The kernel function essentially measures the similarity between x and y. The similarity is higher when x and y are closer, and it decreases exponentially as the distance increases.

γ controls the width of the Gaussian function, affecting how quickly the similarity decreases with distance The exponential function ensures that the kernel values are between 0 and 1, with higher values indicating greater similarity, give

Accuracy: 0.950

convert to 3D

represent as a linear



5th Model: Artificial Neural Network (ANN)

An artificial neural network (ANN) is a computational model inspired by the structure and functioning of the human brain. ANNs consist of interconnected groups of artificial neurons (also called nodes) that process information using a connectionist approach to computation. Here are the key components and concepts associated with artificial neural networks:

◎ **Neurons (Nodes):**

- Basic units that receive input, process it, and pass on the output to other neurons.
- Each neuron performs a weighted sum of its inputs, applies an activation function, and produces an output.

□ **Layers**

□ **Input Layer:** The input layer consists of neurons corresponding to each feature (deg malig, tumor size, inv nodes, age).

□ **Hidden Layers**: Add one or more hidden layers with a suitable number of neurons. The number of hidden layers and neurons can be determined based on experimentation and validation. (100,20)

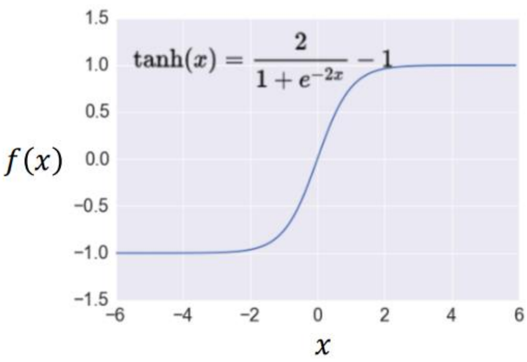
□ **Output Layer**: For binary classification (recurrence vs. no recurrence), the output layer typically has one neuron with a sigmoid activation function.

□ **Weights and Biases:**

- **Weights:** Parameters that are adjusted during training to influence the strength of the connection between neurons.
- **Biases:** Additional parameters that allow the activation function to be shifted, aiding in learning the optimal model.

□ **Activation Function:**

- Functions applied to the output of each neuron to introduce non-linearity. Common activation functions include sigmoid, ReLU, and tanh. We use tanh
- Tanh function: takes a real-valued number and "squashes" it into range between -1 and 1 ♣ Like sigmoid, tanh neurons saturate ♣ Unlike sigmoid, the output is zero-centered o It is therefore preferred than sigmoid ♣ Tanh is a scaled sigmoid: $\tanh(x) = 2 \cdot \sigma(2x) - 1$



$\mathbb{R}^n \rightarrow [-1,1]$

Fig (14)

Solver for weight optimization

- **Adam** (Adaptive Moment Estimation): stochastic gradient-based optimizer

combines insights from the momentum optimizers that accumulate the values of past gradients, and it also introduces new terms based on the second moment of the gradient o Similar to GD with momentum

Learning Process

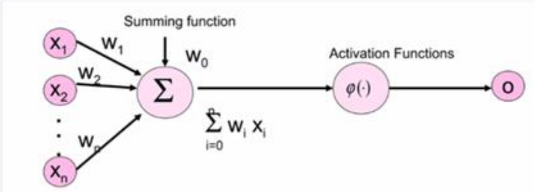


Fig (15)

Backpropagation:

- An algorithm used to update the weights and biases by calculating the gradient of the loss function with respect to each weight. The gradients are used to perform gradient descent or other optimization techniques to minimize the loss function.

• Modern NNs employ the backpropagation method for calculating the gradients of the loss function $\nabla \mathscr{L} \theta = \partial \mathscr{L} \partial \theta i$

♣ Backpropagation is short for "backward propagation" • For training NNs, forward propagation (forward pass) refers to passing the inputs x through the hidden layers to obtain the model outputs (predictions) y

♣ The loss $\mathscr{L} y, y$ function is then calculated

♣ Backpropagation traverses the network in reverse order, from the outputs y backward toward the inputs x to calculate the gradients of the loss $\nabla \mathscr{L} \theta$

♣ The chain rule is used for calculating the partial derivatives of the loss function with respect to the parameters θ in the different layers in the network

• Each update of the model parameters θ during training takes one forward and one backward pass (e.g., of a batch of inputs)

