UNIT 5: Introduction to Big Data and Data Analytics

Title: Introduction to Big Data and Data Analytics	**Approach**: Team discussion, Web search

Summary: Students will delve into the world of **Big Data**, a game-changer in today's digital age. Students gain insights into the various types of data and their unique characteristics, equipping them to understand how this vast information is managed and analysed. The journey continues as students discover the real-world applications of Big Data and Data Analytics in diverse fields, witnessing how this revolutionary concept is transforming how we approach data analysis to unlock new possibilities.

Learning Objectives:

1. Students will develop an understanding of the concept of Big Data and its development in the new digital era.
2. Students will appreciate the role of big data in AI and Data Science.
3. Students will learn to understand the features of Big Data and how these features are handled in Big Data Analytics.
4. Students will appreciate its applications in various fields and how this new concept has evolved to bring new dimensions to Data Analysis.
5. Students will understand the term mining data streams.

Key Concepts:

1. Introduction to Big Data
2. Types of Big Data
3. Advantages and Disadvantages of Big Data
4. Characteristics of Big Data
5. Big Data Analytics
6. Working on Big Data Analytics
7. Mining Data Streams
8. Future of Big Data Analytics

Learning Outcomes:

Students will be able to –

1. Define Big Data and identify its various types.
2. Evaluate the advantages and disadvantages of Big Data.
3. Recognize the characteristics of Big Data.
4. Explain the concept of Big Data Analytics and its significance.
5. Describe how Big Data Analytics works.
6. Exploring the future trends and advancements in Big Data Analytics.

Prerequisites: Understanding the concept of data and reasonable fluency in the English language.

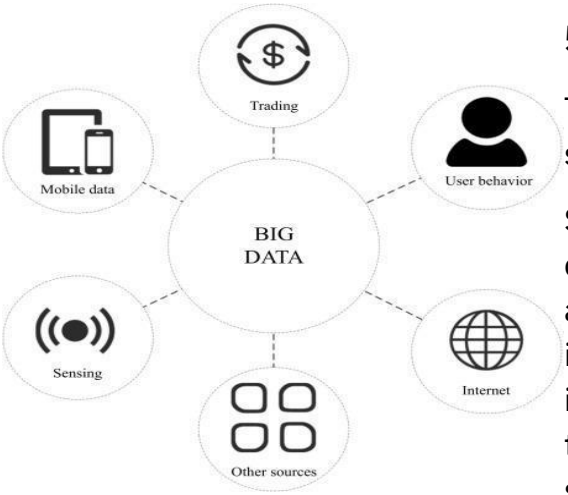


Fig. 5.1 Sources of Big Data

5.1. What is Big Data?

To understand Big Data, let us first understand small data.

Small data refers to datasets that are easily comprehendible by people as they are easily accessible, informative, and actionable, this makes it ideal for individuals and businesses to find useful information and make better choices in everyday tasks. For example, a small store might track daily sales to decide what products to restock.

Big Data refers to extremely large and complex datasets that regular computer programs and databases cannot handle. It comes from three main sources: transactional data (e.g., online purchases), machine data (e.g., sensor readings), and social data (e.g., social media posts). To analyze and use Big Data effectively, special tools and techniques are required. These tools help organizations find valuable insights hidden in the data, which lead to innovations and better decision-making. For example, companies like Amazon and Netflix use Big Data to recommend products or shows based on users' past activities.

5.2. Types of Big Data

Fig. 5.2

Aspect	Structured Data	Semi-Structured Data	Unstructured Data
Definition	Quantitative data with a defined structure	A mix of quantitative and qualitative properties	No inherent structures or formal rules
Data Model	Dedicated data model	May lack a specific data model	Lacks a consistent data model

Aspect	Structured Data	Semi-Structured Data	Unstructured Data
Organization	Organized in clearly defined columns	Less organized than structured data	No organization exhibits variability over time
Accessibility	Easily accessible and searchable	Accessible but may be harder to analyze	Accessibility depends on the specific data format
Examples	Customer information, transaction records, product directories Structured data ID | Name | Age | Degree 1 | John | 18 | B.Sc. 2 | David | 31 | Ph.D. 3 | Robert | 51 | Ph.D. 4 | Rick | 26 | M.Sc. 5 | Michael | 19 | B.Sc.	XML files, CSV files, JSON files, HTML files, semi-structured documents Semi-structured data <University> <Student ID="1"> <Name>John</Name> <Age>18</Age> <Degree>B.Sc.</Degree> </Student> <Student ID="2"> <Name>David</Name> <Age>31</Age> <Degree>Ph.D. </Degree> </Student> .... </University>	Audio files, images, video files, emails, PDFs, social media posts Unstructured data The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

5.3. Advantages and Disadvantages of Big Data:

Big Data is a key to modern innovation. It has changed how organizations analyze and use information. While it offers great benefits, it also comes with challenges that affect its use in different industries. In this section, we will be discussing a few pros and cons of big data.

