How Machine Learning is applied in data science?

- Machine learning analyzes and examines large chunks of data automatically.
- It automates the data analysis process and makes predictions in real-time without any human involvement.
- You can further build and train the data model to make real-time predictions. This point is where you use machine learning algorithms in the data science lifecycle.

The significance of data preprocessing in data science?

Data Preprocessing can be defined as a process of converting raw data into a format that is understandable and usable for further analysis. It is an important step in the Data Preparation stage. It ensures that the outcome of the analysis is accurate, complete, and consistent.

Why is data preprocessing essential?

Data Preprocessing is an important step in the Data Preparation stage of a Data Science development lifecycle that will ensure reliable, robust, and consistent results. The main objective of this step is to ensure and check the quality of data before applying any Machine Learning or Data Mining methods. Let's review some of its benefits -

- Accuracy - Data Preprocessing will ensure that input data is accurate and reliable by ensuring there are no manual entry errors, no duplicates, etc.
- Completeness - It ensures that missing values are handled, and data is complete for further analysis.
- Consistent - Data Preprocessing ensures that input data is consistent, i.e., the same data kept in different places should match.
- Timeliness - Whether data is updated regularly and on a timely basis or not.
- Trustable - Whether data is coming from trustworthy sources or not.
- Interpretability - Raw data is generally unusable, and Data Preprocessing converts raw data into an interpretable format.

Provide examples of common techniques used in preparing data for analysis?

- Quantitative Methods
    1. Cluster analysis - The action of grouping a set of data elements in a way that said elements are more similar (in a particular sense) to each other than to those in other groups – hence the term 'cluster.' Since there is no target variable when clustering, the method is often used to find hidden patterns in the data.
    2. Cohort analysis - This type of data analysis approach uses historical data to examine and compare a determined segment of users' behavior, which can then be grouped with others with similar characteristics.

3. Regression analysis - Regression uses historical data to understand how a dependent variable's value is affected when one (linear regression) or more independent variables (multiple regression) change or stay the same.
4. Neural networks - The neural network forms the basis for the intelligent algorithms of machine learning. It is a form of analytics that attempts, with minimal intervention, to understand how the human brain would generate insights and predict values.
5. Factor analysis - The factor analysis also called “dimension reduction” is a type of data analysis used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors.
6. Data mining - A method of data analysis that is the umbrella term for engineering metrics and insights for additional value, direction, and context. By using exploratory statistical evaluation, data mining aims to identify dependencies, relations, patterns, and trends to generate advanced knowledge.
7. Time series analysis - As its name suggests, time series analysis is used to analyze a set of data points collected over a specified period of time.
8. Decision Trees - The decision tree analysis aims to act as a support tool to make smart and strategic decisions. By visually displaying potential outcomes, consequences, and costs in a tree-like model, researchers and company users can easily evaluate all factors involved and choose the best course of action.
9. Conjoint analysis - This approach is usually used in surveys to understand how individuals value different attributes of a product or service and it is one of the most effective methods to extract consumer preferences.
10. Correspondence Analysis - Also known as reciprocal averaging, correspondence analysis is a method used to analyze the relationship between categorical variables presented within a contingency table.
11. Multidimensional Scaling (MDS) - MDS is a method used to observe the similarities or disparities between objects which can be colors, brands, people, geographical coordinates, and more.

- Qualitative Methods
    1. Text analysis - Text analysis, also known in the industry as text mining, works by taking large sets of textual data and arranging them in a way that makes it easier to manage.
    2. Content Analysis - This is a straightforward and very popular method that examines the presence and frequency of certain words, concepts, and subjects in different content formats such as text, image, audio, or video.
    3. Thematic Analysis - Very similar to content analysis, thematic analysis also helps in identifying and interpreting patterns in qualitative data with the main difference being that the first one can also be applied to quantitative analysis.
    4. Narrative Analysis - narrative analysis is used to explore the meaning behind the stories that people tell and most importantly, how they tell them. By looking into the

words that people use to describe a situation you can extract valuable conclusions about their perspective on a specific topic.

5. Discourse Analysis - Discourse analysis is used to understand the meaning behind any type of written, verbal, or symbolic discourse based on its political, social, or cultural context. It mixes the analysis of languages and situations together.
6. Grounded Theory Analysis - Grounded theory is the only method that doesn't require an initial research question or hypothesis as its value lies in the generation of new theories. With the grounded theory method, you can go into the analysis process with an open mind and explore the data to generate new theories through tests and revisions.

What is exploratory data analysis (EDA)?

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

The role of exploratory data analysis (EDA) in data science?

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

What is exploratory data analysis (EDA) objectives?