• Automatic calculation of the gradients (automatic differentiation) is available in all current deep learning libraries

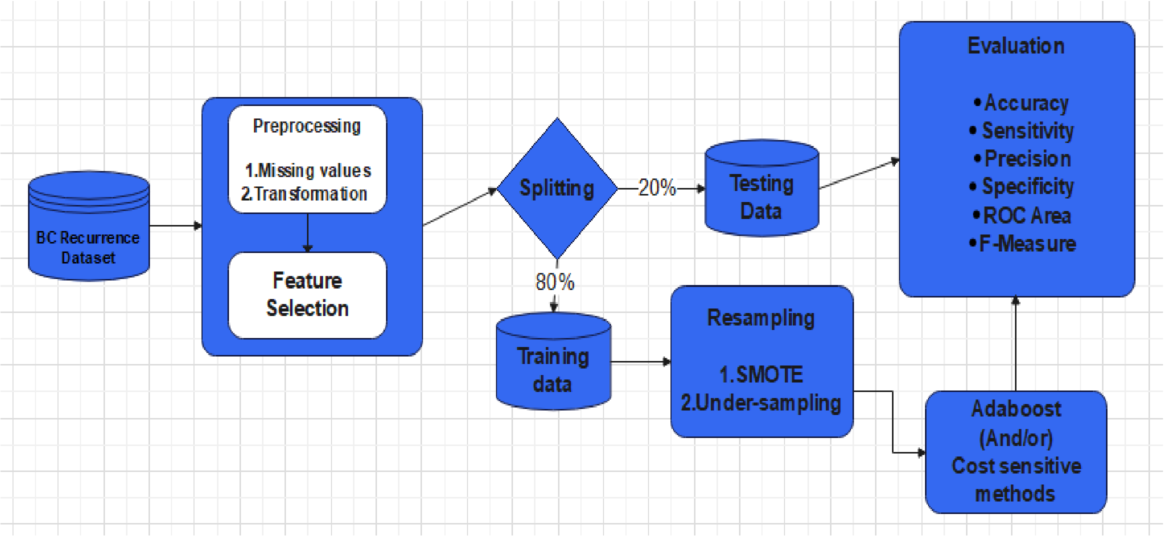
We use the **Multilayer Perceptron (MLPs):**

- A type of feedforward neural network with multiple hidden layers, often used for general-purpose tasks (Prediction of diseases)

Using an MLP with backpropagation to predict breast cancer recurrence involves preparing the data, designing and training the neural network, evaluating and tuning the model, and finally deploying it for practical use. This approach leverages the power of neural networks to capture complex patterns in the data, providing accurate and actionable predictions for clinicians.

Accuracy: 0.957 **the best classifier accuracy**

Chapter 6



System Architecture of Breast cancer recurrence **Fig (14)**

Model Testing:

Model testing is a crucial phase in the development of a predictive model for breast cancer recurrence. It involves evaluating the model using a separate testing dataset and various performance metrics to ensure it generalizes well to new data. By analyzing the confusion matrix and comparing performance metrics, we can refine the model and ensure its robustness, ultimately aiding clinicians in making accurate predictions about breast cancer recurrence based on features like deg malig, tumor size, inv nodes, and age.

Model Evaluation

Evaluating the performance of a machine learning model, particularly for predicting breast cancer recurrence, involves a comprehensive analysis using various metrics and techniques. Here's a detailed explanation:

Performance Metrics

1. **Accuracy:**
   - **Definition:** The proportion of correctly predicted instances (both true positives and true negatives) out of the total instances.
   - **Formula:** Accuracy=TP+TNTP+TN+FP+FN\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}Accuracy=TP+TN+FP+FNTP+TN
   - **Use Case:** Useful for balanced datasets but can be misleading for imbalanced datasets.
2. **Precision:**
   - **Definition:** The proportion of true positive predictions out of all positive predictions.

   - **Formula:** Precision=TPTP+FP\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}Precision=TP+FPTP
   - **Use Case:** Important when the cost of false positives is high, focusing on the accuracy of positive predictions.

