

Experiment 1

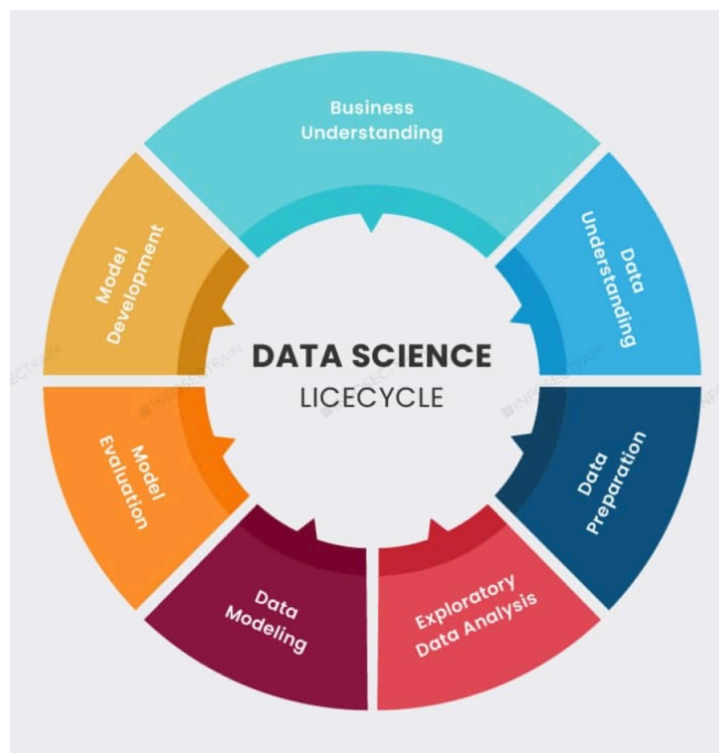
Aim: To study data science tools and programming languages.

Theory:

Data science is the study of data used to extract meaningful insights for business decisions. It combines mathematics, computing and domain knowledge to solve real-world problems and uncover hidden patterns.

It processes raw data to address business challenges and predict future trends.

- Steps involved in Data Science
 1. **Data Collection:** Gathering raw data from various sources, such as databases, sensors or user interactions.
 2. **Data Cleaning:** Ensuring the data is accurate, complete and ready for analysis.
 3. **Data Analysis:** Applying statistical and computational methods to identify patterns, trends or relationships.
 4. **Data Visualization:** Creating charts, graphs and dashboards to present findings clearly.
 5. **Decision-Making:** Using insights to inform strategies, create solutions or predict outcome



Why is Data Science important?

- a. **Helps Business in Decision-Making:** By analyzing data, businesses can understand trends and make informed choices that reduce risks and maximize profits.
- b. **Improves Efficiency:** Organizations can use data science to identify areas where they can save time and resources.
- c. **Personalizes Experiences:** Data science helps create customized recommendations and offers that improve customer satisfaction.
- d. **Predicts the Future:** Businesses can use data to forecast trends, demand and other important factors.
- e. **Drives Innovation:** New ideas and products often come from insights discovered through data science.
- f. **Benefits Society:** Data science improves public services like healthcare, education and transportation by allocating resources more effectively.

Real life examples of Data Science include:

1. **E-Commerce:** Recommender systems suggest products by analyzing browsing history and user behavior patterns.
2. **Streaming Platforms:** Machine learning models recommend movies and videos based on watch duration and engagement metrics.
3. **Banking and Finance:** Fraud detection systems flag unusual transaction activity using anomaly-based classification.
4. **Healthcare Diagnostics:** Predictive models analyze medical scans to assist doctors in early disease detection and treatment decisions.
5. **Ride-Sharing:** Algorithms estimate fares and optimal routes by evaluating traffic conditions and supply-demand ratios.
6. **Social Media:** Sentiment analysis examines user posts to understand public opinion on brands, events and trends.

