Practical No.: 09

Theory:

What Are Open-Source Data Mining Tools?

1. Open-source data mining tools—tools used to extract hidden or unknown information from large datasets—are free to use, can be tailored to individual requirements, and can be redistributed without any constraints.
2. Open-source means that the core functionality of the software, and even the code itself, can be altered; Orgaizations looking to harness advanced analytics without incurring high costs can use these tools to meet data scientists' needs.
3. Data mining tools facilitate the collection of new data points from publicly available resources.
4. This can be done from a variety of sources and through a wide range of techniques, including advanced computational methods that can identify data on a web page or collected through a piece of software or hardware.
5. Insights extracted through data mining can be a potent asset for decision-making, forecasting trends, or making accurate predictions.
6. Its applicability is broad and spans such areas as business intelligence, scientific research, and predictive modeling, making these open source tools invaluable across many sectors.
7. Generally, open source software offers a more transparent and collaborative approach to software development. Defined by its freely available source code, it provides a platform for anyone to inspect, adapt, and share..
8. Similarly, open source data mining tools benefit from this collective effort—with thousands of users and developers worldwide, they're enhanced by the addition of innovative features, bug fixes, and other modifications made available to the broader community.

❖ Top Open Source Data Mining Tools:

A] **WEKA:**

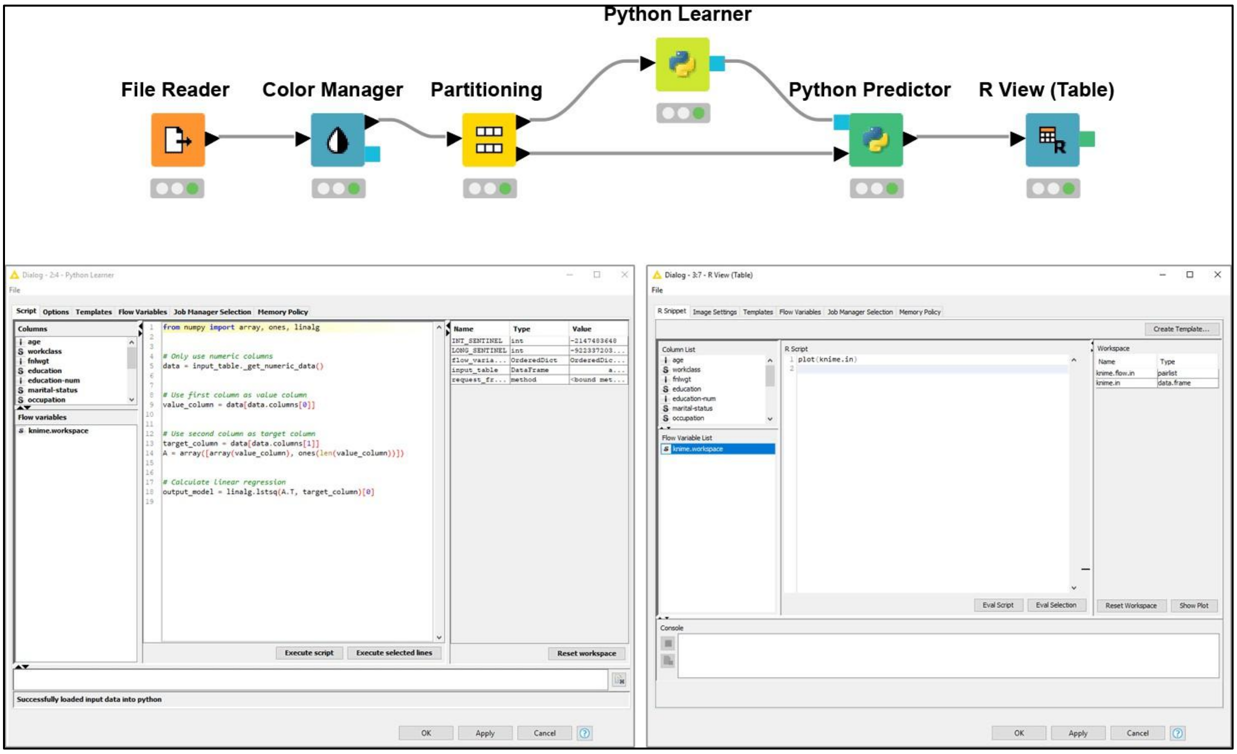
Waikato Environment for Knowledge Analysis
Version 3.9.6
(c) 1999 - 2022
The University of Waikato
Hamilton, New Zealand

1. WEKA is a prominent open source data mining tool created by the University of Waikato in New Zealand.
2. At its core, it's a comprehensive collection of machine learning algorithms tailored for various data

mining tasks.