Advantages:

- **Enhanced Decision Making**: Big Data analytics empowers organizations to make data-driven decisions based on insights derived from large and diverse datasets.
- **Improved Efficiency and Productivity**: By analyzing vast amounts of data, businesses can identify inefficiencies, streamline processes, and optimize resource allocation, leading to increased efficiency and productivity.
- **Better Customer Insights**: Big Data enables organizations to gain a deeper understanding of customer behavior, preferences, and needs, allowing for personalized marketing strategies and improved customer experiences.
- **Competitive Advantage**: Leveraging Big Data analytics provides organizations with a competitive edge by enabling them to uncover market trends, identify opportunities, and stay ahead of competitors.
- **Innovation and Growth**: Big Data fosters innovation by facilitating the development of new products, services, and business models based on insights derived from data analysis, driving business growth and expansion.

Disadvantages:

- **Privacy and Security Concerns**: The collection, storage, and analysis of large volumes of data raise significant privacy and security risks, including unauthorized access, data breaches, and misuse of personal information.
- **Data Quality Issues**: Ensuring the accuracy, reliability, and completeness of data can be challenging, as Big Data often consists of unstructured and heterogeneous data sources, leading to potential errors and biases in analysis.
- **Technical Complexity**: Implementing and managing Big Data infrastructure and analytics tools require specialized skills and expertise, leading to technical challenges and resource constraints for organizations.
- **Regulatory Compliance**: Organizations face challenges in meeting data protection laws like GDPR (General Data Protection Regulation) and The Digital Personal Data Protection Act, 2023. These laws require strict handling of personal data, making compliance essential to avoid legal risks and penalties.
- **Cost and Resource Intensiveness**: The cost of acquiring, storing, processing, and analyzing Big Data, along with hiring skilled staff, can be high. This is especially challenging for smaller organizations with limited budgets and resources.

Activity: Find the sources of big data using the link UNSTATS

5.4. Characteristics of Big Data

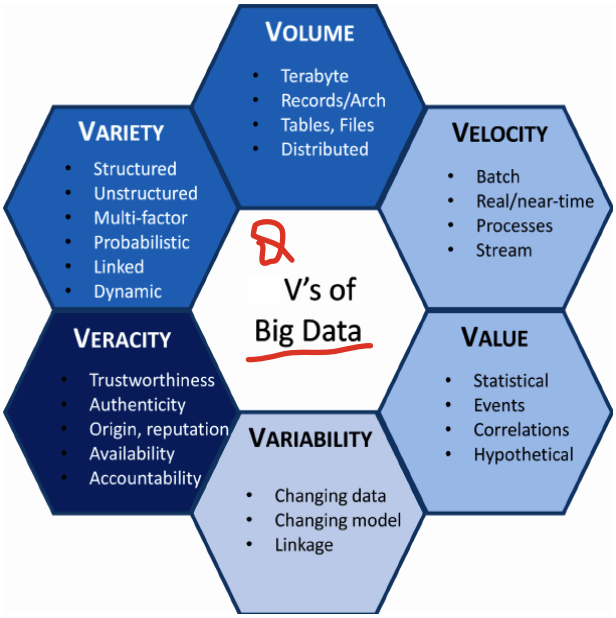


Fig. 5.3 Characteristics of Big Data

The "**characteristics of Big Data**" refer to the defining attributes that distinguish large and complex datasets from traditional data sources. These characteristics are commonly described using the **"3Vs" framework: Volume, Velocity, and Variety.** The **6Vs framework** provides a holistic view of Big Data, emphasizing not only its volume, velocity, and variety but also its veracity, variability, and value. Understanding and addressing these six dimensions are essential for effectively managing, analyzing, and deriving value from Big Data in various domains.

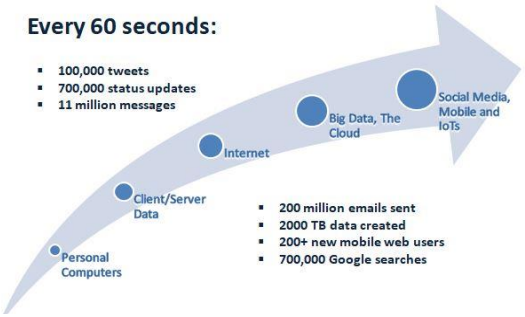


Fig. 5.4 Speed of data generation from various sources

5.4.1. Velocity: Velocity refers to the speed at which data is generated, delivered, and analyzed. In the present world, where millions of people are accessing and storing information online, the speed at which the data gets stored or generated is huge. For example: Google alone generates more than 40,000 search queries per second. See the statistics in the picture provided. Isn't it huge!

5.4.2. Volume: Every day a huge volume of data is generated as the number of people using online platforms has increased exponentially. Such a huge volume of data is considered Big Data. Typically, if the data volume exceeds gigabytes, it falls into the realm of big data. This volume can range from petabytes to terabytes or even exabytes, based on surveys conducted by various organizations. According to the latest estimates, 328.77 million terabytes of data are created each day.