3. **Recall (Sensitivity or True Positive Rate):**
   - **Definition:** The proportion of true positive predictions out of all actual positive instances.
   - **Formula:** Recall=TPTP+FN\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}Recall=TP+FNTP
   - **Use Case:** Important when the cost of false negatives is high, focusing on capturing as many actual positives as possible.
4. **F1 Score:**
   - **Definition:** The harmonic means of precision and recall, providing a balance between the two.
   - **Formula:** F1 Score=2×Precision×RecallPrecision+Recall\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}F1 Score=2×Precision+RecallPrecision×Recall
   - **Use Case:** Useful when the dataset is imbalanced and both precision and recall are important.
5. **ROC-AUC (Receiver Operating Characteristic - Area Under the Curve):**

- **Definition:** Measures the model's ability to distinguish between classes, plotting the true positive rate (recall) against the false positive rate.
- **AUC:** The area under the ROC curve, with values ranging from 0.5 (no discrimination) to 1 (perfect discrimination).
- **Use Case:** Provides a comprehensive measure of model performance across all classification thresholds.

Validation

1. **Cross-Validation:**
    - **Definition:** A technique where the dataset is split into k subsets (folds), and the model is trained and validated k times, each time using a different fold as the validation set and the remaining folds as the training set.
    - **Process:**
        1. Split the dataset into k folds.
        2. Train the model on k-1 folds and validate on the remaining fold.
        3. Repeat for all k folds.
        4. Average the performance metrics across all folds.
    - Provides a robust estimate of model performance, reducing the risk of overfitting and ensuring the model generalizes well to unseen data.
2. **Validation Set:**

   - **Use Case:** Helps in evaluating model performance during the development phase, ensuring that the model performs well on new, unseen data.

Confusion Matrix

- **Definition:** A table used to evaluate the performance of a classification model, showing the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).
- **Structure:**

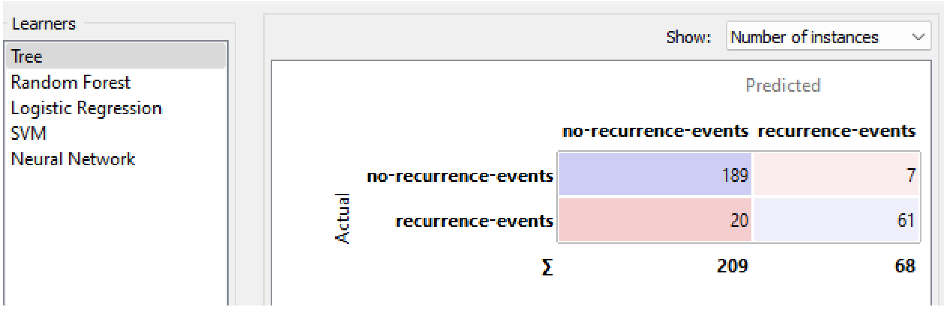
	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- **Analysis:**
   - **True Positives (TP):** Correctly predicted positive cases.
   - **True Negatives (TN):** Correctly predicted negative cases.
   - **False Positives (FP):** Incorrectly predicted positive cases (Type I error).
   - **False Negatives (FN):** Incorrectly predicted negative cases (Type II error).
- **Insights:**
   - **Precision:** Focus on minimizing FP, important in cases where false alarms are costly.
   - **Recall:** Focus on minimizing FN, crucial in cases where missing positive cases is costly.

- **F1 Score:** Balance between precision and recall, useful in scenarios with imbalanced classes.

Evaluating a model's performance involves using various metrics to understand its strengths and weaknesses. Accuracy provides a general measure, while precision, recall, and F1 score offer deeper insights into the model's ability to correctly predict positive and negative cases. ROC-AUC gives an overall performance measure across different thresholds. Cross-validation ensures the model's robustness and generalizability, and the confusion matrix provides a detailed breakdown of prediction outcomes, guiding further improvements.

Confusion Matrix: Create a confusion matrix to gain insights into the model's performance for each emotion class. It shows the number of true positives, false positives, true negatives, and false negatives for each emotion. This can help identify which emotions are more easily confused or misclassified by the model. In testing the model, accuracy is a commonly used metric to evaluate its performance. Accuracy represents the percentage of correctly classified samples out of the total number of samples in the test set. It provides a measure of how well the model can make correct predictions