3. The software, licensed under the GNU General Public License, is designed to help users analyze large datasets and transform them into actionable insights.
4. Key features include the following:
    1. **User-friendly interface**—Offers an intuitive graphical user interface (GUI), simplifying the process of data visualization and analysis for users.
    2. **Comprehensive algorithm suite**—Encompasses a wide range of machine learning algorithms, facilitating tasks like classification, regression, clustering, and association rules mining.
    3. **Data preprocessing tools**—Provides robust capabilities for data transformation, attribute selection, and handling missing values.
    4. **Java-based architecture**—Java-based WEKA is platform-independent and easily integrates with other systems or applications.
    5. **Visualization capabilities**—Contains robust tools for data visualization, such as scatter plots and histograms, assisting users in better understanding their datasets.

B] KNIME:



```python
from numpy import array, ones, linalg


# Only use numeric columns
data = input_table._get_numeric_data()


# Use first column as value column
value_column = data[data.columns[0]]


# Use second column as target column
target_column = data[data.columns[1]]
A = array([array(value_column), ones(len(value_column))])


# Calculate linear regression
output_model = linalg.lstsq(A.T, target_column)[0]
```

Name	Type	Value
INT_SENTINEL	int	-2147483648
LONG_SENTINEL	int	-922337203...
flow_varia...	OrderedDict	OrderedDic...
input_table	DataFrame	a...
request_fr...	method	<bound met...

Successfully loaded input data into python

```r
plot(knime.in)
```

Name	Type
knime.flow.in	pairlist
knime.in	data.frame

1. KNIME is a leading open source platform for data analytics, reporting, integration, and mining.
2. Emerging from the University of Konstanz, it provides users with a visual interface to design data workflows, allowing for a seamless blend of data access, data transformation, model training, and visualization.
3. Key features include the following:
    1. **Drag-and-drop interface**—Visual workflow editor facilitates easy drag-and- drop operations, enabling users to construct sophisticated data workflows without requiring coding.

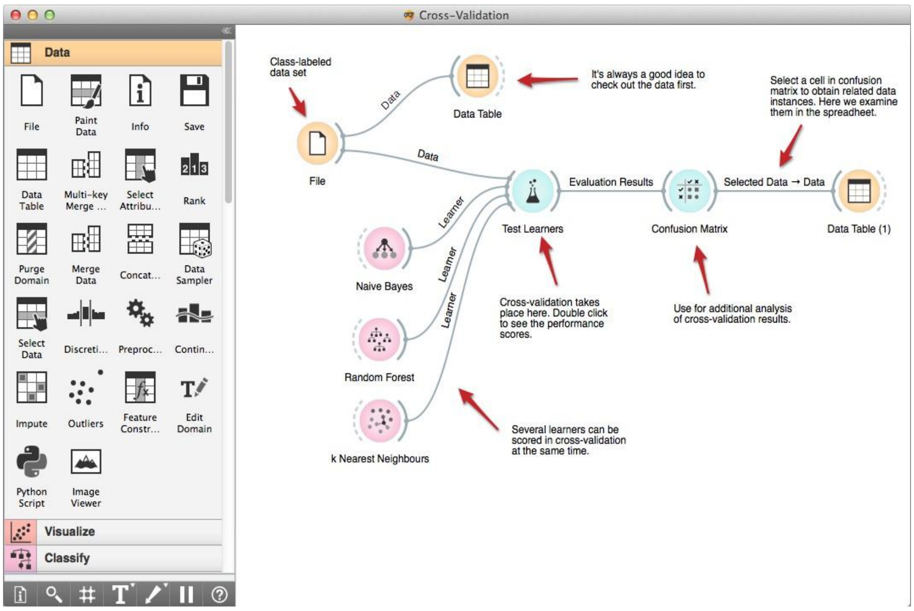
ii. **Modular data pipelining**—Employs a node-based system where each node performs a specific task, ensuring modular and reproducible data workflows.