Data Science Tools and Programming Languages

A. Tools

a. SAS (Statistical Analysis System)

- Developed by the **SAS Institute**, it is a robust software suite used for advanced analytics, multivariate analysis, and business intelligence.
- It is a preferred choice for large-scale enterprises in regulated industries like banking and pharmaceuticals due to its stability and high security.
- **Advanced Statistical Analysis:** Offers powerful procedures (PROC) for everything from descriptive statistics to complex predictive modeling.
- **SAS Studio:** A browser-based development environment that allows users to access datasets and libraries without local client installation.

- **Enterprise-Grade Security:** Includes sophisticated data encryption and centralized metadata management for strict data governance.

b. **Microsoft Power BI**

- A unified business intelligence platform that allows users to connect to disparate data sources, transform raw data into interactive visualizations, and share insights across an organization.
- **Power Query:** A built-in data transformation engine used to clean, reshape, and prepare data for analysis.
- **DAX (Data Analysis Expressions):** A formula language used to create custom calculations and complex data models.
- **AI Integration:** Features like "Q&A" allow users to query data using natural language, while "Key Influencers" identify the main factors driving specific metrics.

c. **BigML**

- A cloud-based machine learning platform designed to democratize AI by providing a user-friendly interface for building and deploying models without deep programming knowledge.
- **1-Click Modeling:** Automates the creation of supervised and unsupervised models (like decision trees or clusters) with a single click.
- **Interpretable Visualizations:** Offers unique visual tools like Sunburst charts and Partial Dependence Plots to explain how models make decisions.
- **WhizzML:** A domain-specific scripting language for automating complex ML workflows and ensuring reproducibility.

d. **Tableau**

- A leading data visualization tool used to turn large datasets into interactive, shareable dashboards. It focuses on "visual storytelling," making it easy for stakeholders to understand complex trends.
- **VizQL Technology:** Translates drag-and-drop actions into data queries, allowing for real-time visual exploration.
- **Interactive Dashboards:** Users can filter, sort, and drill down into data live, supporting rapid ad-hoc analysis.
- **Extensive Data Connectors:** Seamlessly connects to live data in spreadsheets, SQL databases, and cloud platforms like Google BigQuery.

e. Orange

- An open-source data science and machine learning platform that uses **visual** programming to turn raw data into insights and predictive models. It makes complex data analysis accessible through a user-friendly, node-based interface that requires no coding.
- **Visual Workflow Programming (Node-based Engine):**
Users build analysis pipelines by dragging and dropping "widgets" onto a canvas and connecting them. This visual approach simplifies the machine learning process and makes the logic of your data flow easy to see and understand.
- **Interactive Data Visualization:**
Orange features "live" widgets where every chart is connected; selecting a data point in a scatter plot or a branch in a decision tree instantly updates all other connected widgets. This allows for rapid exploratory analysis and deep dives into specific data patterns.
- **Extensive Toolset and Add-ons:**
Beyond standard data preparation, Orange offers specialized toolsets for text mining, image analytics, and bioinformatics. It supports a wide range of data sources, including local spreadsheets, SQL databases, and built-in practice datasets.

B. Languages

a. Python

- Uses libraries like **Pandas, Requests, and BeautifulSoup** to pull data from CSVs, databases, APIs, and websites.
- Integrates easily with cloud data sources and big data tools (Spark, SQL).
- **Pandas and NumPy** handle missing values, filtering, merging, reshaping, and feature engineering.
- Libraries like **OpenRefine** (via APIs) and scikit-learn preprocessing help standardize and normalize data.
- Uses Matplotlib and Seaborn to visualize distributions, correlations, and trends.
- Pandas built-in plotting for quick insights.

a. R Language

- Uses **readr, DBI, and httr** to import data from files, databases, and APIs.
- Excellent support for statistical data formats.
- **dplyr, tidyr, and data.table** for fast data wrangling.
- Strong tools for data transformation and feature engineering.

- **ggplot2** for deep statistical visualization.
- Summary tools like **summary()**, **skimr**, and **corrplot**.