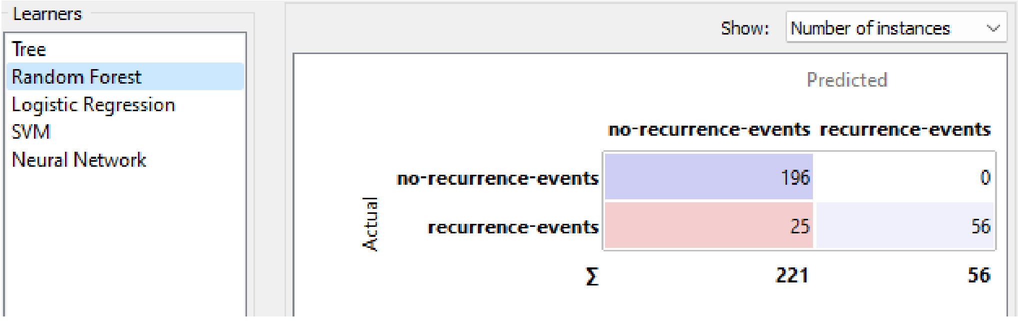
Learners

Tree
Random Forest
Logistic Regression
SVM
Neural Network

Show: Number of instances

		Predicted	
		no-recurrence-events	recurrence-events
Actual	no-recurrence-events	189	7
	recurrence-events	20	61
	∑	209	68

Confusion Matrix of the decision tree model



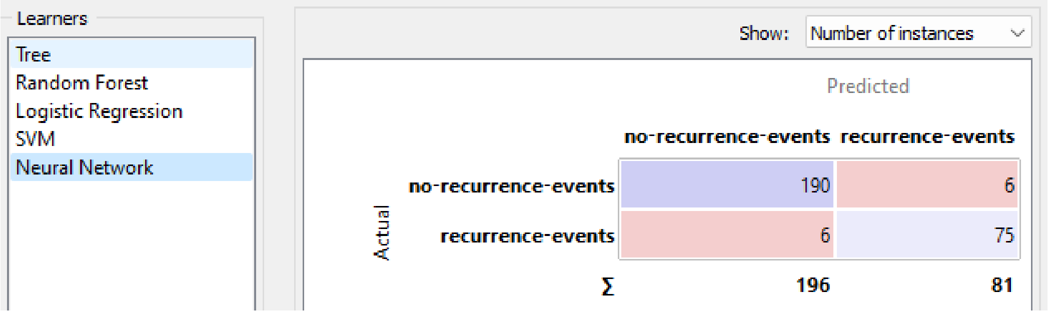
Learners
- Tree
- Random Forest
- Logistic Regression
- SVM
- Neural Network

Show: Number of instances

Predicted

Actual	no-recurrence-events	recurrence-events
no-recurrence-events	196	0
recurrence-events	25	56
Σ	221	56

Confusion Matrix of the random forest model



Learners
- Tree
- Random Forest
- Logistic Regression
- SVM
- Neural Network

Show: Number of instances

Predicted

Actual	no-recurrence-events	recurrence-events
no-recurrence-events	190	6
recurrence-events	6	75
Σ	196	81

Confusion Matrix of the neural network model

Comparison of Models

Test and Score:

Tests learning algorithms on data.

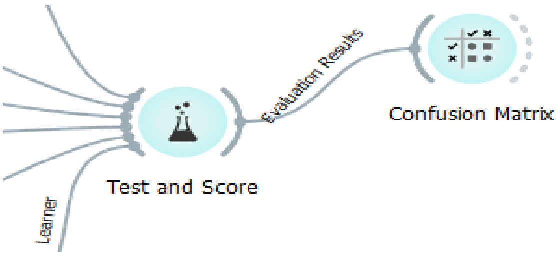


Fig (16)

Inputs

- Data: input dataset
- Test Data: separate data for testing
- Learner: learning algorithm(s)

Outputs

- Evaluation Results: results of testing classification algorithms

The widget tests learning algorithms. Different sampling schemes are available, including using separate test data. The widget does two things. First, it shows a table with different classifier performance measures, such as classification accuracy and area under the curve. Second, it outputs evaluation results, which can be used by other widgets for analyzing the performance of classifiers, such as ROC Analysis or Confusion Matrix.

The *Learner* signal has an uncommon property: it can be connected to more than one widget to test multiple learners with the same procedures.