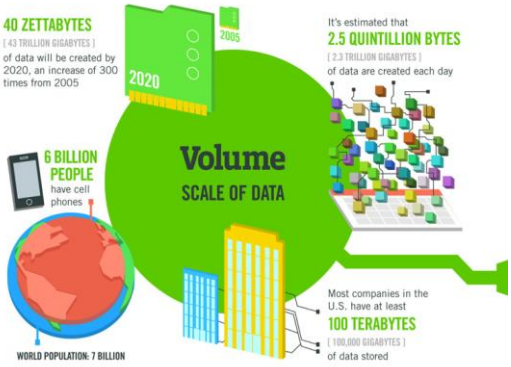


Fig.5.5 Volume of data

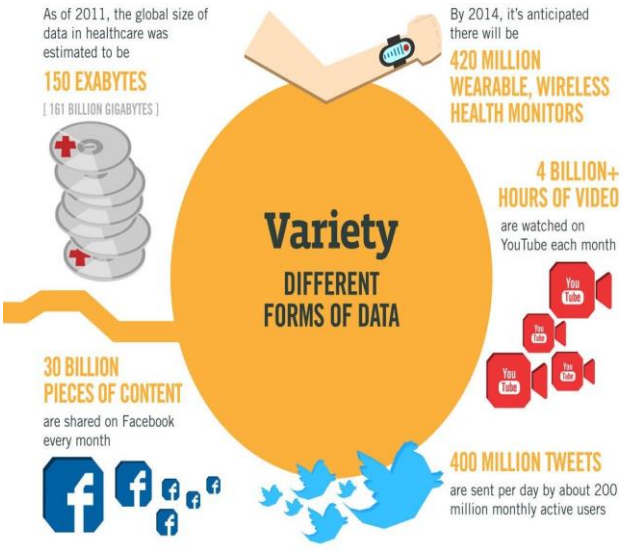


Fig.5.6 Varieties in Big data

5.4.3. Variety: Big data encompasses data in various formats, including structured, unstructured, semi-structured, or highly complex structured data. These can range from simple numerical data to complex and diverse forms such as text, images, audio, videos, and so on. Storing and processing unstructured data through RDBMS is challenging. However, unstructured data often provides valuable insights that structured data cannot offer. Additionally, the variety of data sources within big data provides information on the diversity of data.

5.4.4. Veracity: Veracity is a characteristic in Big Data related to consistency, accuracy, quality, and trustworthiness. Not all data that undergoes processing holds value. Therefore, it is essential to clean data effectively before storing or processing it, especially when dealing with massive volumes. Veracity addresses this aspect of big data, focusing on the accuracy and reliability of the data source and its suitability for analytical models.

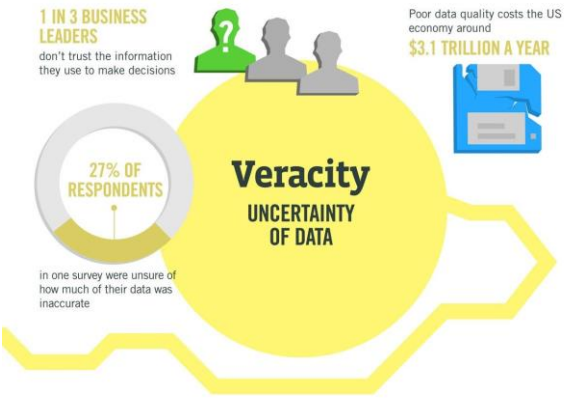


Fig. 5.7

Fig. 5.8 The value of Big Data

5.4.5. Value: The goal of big data analysis lies in extracting business value from the data. Hence, the business value derived from big data is perhaps its most critical characteristic. Without obtaining valuable insights, the other characteristics of big data hold little significance. So, in simple terms Value of Big Data refers to the benefits the big data can provide.

5.4.6. Variability: This refers to establishing if the contextualizing structure of the data stream is regular and dependable even in conditions of extreme unpredictability. It defines the need to get meaningful data considering all possible circumstances.

Fig. 5.9

Case Study: How a Company Uses 3V and 6V Frameworks for Big Data
Company: An OTT Platform 'OnDemandDrama'

3V Framework:

Volume: OnDemandDrama processes huge amounts of data from millions of users, including watch history, ratings, searches, and preferences to offer personalized content recommendations.

Velocity: Data is processed in real-time, allowing OnDemandDrama to immediately adjust recommendations, track the patterns of the users, and offer trending content based on their activity.

Variety: The platform handles diverse data such as user profiles, watch lists, video content, and user reviews which are categorized as structured, semi-structured, and unstructured data.

6V Framework:

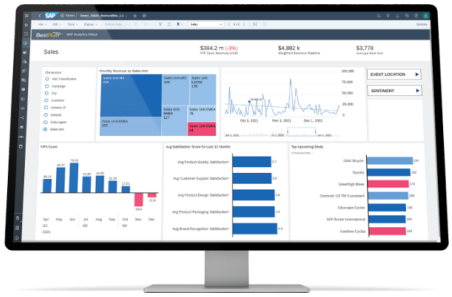
Along with the above 3 V of big data, the 6V Framework involves 3 more features of big data named Veracity, Value, and Variability.

Veracity: OnDemandDrama filters out irrelevant or low-quality data (such as incomplete profiles) to ensure accurate content recommendations.
Value: OnDemandDrama uses the data to personalize user experiences, driving engagement and retention by recommending shows and movies that match individual tastes.
Variability: OnDemandDrama handles changes or inconsistencies in data streams caused by factors like user behavior, trends, or any other external events. For example, user preferences can vary based on region, time, or trends.
By using the 3V and 6V frameworks, OnDemandDrama can manage, process, and derive valuable insights from its Big Data, which enhances customer satisfaction and drives business decisions.