iii. **Extensibility**—Supports thousands of plugins, extending functionalities to various domains such as text processing, image analysis, and machine learning.

iv. **Integrated analytics**—Rich array of built-in algorithms and tools for data mining and machine learning cater to both classic statistical models and cutting-edge artificial intelligence (AI) techniques.

v. **Open platform**—While KNIME offers a free open source version, it also provides an enterprise version with advanced features, ensuring it suits both individual users and large corporations.

C] Orange:



Class-labeled data set

It's always a good idea to check out the data first.

Select a cell in confusion matrix to obtain related data instances. Here we examine them in the spreadheet.

Cross-validation takes place here. Double click to see the performance scores.

Use for additional analysis of cross-validation results.

Several learners can be scored in cross-validation at the same time.

Orange is a powerful, open source data visualization and analysis tool tailored for novice and expert data miners alike.

1. Hailing from the University of Ljubljana in Slovenia, it brings forth a component- based approach to data analytics, making the exploration of quantitative and qualitative data both interactive and fun.
2. Key features include the following:

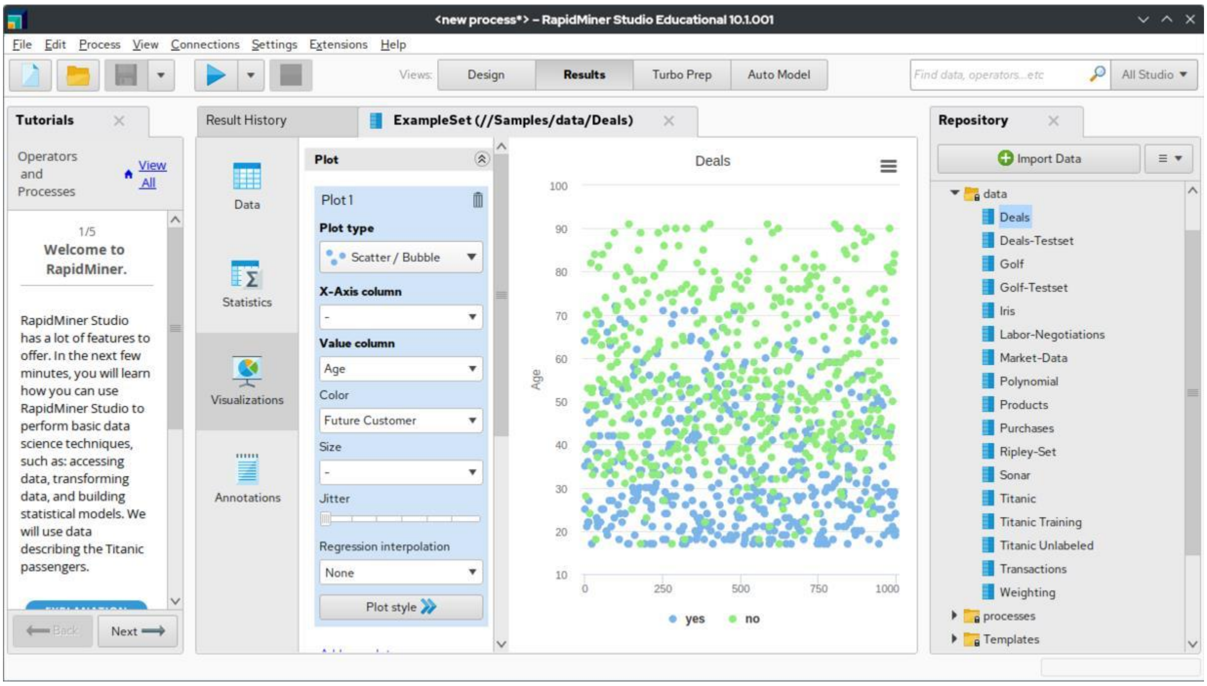
i. **Visual programming**—User-friendly interface enables users to design data workflows visually; users can establish a data analysis pipeline without writing a single line of code by simply dragging and dropping widgets.