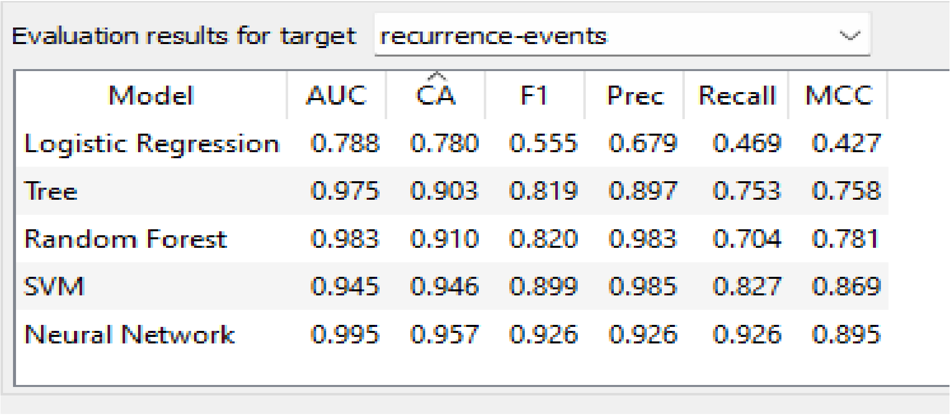
- The widget supports various sampling methods.
  - Cross-validation splits the data into a given number of folds (usually 5 or 10). The algorithm is tested by holding out examples from one fold at a time; the model is induced from other folds and examples from the held out fold are classified. This is repeated for all the folds.
  - **Cross validation by feature** performs cross-validation but folds are defined by the selected categorical feature from meta-features.
  - **Random sampling** randomly splits the data into the

training and testing set in the given proportion (e.g. 70:30); the whole procedure is repeated for a specified number of times.

  - **Leave-one-out** is similar, but it holds out one instance at a time, inducing the model from all others and then classifying the held-out instances. This method is obviously very stable, reliable... and very slow.
  - **Test on train data** uses the whole dataset for training and then for testing. This method practically always gives wrong results.
  - **Test on test data**: the above methods use the data from *Data* signal only. To input another dataset with testing examples (for instance from another file or some data selected in another widget), we select *Separate Test Data* signal in the communication channel and select Test on test data.

- For classification, *Target class* can be selected at the bottom of the widget. When *Target class* is (Average over classes), methods return scores that are weighted averages over all classes. For example, in the case of the classifier with 3 classes, scores are computed for class 1 as a target class, class 2 as a target class, and class 3 as a target class. Those scores are averaged with weights based on the class size to retrieve the final score.
- The widget will compute several performance statistics. A few are shown by default. To see others, right-click on the header and select the desired statistic.
  - Classification

- Area under ROC is the area under the receiver-operating curve.
- Classification accuracy is the proportion of correctly classified examples.
- F-1 is a weighted harmonic mean of precision and recall (see below).
- Precision is the proportion of true positives among instances classified as positive, e.g. the proportion of IRIS VIRGINICA correctly identified as Iris virginica.
- Recall is the proportion of true positives among all positive instances in the data, e.g. the number of sick among all diagnosed as sick.
- Specificity is the proportion of true negatives among all negative instances, e.g. the number of non-sick among all diagnosed as non-sick.
- LogLoss or cross-entropy loss takes into account the uncertainty of your prediction based on how much it varies from the actual label.
- Matthews correlation coefficient takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.
- Train time - cumulative time in seconds used for training models.
- Test time - cumulative time in seconds used for testing models.



Evaluation results for target recurrence-events

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.788	0.780	0.555	0.679	0.469	0.427
Tree	0.975	0.903	0.819	0.897	0.753	0.758
Random Forest	0.983	0.910	0.820	0.983	0.704	0.781
SVM	0.945	0.946	0.899	0.985	0.827	0.869
Neural Network	0.995	0.957	0.926	0.926	0.926	0.895

Fig (16)

Based on test and score and comparing between models, the neural network is the best for all other models

Model Deployment

Model deployment is the process of integrating a trained machine learning model into a production environment where it can be used to make predictions on new, unseen data. Here is a detailed guide on deploying a model for predicting breast cancer recurrence based on features like degree of malignancy (deg malig), tumor size, involved lymph nodes (inv nodes), and age.

1. Preparation

1. **Model Selection:**
   - Ensure the best-performing model is chosen after extensive training and testing. The model should have high accuracy, precision, recall, and other relevant metrics on the testing set.
   - Ex: A well-tuned neural network
2. **Exporting the Model:**
   - Serialize the trained model using a standard format like Pickle or joblib in Python.