5.5. Big Data Analytics

Data Analytics

Data analytics involves analyzing datasets to uncover insights, trends, and patterns. It can be applied to datasets of any size, from small to moderate volumes. Technologies commonly used in data analytics include statistical analysis software, data visualization tools, and relational database management systems (RDBMS).



Big data analytics uses advanced analytic techniques against huge, diverse datasets that include structured, semi-structured, and unstructured data, from different sources, and in various sizes from terabytes to zettabytes.

Big-Data Analytics encompasses the methodologies, tools, and practices involved in analyzing and managing data, covering tasks such as data collection, organization, and storage. The primary objective of data analytics is to utilize statistical analysis and technological methods to uncover patterns and address challenges. In the business realm, big data analytics has gained significance as a means to assess and refine business processes, as well as enhance decision-making and overall business performance. It provides valuable insights and forecasts that help businesses make informed decisions to improve their operations and outcomes. Different types of Big Data Analytics can help businesses and organizations find insights from large and complex datasets. Some of the common types are: Descriptive analytics, Diagnostic analytics, Predictive analytics, and Prescriptive analytics, which we have discussed in Unit 2 of Data Science Methodology.

Big Data Analytics emerges as a consequence of four significant global trends:

1. **Moore's Law**: The exponential growth of computing power as per Moore's Law has enabled the handling and analysis of massive datasets, driving the evolution of Big Data Analytics.
2. **Mobile Computing**: With the widespread adoption of smartphones and mobile devices, access to vast amounts of data is now at our fingertips, enabling real-time connectivity and data collection from anywhere.
3. **Social Networking**: Platforms such as Facebook, Foursquare, and Pinterest facilitate extensive networks of user-generated content, interactions, and data sharing, leading to the generation of massive datasets ripe for analysis.
4. **Cloud Computing**: This paradigm shift in technology infrastructure allows organizations to access hardware and software resources remotely via the Internet on a pay-as-you-go basis, eliminating the need for extensive on-premises hardware and software investments.

5.6. Working on Big Data Analytics

Big data analytics involves collecting, processing, cleaning, and analyzing enormous datasets to improve organizational operations. The working process of big data analytics includes the following steps –

Step 1. Gather data Each company has a unique approach to data collection. Organizations can now collect structured and unstructured data from various sources, including cloud storage, mobile apps, and IoT sensors.
Step 2. Process Data Once data is collected and stored, it must be processed properly to get accurate results on analytical queries, especially when it's large and unstructured. Various processing options are available: • **Batch processing** which looks at large data blocks over time. • **Stream processing** looks at small batches of data at once, shortening the delay time between collection and analysis for quicker decision-making.
Step 3. Clean Data Scrubbing all data, regardless of size, improves quality and yields better results. Correct formatting and elimination of duplicate or irrelevant data are essential. Erroneous and missing data can lead to inaccurate insights.
Step 4. Analyze Data Getting big data into a usable state takes time. Once it's ready, advanced analytics processes can turn big data into big insights.

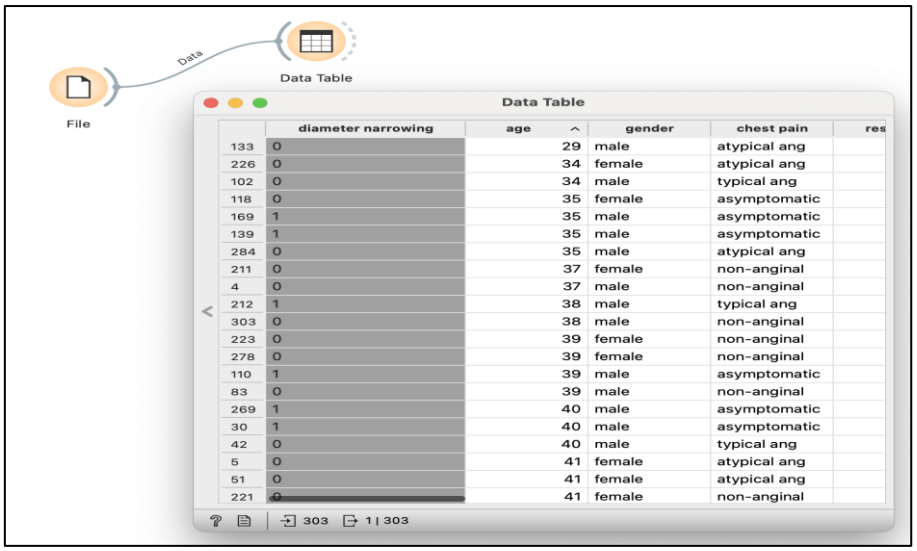
Example: **Data Analytics Tools – Tableau, APACHE Hadoop, Cassandra, MongoDB, SaS**