Data visualization using Orange Data Analytics

Data Collection:

- **Multiple Connectors:** Connects to local files (Excel, CSV, Tab), SQL databases, and cloud-based URLs.
- **Built-in Datasets:** Includes a library of classic datasets (like Titanic, Iris, and Employee Attrition) for instant practice and demonstration.
- **Live Data Loading:** The **File** widget monitors your source file and automatically updates the workflow when the data changes.

Data Cleaning:

- **Transformation Tools:** Widgets like **Select Columns**, **Transpose**, and **Merge Data** allow for easy manipulation of the dataset structure.
- **Preprocessing:** The **Preprocess** widget automates handling of missing values, data normalization, and feature scaling.
- **Data Filtering:** Users can use the **Data Sampler** or **Filter** widgets to focus on specific subsets of data.

Data Analysis:

- **Extensive Modeling:** Offers a full suite of algorithms including **Decision Trees**, **Linear Regression**, **Random Forests**, and **K-Means Clustering**.
- **Performance Evaluation:** The **Test and Score** widget provides instant metrics like **Accuracy (CA)**, **AUC**, and **F1 scores** to validate your models.
- **Statistical Insights:** Supports basic statistical summaries and correlation analysis through the **Data Table** and **Correlations** widgets.

Data Visualization:

- **Interactive Charts:** Features drag-and-drop widgets for **Distributions**, **Scatter Plots**, **Box Plots**, and **Heat Maps**.
- **Visual Data Exploration:** Selecting a group of points in one chart (e.g., a Scatter Plot) automatically highlights or filters those same points in every other connected widget.
- **Model Visualization:** Specialized viewers, such as the **Tree Viewer**, turn complex machine learning logic into easy-to-read flowcharts.

Decision Making:

- **Predictive Insights:** The **Predictions** widget allows users to input new, "unseen" data into a trained model to forecast outcomes (like which employees are likely to leave).
- **Explainable AI:** By visualizing the rules inside a **Decision Tree**, managers can understand exactly what factors (like Overtime or Salary) are driving the results.
- **Workflow Sharing:** Workflows can be saved and shared as visual files, allowing teams to replicate the exact decision-making logic without needing to read code.