2. Setting Up the Deployment Environment

Infrastructure:

Choose a deployment environment such as cloud services (AWS, Google Cloud, Azure), on-premises servers, or edge devices.

Ensure the infrastructure can handle the computational requirements of the model and can scale as needed.

Environment Configuration:

Set up the necessary software environment. This often involves creating a virtual environment with dependencies using tools like pip or conda.

Deploying a model for predicting breast cancer recurrence involves several steps, from model selection and API development to testing, monitoring, and maintenance. By carefully setting up and managing the deployment environment, ensuring data privacy, and maintaining continuous monitoring and updates, the deployed model can provide valuable predictions to aid clinicians in early intervention and personalized treatment planning.

Chapter 7

Orange

Introduction to Orange Tool

Orange is an open-source data visualization, machine learning, and data mining toolkit. It features a visual programming front-end for explorative data analysis and interactive data visualization, and can also be used as a Python library.

Orange allows you to:

– Show a data table and select features
–Read the data
– Compare learning algorithms and train predictors
–Visualize data elements

Some important terms are:

- **Widgets**: The various components present in Orange are known as widgets and they are divided into various categories like Data, Visualize, Model, Evaluate, and so on.

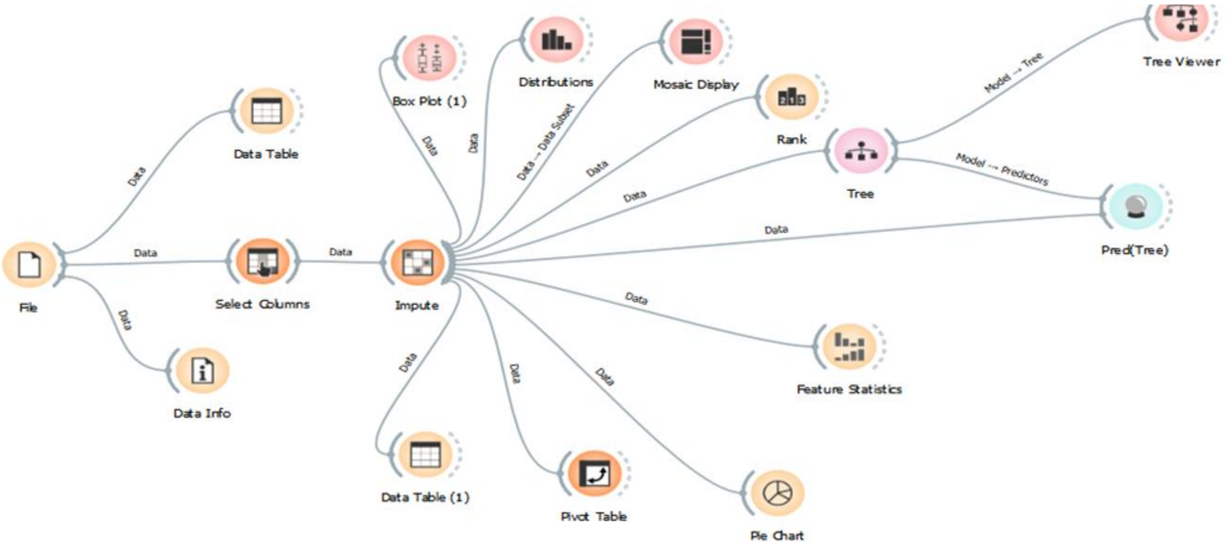
- **Workflows**: Orange workflows consist of components that read, process, and visualize data. We call them “widgets.” We place the widgets on a canvas. Widgets communicate by sending information along with a communication channel. An output from one widget is used as input to another.

The objective of Orange is to provide a platform for experiment-based selection, predictive modeling, and recommendation system. It is primarily used in bioinformatics, genomic research, biomedicine, and teaching. In education, it is used for providing better teaching methods for data mining and machine learning to students of biology, biomedicine, and informatics.

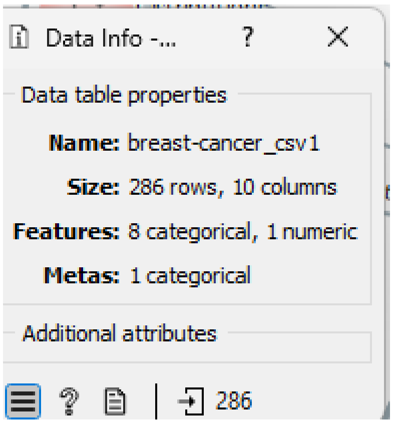
How to use workflows in Orange?