ii. **Widgets system**—Operates on a system of widgets, with each widget performing a specific function from data input and preprocessing to visualization and predictive modeling.

iii. **Interactive visualizations**—Offers a diverse set of visualization tools such as scatter plots, box plots, and tree viewers, facilitating deep exploratory data analysis.

iv. **Extensible framework**—Core functionalities can be extended through add-ons, catering to specialized data analysis tasks like bioinformatics, text mining, and more.

D] RapidMiner:



1/5

Welcome to RapidMiner.

RapidMiner Studio has a lot of features to offer. In the next few minutes, you will learn how you can use RapidMiner Studio to perform basic data science techniques, such as: accessing data, transforming data, and building statistical models. We will use data describing the Titanic passengers.

Plot

Plot 1

Plot type: Scatter / Bubble

X-Axis column: -

Value column: Age

Color: Future Customer

Size: -

Jitter

Regression interpolation: None

1. RapidMiner is a highly acclaimed data science platform that seamlessly integrates data preparation, machine learning, and model deployment into a single cohesive environment.
2. Originating in the research community of the Technical University of Dortmund, Germany, it has since burgeoned into one of the leading data mining tools favored by businesses, analysts, and researchers worldwide.
3. Key features include the following:
   i. **Unified environment**—All-in-one platform simplifies the data science process by consolidating data access, data preparation, machine learning, and model deployment.
   ii. **Visual workflow designer**—Intuitive drag-and-drop interface enables users to design complex data workflows visually, ensuring clarity and efficiency even for those with minimal coding experience.
   iii. **Extensive algorithms library**—Comprehensive library of prebuilt machine learning algorithms and models caters to a multitude of data analytics tasks, from regression and clustering to advanced predictive modeling.
   iv. **Scalability and integration**—Designed for both small-scale projects and large- scale industrial applications, can be easily integrated with other tools and databases and offers cloud solutions to ensure scalability.
   v. **Collaborative data science**—Collaborative features let teams share data, models, and results, enabling synchronized work across different sectors of an organization and facilitating decision-making.

E] Apache Mahout:

All Applications

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved	Active Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes
0	0	0	0	0	0 B	8 GB	0 B	0	8	0	1	0	0	0	0

Show 20 entries Search:

ID	User	Name	Application Type	Queue	StartTime	FinishTime	State	FinalStatus	Progress	Tracking UI
No data available in table										

Showing 0 to 0 of 0 entries

1. Apache Mahout is an open source project focused on producing scalable machine- learning algorithms to be used in data mining.
2. Rooted in the Apache Software Foundation, Mahout primarily operates in the Hadoop ecosystem, utilizing the MapReduce paradigm to effectively process large datasets.
3. It’s capable of handling extensive data mining challenges, aiding businesses and researchers in extracting meaningful insights from vast data reservoirs.
4. Key features include the following:
   i. **Scalable machine learning**—Adept at handling gigantic datasets thanks to its tight integration with Hadoop and ability to run atop distributed storage and processing environments.
   ii. **Diverse algorithms library**—Rich repository of machine learning algorithms span various domains like clustering, classification, and collaborative filtering, catering to a wide spectrum of data analytics needs.
   iii. **Linear algebra framework**—Incorporates a specialized linear algebra framework, known as Samsara, which acts as a foundation for many of its machine learning algorithms, ensuring mathematical accuracy and computational efficiency.
   iv. **Modular and extensible**—Inherently modular architecture lets users effortlessly incorporate new algorithms or extend existing ones, tailoring the tool to their specific requirements.
   v. **Native support for Spark**—While Mahout originally leveraged MapReduce, it has evolved to natively support other distributed backends, notably Apache Spark, allowing for faster processing speeds and broader application.

Conclusion:

In conclusion, the study of data mining with open-source tools has provided valuable insights into the potential of these tools in real-world scenarios. The versatility,accessibility, and community support make them an excellent choice for researchers, analysts, and practitioners alike.