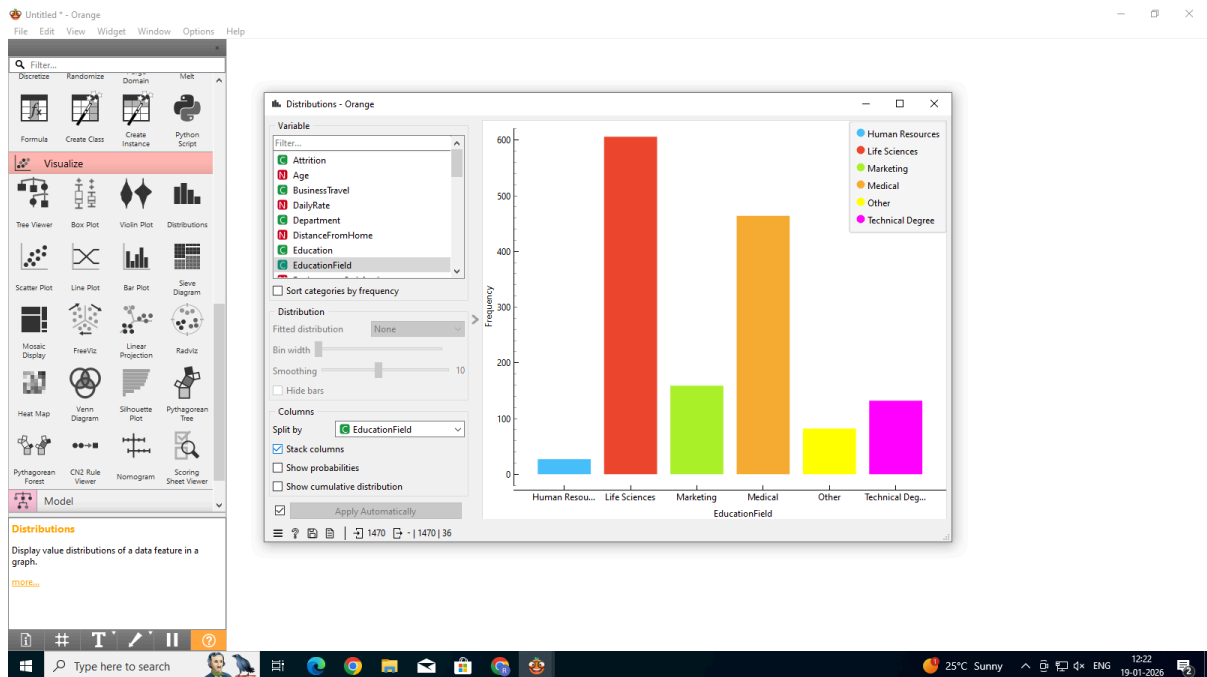


Fig 1: A Distributions Bar Chart showing the count of employees across different "Education Fields" within the organization.

Insights: This bar chart identifies "Life Sciences" and "Medical" as the most common education backgrounds among the staff. It helps HR understand the composition of their workforce before drilling down into which specific groups are quitting most often.

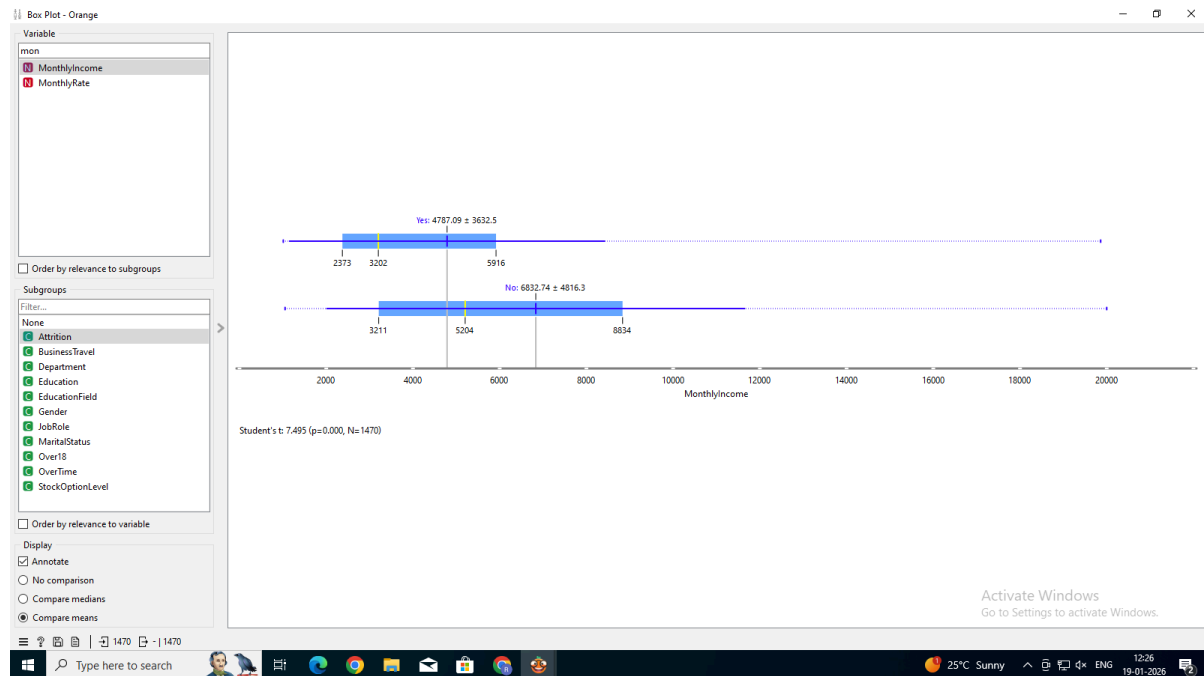


Fig 2: A Box Plot analysis comparing "Monthly Income" distributions between employees who stayed versus those who left.