I have created a simple workflow wherein the inbuilt Iris dataset provided by Orange is being used. The workflow is such that data from the dataset is sent to the data table, to Distributions for creating a distribution and a Scatter Plot is plotted from the dataset. To create this workflow, we load the dataset using the File widget, and then flow between File-Data Info, File-Data Table, File-Distributions, and File-Scatter Plot is created.

We have created a simple workflow wherein the inbuilt breast cancer dataset . The workflow is such that data from the dataset is sent to the data table, to Distributions for creating a distribution and a Scatter Plot is plotted from the dataset. To create this workflow we load the dataset using the File widget, and then flow between File-Data Info **Fig (18)**



Data Info



Data table properties

Name: breast-cancer_csv1

Size: 286 rows, 10 columns

Features: 8 categorical, 1 numeric

Metas: 1 categorical

Additional attributes

Fig (19)

Other Tools:

Perspective Analytics:

Perspective analytics involves examining data from various angles to gain deeper insights and make informed decisions. In the context of a breast cancer recurrence detection project, perspective analytics can provide valuable insights into the factors influencing recurrence and help develop more effective predictive models.

□ Objective**:** Recommend actions to achieve desired outcomes.

□ Application**:** Suggesting personalized treatment plans and follow-up schedules based on the predicted risk of recurrence.

□ Tools**:** opt Perspective analytics in breast cancer recurrence detection provides a comprehensive approach to understanding and predicting recurrence by leveraging descriptive, diagnostic, predictive, and prescriptive analytics. By integrating these different perspectives, healthcare providers can make more informed decisions, improve patient outcomes, and optimize resource allocation. This holistic approach ensures that the models and insights derived from the data are robust, actionable, and continuously refined to adapt to new information and evolving healthcare medialization algorithms and simulation models.

Conclusion:

Detecting breast cancer recurrence is a critical task in oncology, as early detection can significantly improve patient outcomes. By leveraging various machine learning algorithms and data mining techniques, such as those provided by the Orange tool, we can build robust predictive models using features like age, involved lymph nodes (inv nodes), and degree of malignancy (deg malig).

Using machine learning algorithms and data mining techniques like those available in Orange, significant advancements can be made in predicting breast cancer recurrence. By effectively leveraging features such as age, involved lymph nodes, and degree of malignancy, we can build predictive models that provide valuable insights for early intervention and personalized treatment, ultimately improving patient outcomes. The ongoing evolution of these models, coupled with their integration into clinical workflows, holds great promise for the future of oncology.

Key Findings:

1. **Feature Importance:**
2. **Model Comparison:**
3. **Data Mining with Orange:**
   - **Ease of Use:** Orange's visual programming interface allows for easy implementation and comparison of different models.
   - **Interactive Exploration:** Enables quick adjustments and visualization of data, aiding in understanding and refining the model.
   - **Integration:** Supports integration of various machine learning techniques, enhancing the ability to experiment with different approaches.

Practical Implications:

1. **Early Intervention:** Accurate models help in identifying high-risk patients early, enabling timely interventions and personalized treatment plans.
2. **Resource Allocation:** Helps in optimizing resource allocation in healthcare by focusing on patients with higher recurrence risk.

3. **Clinical Decision Support:** Provides clinicians with valuable insights and predictive tools to aid in decision-making.

Future Directions:

1. **Model Enhancement:**
   - **Incorporating More Features:** Including additional clinical and genetic features could further improve model accuracy.
   - **Advanced Algorithms:** Exploring more advanced algorithms like deep learning or ensemble methods for better performance.
2. **Continuous Learning:**
   - **Regular Updates:** Continuously updating the models with new patient data to maintain and improve accuracy.
   - **Feedback Loops:** Implementing feedback mechanisms from clinical use to refine models.
3. **Real-World Implementation:**
   - **Integration with EHRs:** Seamless integration with electronic health records (EHRs) for real-time predictions.
   - **User Training:** Training healthcare providers to effectively use and interpret model predictions.