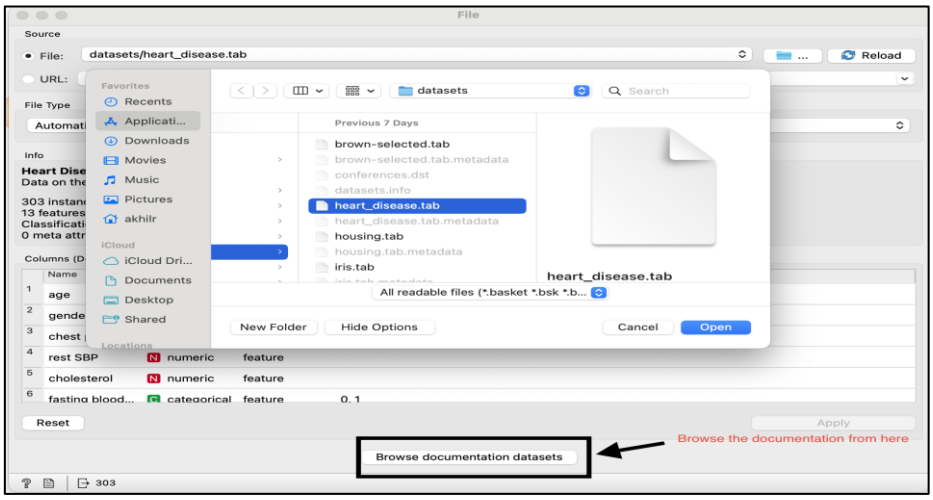
Using Orange Data Mining for Big Data Analytics

We will explore how big data analysis can be performed using Orange Data Mining.

Step 1: Gather Data

1. Use the **File** widget to load data into Orange.
2. Load the desired dataset. For demonstration, we will use the built-in **Heart Disease** dataset.





It is important to carefully study the dataset and understand the **features** and **target** variable.

- **Features**: age, gender, chest pain, resting blood pressure (rest_spb), cholesterol, resting ECG (rest_ecg), maximum heart rate (max_hr), etc.
- **Target**: **diameter narrowing**.

> If the value for **diameter narrowing** is **1**, it signifies significant narrowing of the arteries, which is a risk factor for heart disease. If the value is **0**, it indicates healthier arteries with minimal or no narrowing.

Step 2: Process Data

Data processing involves preparing the data for accurate analysis. There are two methods:

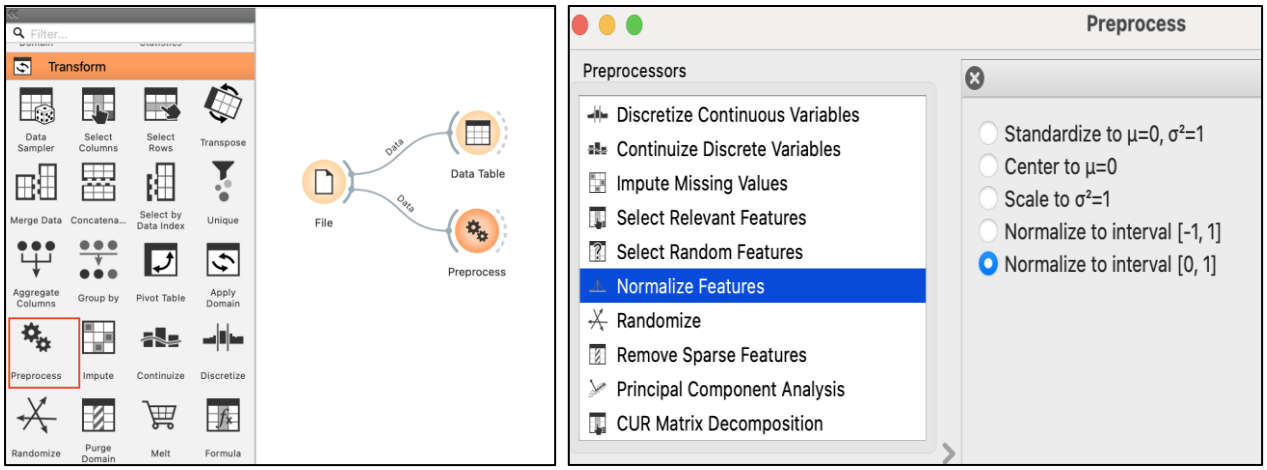
1. **Batch Processing**: Use the **Preprocess** widget to normalize large chunks of structured data at once.
2. **Stream Processing (near-real-time)**: While Orange does not natively support live stream data, you can incrementally process smaller subsets of the data in parallel workflows.

Here, we will focus on the **Normalization** technique.

Normalization in data preprocessing refers to scaling numerical values to a specific range (e.g., 0–1 or -1–1), making them comparable and improving the performance of machine learning algorithms.