Insights: The Box Plot clearly shows that employees who stay (No) have a higher median monthly income than those who leave (Yes). This visual gap provides a clear "Decision Making" insight: salary dissatisfaction is a primary driver of staff turnover.

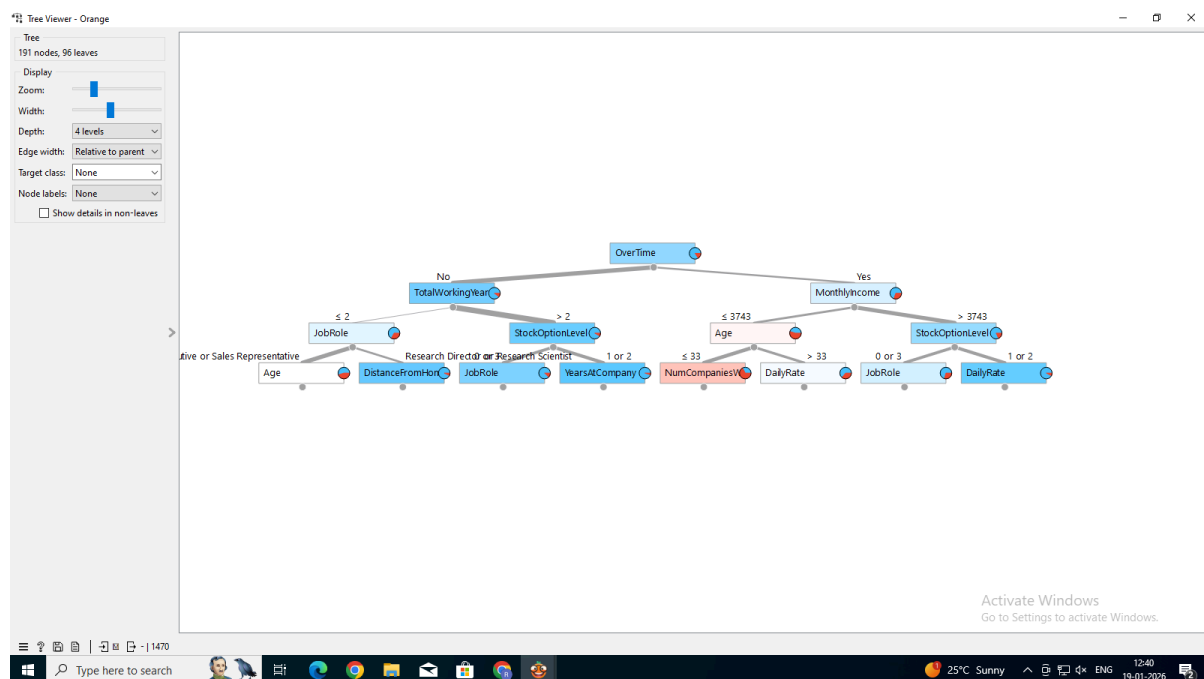


Fig 3: The Tree Viewer visualization, displaying the automated decision logic and "OverTime" as the primary root factor for attrition.

Insights: The tree identifies "OverTime" as the most critical factor, as it sits at the very top of the decision hierarchy. This visualization proves that management can reduce attrition by focusing specifically on workload and compensation rules shown in the lower branches.

