Name - Jayant Goel

Enrollment number - 18103255

Batch - B8

Machine Learning & Big Data
Tutorial - 2

Ans1:

Supervised learning is the type of machine learning in which machines are trained using well "labelled" training data, and on the basis of that data, machines predict the output. The labelled data means some input data is already tagged with the correct output.

In supervised learning, the training data provided to the machines work as the supervisor that teaches the machines to predict the output correctly. It applies the same concept as a student learns in the supervision of the teacher.

Algorithms:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

Ans2:

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things.

Algorithms:

- K-means for clustering problems.
- Apriori algorithm for association rule learning problems.

Ans3:

1) Xplenty - It is a complete toolkit for building data pipelines with low-code and no-code capabilities. It has solutions for marketing, sales, support, and developers. Xplenty will help you make the most out of your data without investing in hardware, software, or related personnel. Xplenty provides support through email, chats, phone, and an online meeting.
2) Adverity - It is a flexible end-to-end marketing analytics platform that enables marketers to track marketing performance in a single view and effortlessly uncover new insights in real-time. Thanks to automated data integration from over 600 sources, powerful data visualizations, and AI-powered predictive analytics, Adverity enables marketers to track marketing performance in a single view and effortlessly uncovers new insights in real-time
3) Apache Hadoop - It is a software framework employed for clustered file system and handling of big data. It processes datasets of big data by means of the MapReduce programming model. The core strength of Hadoop is its HDFS (Hadoop Distributed File System) which has the ability to hold all type of data – video, images, JSON, XML, and plain text over the same file system.
4) CDH - It aims at enterprise-class deployments of that technology. It is totally open source and has a free platform distribution that encompasses Apache Hadoop, Apache Spark, Apache Impala, and many more.
5) Apache Cassandra - It is free of cost and open-source distributed NoSQL DBMS constructed to manage huge volumes of data spread across numerous commodity servers, delivering high availability. It employs CQL (Cassandra Structure Language) to interact with the database. No single point of failure. Handles massive data very quickly. Log-structured storage
6) KNIME - It stands for Konstanz Information Miner which is an open source tool that is used for Enterprise reporting, integration, research, CRM, data mining, data analytics, text mining, and business intelligence. It supports Linux, OS X, and Windows operating systems. It can be considered as a good alternative to SAS. Some of the top companies using Knime include Comcast, Johnson & Johnson, Canadian Tire, etc.

7) Datawrapper - It is an open source platform for data visualization that aids its users to generate simple, precise and embeddable charts very quickly. Brings all the charts in one place. Great customization and export options. Requires zero coding.

8) MongoDB - It is a NoSQL, document-oriented database written in C, C++, and JavaScript. It is free to use and is an open source tool that supports multiple operating systems including Windows Vista (and later versions), OS X (10.7 and later versions), Linux, Solaris, and FreeBSD. Its main features include Aggregation, Adhoc-queries, Uses BSON format, Sharding, Indexing, Replication, Server-side execution of javascript, Schemaless, Capped collection, MongoDB management service (MMS), load balancing and file storage.

9) Lumify - It is a free and open source tool for big data fusion/integration, analytics, and visualization. Its primary features include full-text search, 2D and 3D graph visualizations, automatic layouts, link analysis between graph entities, integration with mapping systems, geospatial analysis, multimedia analysis, real-time collaboration through a set of projects or workspaces.

10) HPCC - It stands for High-Performance Computing Cluster. This is a complete big data solution over a highly scalable supercomputing platform. HPCC is also referred to as DAS (Data Analytics Supercomputer). This tool was developed by LexisNexis Risk Solutions. This tool is written in C++ and a data-centric programming language known as ECL(Enterprise Control Language). It is based on a Thor architecture that supports data parallelism, pipeline parallelism, and system parallelism. It is an open-source tool and is a good substitute for Hadoop and some other Big data platforms.

Ans4:

Three characteristics define Big Data: volume, variety, and velocity.

1) The Volume of Data - The sheer volume of data being stored today is exploding. In the year 2000, 800,000 petabytes (PB) of data were stored in the world. Of course, a lot of the data that's being created today isn't analyzed at all and that's another problem that needs to be considered. This number is expected to reach 35 zettabytes (ZB) by 2020. Twitter alone generates more than 7 terabytes (TB) of data every day, Facebook 10 TB, and some enterprises generate terabytes of data every hour of every day of the year. It's no longer unheard of for individual enterprises to have storage clusters holding petabytes of data.
2) The Variety of Data - With the explosion of sensors, and smart devices, as well as social collaboration technologies, data in an enterprise has become complex, because it includes not only traditional relational data, but also raw, semi-structured, and unstructured data from web pages, weblog files (including click-stream data), search indexes, social media forums, e-mail, documents, sensor data from active and passive systems, and so on.
3) The Velocity of Data - Just as the sheer volume and variety of data we collect and the store has changed, so, too, has the velocity at which it is generated and needs to be handled. A conventional understanding of velocity typically considers how quickly the data is arriving and stored, and its associated rates of retrieval. Managing all of that quickly is good—and the volumes of data that we are looking at are a consequence of how quickly the data arrives.