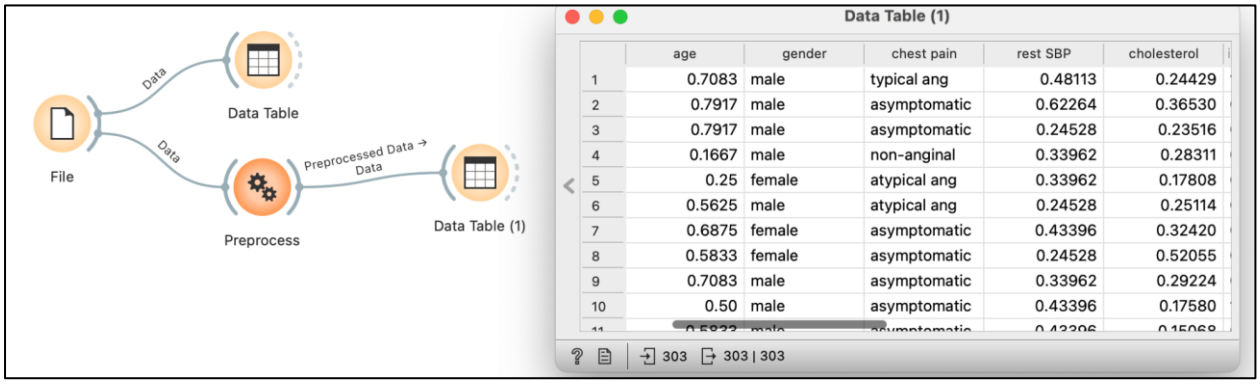
Step 2.1: Normalize Data

1. Connect the **Preprocess** widget to the **File** or **Data Table** widget.
2. Double-click on the **Preprocess** widget and select **"Normalize Features"**.
3. Choose an interval, such as **0–1** or **-1–1**.



Step 2.2: Verify Normalized Data

1. Connect the **Data Table** widget to the **Preprocess** widget.
2. Open the Data Table to observe the differences in values.



You will see that all numerical values are now scaled between **0 and 1**.

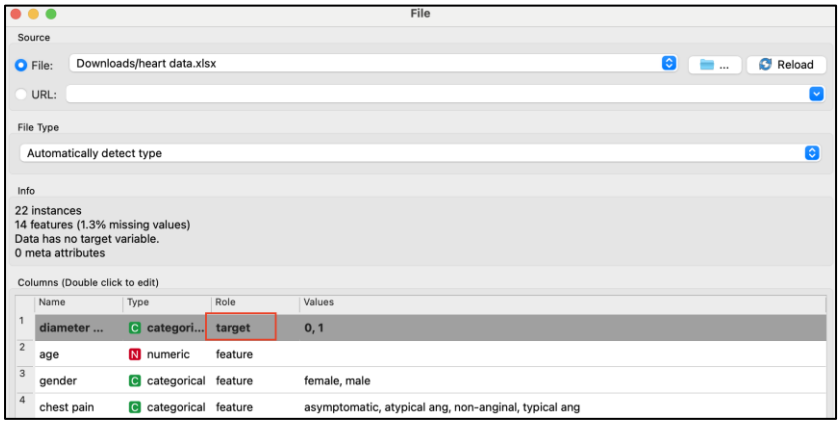
Step 3: Clean Data

Data cleaning is essential to ensure quality results. We will use the **Impute** widget to handle missing values by replacing them with the **mean**, **median**, **mode**, or a custom value. In this data we all can see that some values are missing in the figure below. This missing value data set is being saved as heart data.xlsx in the computer folder.

	A	B	C	D	E	F	G	H	I	J	K	L
1	diameter nai	age	gender	chest pain	rest SBP	cholestrol	fasting blooc	rest ECG	max HR	exerc incd ar	ST by exercis	slope peak e m
2	0	29	male	atypical ang	130	204	0	left vent hyp	202	0	0	upsloping
3	0	34	female	atypical ang	118		0	normal	192	0	0.7	upsloping
4	0	34	male	typical ang		182	0	left vent hyp	174	0	0	upsloping
5	0	35	female	asymptomat	138	183	0	normal	182	0	1.4	upsloping
6	1	35	male	asymptomat	126	282	0	left vent hyp	156	1	0	upsloping
7	1	35	male	asymptomat	120	198		normal	130	1	1.6	flat
8	0	35	male	atypical ang	122	192	0	normal	174	0	0	upsloping
9	0	37	female	non-anginal	120	215	0	normal	170	0	0	upsloping
10	0	37	male	non-anginal	130	250	0	normal	187	0	3.5	downsloping
11	1	38	male	typical ang	120	231	0	normal	182	1	3.8	flat
12	0	38	male	non-anginal	138	175	0	normal	173	0	0	upsloping ?
13	0	39	female	non-anginal	94	199	0	normal	179	0	0	upsloping
14	0	39	female	non-anginal	138	220	0	normal	152	0	0	flat

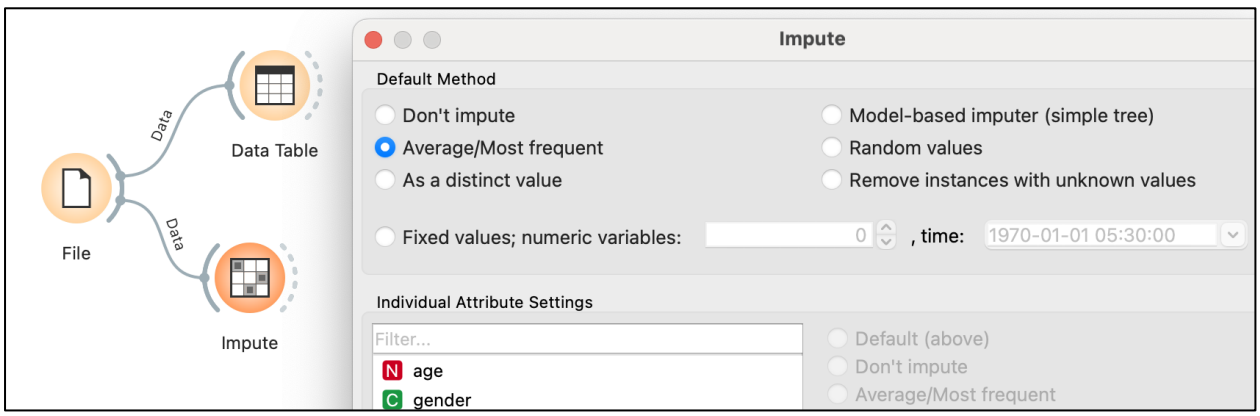
Step 3.1: Upload Data

1. Use the **File** widget to upload a dataset with missing values.
2. Assign the role of **"Target"** to the feature you want to predict.