References: -

- Mojrian, S., Pinter, G., Hassannataj Joloudari, J., Felde, I., Nabipour, N., Nadai, L., & Mosavi, A., 2019. Hybrid Machine Learning Model of Extreme Learning Machine Radial basis function for Breast Cancer Detection and Diagnosis; a Multilayer Fuzzy Expert System. [PDF]

- Al-Quraishi, Tahsien Ali Hussein. (2019). *Predicting breast cancer risk, recurrence and survivability*

- Rabinovici-Cohen, S., M. Fernández, X., Grandal Rejo, B., Hexter, E., Hijano Cubelos, O., Pajula, J., Pölönen, H., Reyal, F., & Rosen-Zvi, M., 2022. Multimodal Prediction of Five-Year Breast Cancer Recurrence in Women Who Receive Neoadjuvant Chemotherapy. ncbi.nlm.nih.gov

- Kumar Mondol, R., K. A. Millar, E., Sowmya, A., & Meijering, E., 2024. BioFusionNet: Deep Learning-Based Survival Risk Stratification in ER+ Breast Cancer Through Multifeature and Multimodal Data Fusion. [PDF]

- Y. Ling, A., W. Kurian, A., L. Caswell-Jin, J., W. Sledge, G., H. Shah, N., & R. Tamang, S., 2019. A Semi-Supervised Machine Learning Approach to Detecting Recurrent Metastatic Breast Cancer Cases Using Linked Cancer Registry and Electronic Medical Record Data. [PDF]

- https://www.techtarget.com/searchbusinessanalytics/definition/logisc-regression#:~:text=Logistic%20regression%2C%20also%20known%20as,or%20more%20existing%20independent%20variables.

- https://www.geeksforgeeks.org/understanding-logistic-regression/

- https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/

- https://www.geeksforgeeks.org/support-vector-machine-algorithm/

- https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

- https://www.freecodecamp.org/news/svm-machine-learning-tutorial-what-is-the-support-vector-machine-algorithm-explained-with-code-examples/

- https://www.datacamp.com/tutorial/svm-classification-scikit-learn-python

- https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

- hen SI, Tseng HT, Hsieh CC. Evaluating the impact of soy compounds on breast cancer using the data mining approach. *Food & function* . 2020;11(5):4561–70. doi: 10.1039/C9FO00976K. [PubMed] [CrossRef] [Google Scholar]

- Chaurasia V, Pal S, Tiwari BB. Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms & Computational Technology* . 2018;12(2):119–26. doi: 10.1177/1748301818756225. [CrossRef] [Google Scholar]

- 8. Burnside ES, Liu J, Wu Y, Onitilo AA, McCarty CA, Page CD, et al. Comparing Mammography Abnormality Features to Genetic Variants in the Prediction of Breast Cancer in Women Recommended for Breast Biopsy. *Acad Radiol* . 2016;23(1):62–9. doi: 10.1016/j.acra.2015.09.007. [PMC Free Article] [PMC free article] [PubMed] [CrossRef] [Google Scholar]

- Feld SI, Fan J, Yuan M, Wu Y, Woo KM, Alexandridis R, Burnside ES. Utility of Genetic Testing in Addition to Mammography for Determining Risk of Breast Cancer Depends on Patient Age. *AMIA Jt Summits Transl Sci Proc* . 2018;2017:81–90. [PMC Free Article] [PMC free article] [PubMed] [Google Scholar]

- othi N, Husain W. Data mining in healthcare-a review. *Procedia Computer Science* . 2015;72:306–13. doi: 10.1016/j.procs.2015.12.145. [CrossRef] [Google Scholar]

- ent CK, Bassett LW, D'Orsi CJ, Sayre JW. The positive predictive value of BI-RADS microcalcification descriptors and final assessment categories. *AJR Am J Roentgenol* . 2010;194(5):1378–83. doi: 10.2214/AJR.09.3423. [PubMed] [CrossRef] [Google Scholar]

- Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction. *Radiology* . 2019;292(1):60–6. doi: 10.1148/radiol.2019182716. [PubMed] [CrossRef] [Google Scholar]

- Dai B, Chen RC, Zhu SZ, Zhang WW. Using random forest algorithm for breast cancer diagnosis. 2018 International Symposium on Computer, Consumer and Control (IS3C); Taichung, Taiwan: IEEE; 2018. p. 449-52. doi: 10.1109/IS3C.2018.00119. [CrossRef] [Google Scholar]