The goal of EDA is to allow data scientists to get deep insight into a data set and at the same time provide specific outcomes that a data scientist would want to extract from the data set. It includes:

- List of outliers
- Estimates for parameters
- Uncertainties about those estimates
- List of all important factors
- Conclusions or assumptions as to whether certain individual factors are statistically essential.
- Optimal settings

- A good predictive model

Some common techniques used in exploratory data analysis (EDA) ?

EDA techniques may include calculating summary statistics, visualizing data distributions, identifying outliers, exploring relationships between variables, and performing hypothesis testing.

Can you provide an overview of what big data is?

Big Data is a collection of data that is huge in volume, yet growing exponentially with time. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data is also a data but with huge size.

Following are the types of Big Data:

- Structured - Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- Unstructured - Any data with unknown form or the structure is classified as unstructured data.
- Semi-structured - Semi-structured data can contain both the forms of data.

What're the characteristics of big data?

- Volume - The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data.
- Variety - Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications.
- Velocity - Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors, Mobile devices, etc. The flow of data is massive and continuous.
- Variability - This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

What're the challenges and opportunities it presents in the field of data science?

Challenges:

1. Identifying the data problem - One of the toughest challenges of data science is identifying the problem or the issue. Data scientists mostly start off with a huge data set that is often unstructured. They have to understand what they have to do with this data.
2. Finding the most appropriate data - As companies produce huge amounts of data every second, it is a daunting task to get your hands on the right data for analysis. This is because the correct data set will be crucial for developing the most appropriate data model.

3. Lack of skilled workforce - As more and more companies are becoming dependent on data science, the demand for skilled data professionals is increasing.
4. Data cleansing - Data cleansing or removing unwanted data from a data set is one of the pressing challenges of data science

Opportunities:

1. Data Scientist

Average Salary: $123,300

Typical Job Requirements: Find, clean, and organize data for companies. Data scientists need to be able to analyze large amounts of complex raw and processed information to find patterns that benefit an organization and help drive its strategic business decisions. Compared to data analysts, data scientists are much more technical.

2. Machine Learning Engineer

Typical Job Requirements: Machine learning engineers create data funnels and deliver software solutions. They typically need strong statistics and programming skills, as well as a knowledge of software engineering. In addition to designing and building machine learning systems, they're also responsible for running tests and experiments to monitor the performance and functionality of such systems.

3. Machine Learning Scientist

Average Salary: $135,000

Typical Job Requirements: Research new data approaches and algorithms to be used in adaptive systems including supervised, unsupervised, and deep learning techniques. Machine learning scientists often go by titles like Research Scientist or Research Engineer.

4. Applications Architect

Average Salary: $146,200

Typical Job Requirements: Track the behavior of applications used within a business and how they interact with each other and with users. Applications architects are focused on designing the architecture of applications as well, including building components like user interface and infrastructure.

5. Enterprise Architect

Average Salary: $150,400

Typical Job Requirements: An enterprise architect is responsible for aligning an organization's strategy with the technology needed to execute its objectives. To do so, they must have a complete understanding of the business and its technology needs in order to design the systems architecture required to meet them.

6. Data Architect

Average Salary: $137,000

Typical Job Requirements: Ensure data solutions are built for performance and design analytics applications for multiple platforms. In addition to creating new database systems, data architects often find ways to improve the performance and functionality of existing systems, as well as working to provide access to database administrators and analysts.

7. Infrastructure Architect

Average Salary: $145,200

Typical Job Requirements: Oversee that all business systems are working optimally and can support the development of new technologies and system requirements. A similar job title is Cloud Infrastructure Architect, which oversees a company's cloud computing strategy.

8. Data Engineer

Average Salary: $129,900

Typical Job Requirements: Perform batch processing or real-time processing on gathered and stored data. Data engineers are also responsible for building and maintaining data pipelines that create a robust and interconnected data ecosystem within an organization, making information accessible for data scientists.

9. Business Intelligence (BI) Developer

Average Salary: $105,300

Typical Job Requirements: BI developers design and develop strategies to assist business users in quickly finding the information they need to make better business decisions. Extremely data-savvy, they use BI tools or develop custom BI analytic applications to facilitate the end-users' understanding of their systems.

10. Statistician

Average Salary: $87,300

Typical Job Requirements: Statisticians collect, analyze, and interpret data in order to identify trends and relationships which can be used to inform organizational decision-

making. The daily responsibilities of statisticians often include designing data collection processes, communicating findings to stakeholders, and advising organizational strategy.

11. Data Analyst

Average Salary: $80,800

Typical Job Requirements: Transform and manipulate large data sets to suit the desired analysis for companies. For many companies, this role can also include tracking web analytics and analyzing A/B testing. Data analysts also aid in the decision-making process by preparing reports for organizational leaders which effectively communicate trends and insights gleaned from their analysis.