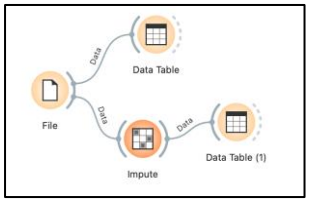
Step 3.2: Handle Missing Values

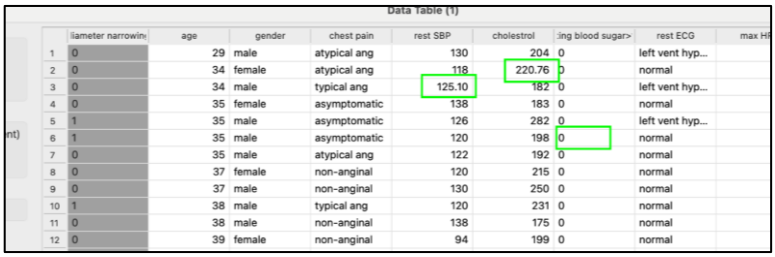
1. Connect the **Impute** widget to the **File** widget.
2. Double-click the **Impute** widget and select an imputation strategy: Average (mean), Most frequent (mode), Fixed value, Random value



Step 3.3: Verify Cleaned Data

1. Connect the **Data Table** widget to the **Impute** widget.
2. Open the Data Table to confirm the missing values have been replaced.




Data Table (1)

	liameter narrowin	age	gender	chest pain	rest SBP	cholestrol	ing blood sugar>	rest ECG	max HF
1	0	29	male	atypical ang	130	204	0	left vent hyp...	
2	0	34	female	atypical ang	118	220.76	0	normal	
3	0	34	male	typical ang	125.10	182	0	left vent hyp...	
4	0	35	female	asymptomatic	138	183	0	normal	
5	1	35	male	asymptomatic	126	282	0	left vent hyp...	
6	1	35	male	asymptomatic	120	198	0	normal	
7	0	35	male	atypical ang	122	192	0	normal	
8	0	37	female	non-anginal	120	215	0	normal	
9	0	37	male	non-anginal	130	250	0	normal	
10	1	38	male	typical ang	120	231	0	normal	
11	0	38	male	non-anginal	138	175	0	normal	
12	0	39	female	non-anginal	94	199	0	normal	

Missing values are now filled with the chosen method (e.g., average values).

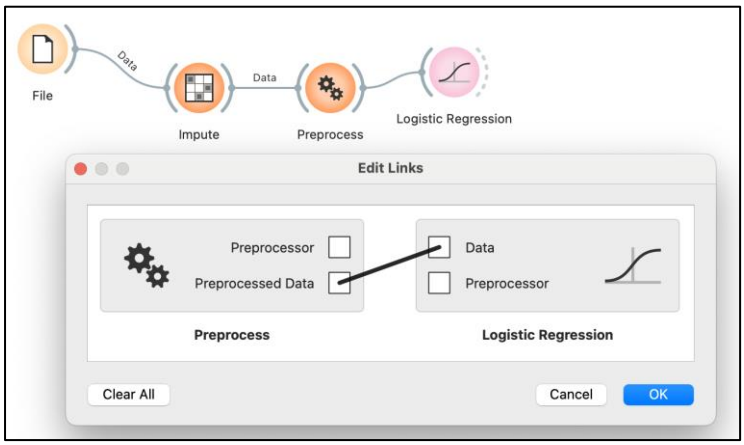
Step 4: Analyze Data

After cleaning, Orange provides various advanced analytics tools to extract insights:

- **K-Means**: For segmenting data into clusters.
- **Logistic Regression / Decision Tree**: For predicting outcomes using labeled data.
- **Scatter Plot / Box Plot / Heat Map**: For visualizing data patterns and relationships.

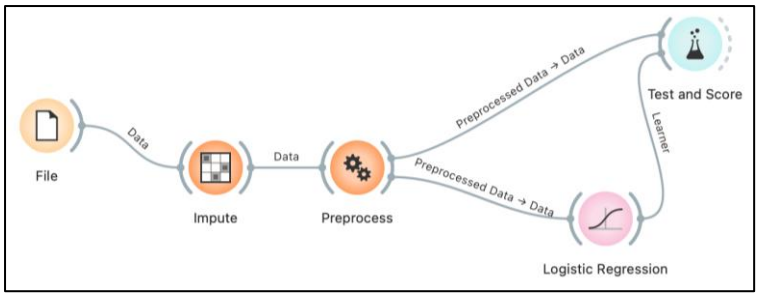
Step 4.1: Build a Logistic Regression Model