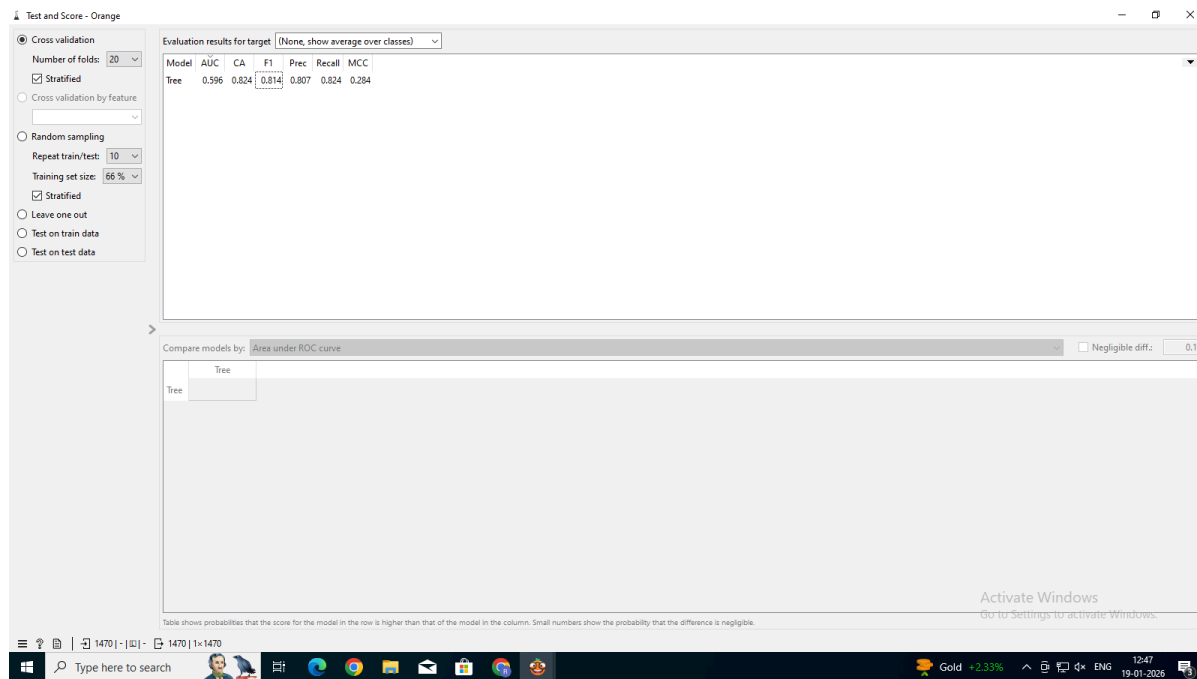


Fig 4: Model evaluation results in the Test and Score widget, showing a Classification Accuracy of 82.4% for the Tree model.

Insights: The model shows an 82.4% Classification Accuracy (CA), meaning it correctly predicts employee behavior in over eight out of ten cases. The F1 score of 0.814 confirms that the model is balanced and reliable enough for HR to use in real-world decision-making.

