Data Mining and Business Intelligence Tools

1.1 Data Mining Introduction:

Data mining is the term related to discovering various patterns among huge amount of data and transforming the acquired data into more useful form. For this purpose, various data mining tools, algorithms, statistical analysis and also advanced techniques like artificial intelligence.

Here is a list of some of the most useful and popular data mining tools.

- Rapid Miner
- Orange
- Weka
- Knime
- Sisense
- Sql Server Data Tools
- Apache Mahout

1. Rapid Miner

Rapid Miner is one of the best predictive analysis system developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis.

The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.

Rapid Miner offers the server as both on premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template based frameworks that enable speedy delivery with reduced number of errors (which are quite commonly expected in manual code writing process).

Rapid Miner constitutes of three modules, namely

1. Rapid Miner Studio- This module is for workflow design, prototyping, validation etc.
2. Rapid Miner Server- To operate predictive data models created in studio
3. Rapid Miner Radoop- Executes processes directly in Hadoop cluster to simplify predictive analysis.

Pros :-

- Flow based programming allows visualization of pipelines
- Contains modules for statistical analysis, machine learning, etc
- No coding required
- Easy to setup

Cons :-

- No coding required-Challenging to use for coders. Although it does contain Java/Python modules you must use flow programming interface.
- Commercial-Expensive licenses need to be purchased.
- Unintuitive-Its very easy to get lost in the sea of modules
- Limited-Its use case is limited to the set of processors/modules it contains

2. Orange

Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component based software. It has been written in Python computing language.

As it is component-based software, the components of orange are called ‘widgets'. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modeling.

Widgets offer major functionalities like

- Showing data table and allowing to select features
- Reading the data
- Training predictors and to compare learning algorithms
- Visualizing data elements etc.

Pros :-

Open Source

- Interactive Data Visualization
- Visual Programming
- Supports Hands-on Training and Visual Illustrations
- Add-ons Extend Functionality

3. Weka

Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning.
Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language.

Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file.

Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

4. Knime

KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together.

KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence.

KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time and it has made predictive analysis

accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

5. Sisense

Sisense is extremely useful and best suited BI software when it comes to reporting purposes within the organization. It is developed by the company of same name 'Sisense'. It has a brilliant capability to handle and process data for the small scale/large scale organizations.

It allows combining data from various sources to build a common repository and further, refines data to generate rich reports that get shared across departments for reporting.

Sisense got awarded as best BI software is 2016 and still holds a good position.
Sisense generates reports which are highly visual. It is specially designed for users that are non-technical. It allows drag & drop facility as well as widgets.

Different widgets can be selected to generate the reports in form of pie charts, line charts, bar graphs etc. based on the purpose of an organization. Reports can be further drilled down by simply clicking to check details and comprehensive data.

6. SSDT (Sql Server Data Tools)

SSDT is a universal, declarative model that expands all the phases of database development in the Visual Studio IDE. BIDS was the former environment developed by Microsoft to do data analysis and provide business intelligence solutions. Developers use SSDT transact- a design capability of SQL, to build, maintain, debug and refactor databases.

A user can work directly with a database or can work directly with a connected database, thus, providing on or off-premise facility.

Users can use visual studio tools for development of databases like IntelliSense, code navigation tools, and programming support via C#, visual basic etc. SSDT provides Table Designer to create new tables as well as edit tables in direct databases as well as connected databases.

7. Apache Mahout

Apache Mahout is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering.
Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates.

To key up, Mahout has following major features

- Extensible programming environment
- Pre-made algorithms
- Math experimentation environment
- GPU computes for performance improvement.

8. DataMelt

DataMelt, also known as DMelt is a computation and visualization environment that provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students.

DMelt is written in JAVA and it is a multi-platform utility. It can run on any operating system which is compatible with JVM(Java Virtual Machine).

It contains Scientific & mathematical libraries.

Scientific libraries: To draw 2D/3D plots.
Mathematical libraries: To generate random numbers, curve fitting, algorithms etc.
DataMelt can be used for analysis of large data volumes, data mining, and stat analysis. It is widely used in the analysis of financial markets, natural sciences & engineering.

9. IBM Cognos

BM Cognos BI is an intelligence suite owned by IBM for reporting and data analysis, score carding etc. It consists of sub-components that meet specific organizational requirements Cognos Connection, Query Studio, Report Studio, Analysis Studio, Event studio & Workspace Advance.

Cognos Connection: A web portal to gather and summarize data in scoreboard/reports.
Query Studio: Contains queries to format data & create diagrams.
Report Studio: To generate management reports.
Analysis Studio: To process large data volumes, understand & identify trends.
Event Studio: Notification module to keep in sync with events.
Workspace Advanced: User-friendly interface to create personalized & user-friendly documents.

10.XLMiner

XLMiner is a comprehensive data mining add-in for Excel. Data mining is a discovery-driven data analysis technology used for identifying patterns and relationships in data sets. With overwhelming amounts of data now available from transaction systems and external data sources, organizations are presented with increasing opportunities to understand their data and gain insights into it. Data mining is still an emerging field, and is a convergence of fields like statistics, machine learning, and artificial intelligence.

XLMiner is a tool belt offering a variety of methods to analyze data. It has extensive coverage of statistical and machine-learning techniques for classification, prediction, affinity analysis, data exploration, and reduction.