1. Drag and drop the **Logistic Regression** widget.
2. Connect it to the cleaned and normalized data.



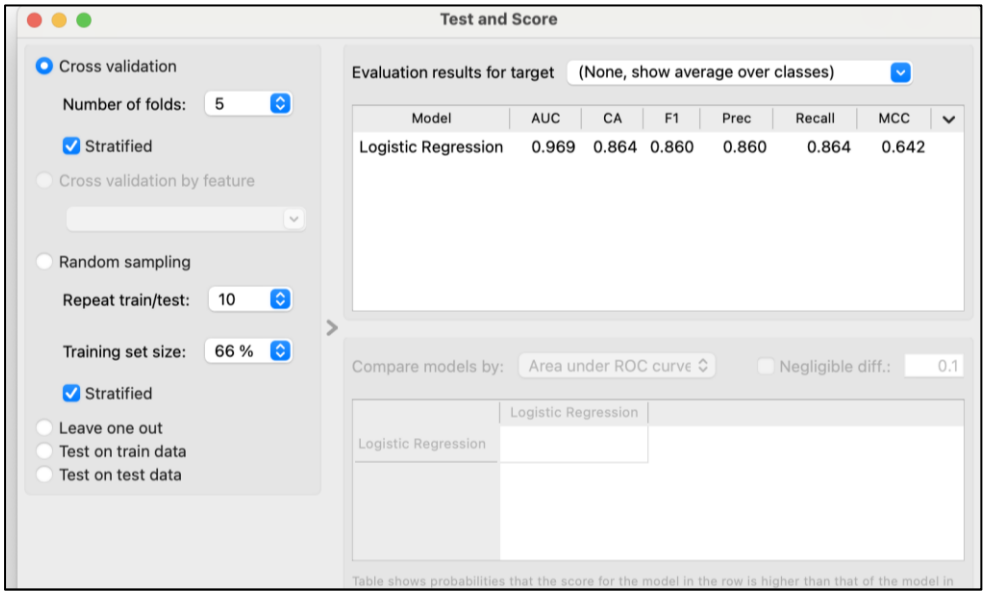
Step 4.2: Test the Model

1. Add the **Test and Score** widget.
2. Connect the **Test and Score** widget to:
   a. The **Logistic Regression** widget (learner data)
   b. The processed data.



Step 4.3: Choose a Validation Method

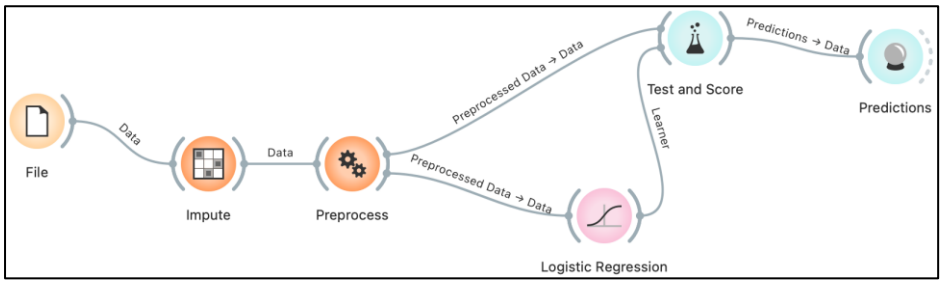
1. Double-click the **Test and Score** widget. Select a validation method (e.g., **Cross-Validation**).



Step 4.4: Generate Predictions

Connect the **Predict** widget to the **Test and Score** widget.

Check the predictions generated using the Logistic Regression model.



5.7. Mining Data Streams

To understand mining data streams, we first understand what data stream is. A data stream is a continuous, real-time flow of data generated by various sources. These sources can include sensors, satellite image data, Internet and web traffic, etc.

Mining data streams refers to the process of extracting meaningful patterns, trends, and knowledge from a continuous flow of real-time data. Unlike traditional data mining, it processes data as it arrives, without storing it completely. An example of an area where data

stream mining can be applied is website data. Websites typically receive continuous streams of data daily. For instance, a sudden spike in searches for "election results" on a particular day might indicate that elections were recently held in a region or highlight the level of public interest in the results.

5.8. Future of Big Data Analytics

The future of Big Data Analytics is highly influenced by several key technological advancements that will shape the way data is processed and analyzed. A few of them are:

Real-Time Analytics: It will allow businesses to process data instantaneously, providing immediate insights for decision-making and enabling actions based on live data, such as monitoring customer behavior or tracking supply chain activities.

Development of Advanced Models in Predictive Analytics: Predictive analytics will evolve with the integration of more sophisticated machine learning and AI algorithms, enabling organizations to forecast trends and behaviors with greater precision.

Quantum Computing: Quantum computing promises to revolutionize Big Data analytics by offering unprecedented processing power. Quantum computers will be able to solve complex problems much faster than classical computers.

Activity 1: Note – <u>This is a research-based group activity</u>

i) Watch this video using the link https://www.youtube.com/watch?v=37x5dKW-X5U

ii) Form a group, explore the applications of Big Data & Data Analytics in the following fields, and fill in the table given below:

Field	Video resource	Insights are drawn about this field and its futuristic development
Education		
Environmental Science		
Media and Entertainment		