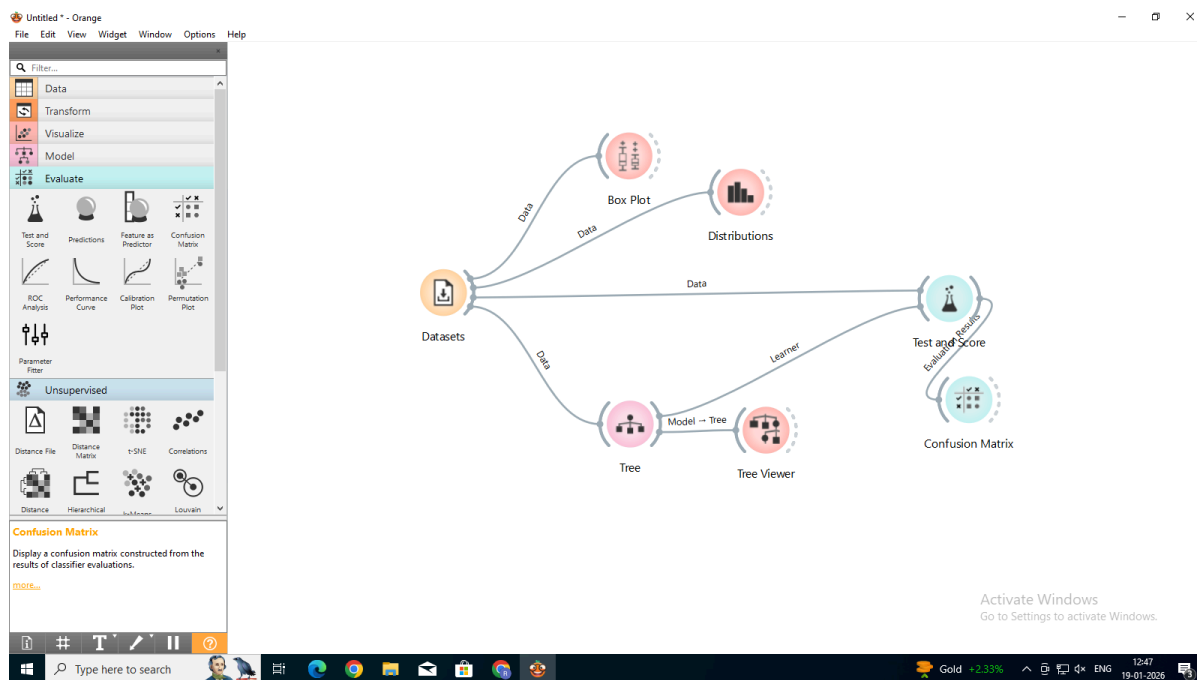


Fig 5: The finalized Predictive Workflow, featuring the correct connection between the Decision Tree learner and the Evaluation widgets.

Insights: By adding the Confusion Matrix and connecting the Learner directly to Test and Score, the pipeline is now fully functional. This represents a complete "End-to-End" data science process from raw data to model validation.

Conclusion :

The study of data science tools and programming languages demonstrates that the efficacy of modern data analysis depends on a well-integrated ecosystem of software and syntax. Through this experiment, it is evident that Python and R serve as the foundational pillars, offering the flexibility required for complex algorithm development and statistical rigor. However, these languages are most effective when paired with specialized libraries—such as Pandas for manipulation, Matplotlib for visualization, and Scikit-learn for machine learning—which streamline the transition from raw data to actionable insights.

Orange Data Mining is a powerful and intuitive open-source data science platform that supports the entire data analysis and machine learning lifecycle. With its versatile data connectivity, robust preprocessing widgets, advanced machine learning models, and interactive visual programming, it enables users to analyze complex information and make data-driven decisions through a visual workflow. Its no-code, drag-and-drop interface makes advanced data science and predictive analytics accessible to both beginners and experts, helping users transform raw data into visual insights and predictive models efficiently. It is a visual programming platform that converts complex datasets into interactive reports and decision-making logic, making data exploration simple and effective.

POST LAB QUESTION/ANSWER

Q1. What are the various applications of Data Science?

Real-world Applications of Data Science

1. In Search Engines

The most useful application of Data Science is Search Engines. As we know when we want to search for something on the internet, we mostly use Search engines like Google, Yahoo, DuckDuckGo and Bing, etc. So Data Science is used to get Searches faster.

For Example, When we search for something suppose "Data Structure and algorithm courses " then at that time on Internet Explorer we get the first link of GeeksforGeeks Courses. This happens because the GeeksforGeeks website is visited most in order to get information regarding Data Structure courses and Computer related subjects. So this analysis is done using Data Science, and we get the Topmost visited Web Links.

2. In Transport

Data Science is also entered in real-time such as the Transport field like Driverless Cars. With the help of Driverless Cars, it is easy to reduce the number of Accidents.

For Example, In Driverless Cars the training data is fed into the algorithm and with the help of Data Science techniques, the Data is analyzed like what as the speed limit in highways, Busy Streets, Narrow Roads, etc. And how to handle different situations while driving etc.

3. In Finance

Data Science plays a key role in Financial Industries. Financial Industries always have an issue of fraud and risk of losses. Thus, Financial Industries needs to automate risk of loss analysis in order to carry out strategic decisions for the company. Also, Financial Industries uses Data Science Analytics tools in order to predict the future. It allows the companies to predict customer lifetime value and their stock market moves.

For Example, In Stock Market, Data Science is the main part. In the Stock Market, Data Science is used to examine past behavior with past data and their goal is to examine the future outcome. Data is analyzed in such a way that it makes it possible to predict future stock prices over a set timetable.

4. In E-Commerce

E-Commerce Websites like Amazon, Flipkart, etc. uses data Science to make a better user experience with personalized recommendations.

For Example, When we search for something on the E-commerce websites we get suggestions similar to choices according to our past data and also we get recommendations according to most buy the product, most rated, most searched, etc. This is all done with the

help of Data Science.

5. In Health Care

In the Healthcare Industry data science acts as a boon. Data Science is used for:

- Detecting Tumor.
- Drug discoveries.
- Medical Image Analysis.
- Virtual Medical Bots.
- Genetics and Genomics.
- Predictive Modeling for Diagnosis etc.

6. Image Recognition

Currently, Data Science is also used in Image Recognition. For Example, When we upload our image with our friend on Facebook, Facebook gives suggestions Tagging who is in the picture. This is done with the help of machine learning and Data Science. When an Image is Recognized, the data analysis is done on one's Facebook friends and after analysis, if the faces which are present in the picture matched with someone else profile then Facebook suggests us auto-tagging.

7. Targeting Recommendation

Targeting Recommendation is the most important application of Data Science. Whatever the user searches on the Internet, he/she will see numerous posts everywhere. This can be explained properly with an example: Suppose I want a mobile phone, so I just Google search it and after that, I changed my mind to buy it offline. In Real -World Data Science helps those companies who are paying for Advertisements for their mobile. So everywhere on the internet, in social media, in the websites, in the apps everywhere I will see the recommendation of that mobile phone which I searched for. So this will force me to buy online.

8. Airline Routing Planning

With the help of Data Science, Airline Sector is also growing like with the help of it, it becomes easy to predict flight delays. It also helps to decide whether to directly land into the destination or take a halt in between like a flight can have a direct route from Delhi to the U.S.A or it can halt in between after that reach at the destination.

Q2. Write a short note on the following: Hadoop, GO, GIT.

1) Hadoop

Apache Hadoop (Distributed Systems)

Overview: Apache Hadoop is an open-source framework developed in Java, engineered to facilitate the distributed processing of massive datasets. It operates across clusters of commodity hardware using a parallel computing model.

Functionality: Hadoop addresses the challenges of "Big Data" (Volume, Velocity, and Variety) by partitioning data into smaller blocks and distributing them across a cluster. The core architecture relies on:

HDFS (Hadoop Distributed File System): A fault-tolerant storage system.

MapReduce: A programming model for processing data in parallel.

Industrial Applications:

Data Warehousing: Aggregating large-scale unstructured data for business intelligence.

Log Analysis: Processing server logs to derive user behavioral patterns.

Predictive Modeling: Enhancing fraud detection algorithms in financial sectors.

2) Go Programming Language (Golang)

Overview: Go is a statically typed, compiled language developed by Google. It was designed to bridge the gap between the performance of low-level languages like C++ and the productivity of high-level languages like Python.

Functionality: Go is optimized for backend systems and cloud-native infrastructure. Its standout feature is Concurrency, managed through "Goroutines"—lightweight execution threads that allow the system to handle thousands of simultaneous tasks with minimal memory overhead.

Industrial Applications:

Cloud Infrastructure: Powering containerization tools like Docker and Kubernetes.

Microservices: Implementing highly scalable, independent services for high-traffic platforms (e.g., Uber, Netflix).

Network Programming: Building high-performance web servers and APIs.

3) Git (Version Control System)

Overview: Git is a Distributed Version Control System (VCS) used to track chronological changes in source code during software development.

Functionality: Unlike centralized systems, Git allows every developer to have a full local copy of the project history. Key features include:

Branching/Merging: Enabling parallel development streams.

Reversion: Providing the ability to roll back to any previous state of the project in case of errors.

Industrial Applications:

Collaborative Development: Facilitating teamwork and code reviews via platforms like GitHub and GitLab.

DevOps Pipelines: Serving as the foundation for Continuous Integration and Continuous Deployment (CI/CD) workflows.