DATA SCIENCE

What do you understand about data science?

Data science is a field that deals with analyzing, manipulating and visualizing data using different methods such as Python,R , Sql, Excel, Tableau etc…

Data science helps us make informed decisions based on the data that is worked on. It enables us to find solutions to our everyday problems and make better decisions.

Data science finds applications across multiple domains and industries, such as healthcare, finance, marketing, transportation, and more. It plays a crucial role in aiding decision-making processes, optimizing operations, improving customer experiences, and driving innovation.

Subfields of Data Science?

1. Business Analytics: Business analytics applied data science techniques to solve business problems and drive data-informed decision-making. It involves analyzing business data to gain insights into performance, customer behavior, market trends, and optimizing business strategies.
2. Predictive Analytics: Predictive analytics focuses on using historical data and statistical models to make predictions about future events or outcomes. It involves techniques such as regression analysis, time series forecasting, and machine learning algorithms.
3. Computational Biology/Bioinformatics: Computational biology and bioinformatics use data science to analyze biological data, including genomic data, protein sequences, and molecular structures. It aims to uncover patterns, understand biological processes, and develop computational models for biological systems.
4. Social Network Analysis: Social network analysis examines the relationships and interactions between individuals or entities in a network. It involves analyzing network structures, identifying influential nodes, and understanding information flow and social dynamics.
5. Image and Video Processing: This subfield focuses on extracting meaningful information from images and videos. It involves techniques such as image recognition, object detection, image segmentation, and video analysis.
6. Natural Language Processing (NLP): NLP focuses on understanding and processing human language. It includes tasks such as text classification, sentiment analysis, named entity recognition, language translation, and chatbot development.
7. Recommender Systems: Recommender systems are used to provide personalized recommendations to users based on their preferences and behavior. They are widely used in e-commerce, content streaming platforms, and personalized marketing.
8. Fraud Detection: Fraud detection applies data science techniques to identify fraudulent activities or patterns. It involves analyzing large volumes of data, identifying anomalies, and developing algorithms to detect fraudulent behavior.
9. Operations Research: Operations research uses mathematical and statistical modeling techniques to optimize complex systems and decision-making processes. It is often used to solve problems related to logistics, supply chain management, resource allocation, and production planning.
10. Spatial Analysis: Spatial analysis focuses on analyzing data with geographic or spatial components. It involves techniques such as geographic information systems (GIS), spatial statistics, and spatial data visualization.

Examples of where/how/when data science is applied in different industries?

1. Transportation and Logistics:
   * Route optimization and fleet management for transportation companies.
   * Demand forecasting for logistics planning and inventory management.
   * Predictive maintenance for vehicle fleets.
   * Real-time traffic analysis and congestion prediction.
2. Energy and Utilities:
   * Smart grid optimization and demand response management.
   * Predictive maintenance for power generation and distribution systems.
   * Energy consumption analysis and optimization for cost savings.
   * Renewable energy forecasting and resource optimization.
3. Marketing and Advertising:
   * Customer segmentation and targeting based on demographic and behavioral data.
   * Sentiment analysis of social media data for brand reputation management.
   * A/B testing and conversion rate optimization.
   * Marketing campaign attribution and ROI analysis.
4. Telecommunications:
   * Churn prediction and customer retention strategies.
   * Network optimization and capacity planning.
   * Fraud detection for telecommunications services.
   * Customer sentiment analysis from call center interactions.

Explain your understanding of Machine Learning, AI and Data Analytics?

1. Machine Learning: Machine Learning is the development of algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed.
2. Artificial Intelligence (AI): AI refers to the field of creating intelligent machines that can perform tasks requiring human-like intelligence. It encompasses techniques like Machine Learning, natural language processing, and computer vision.
3. Data Analytics: Data Analytics involves extracting insights and actionable information from data through techniques like statistical analysis, data mining, and visualization to make data-driven decisions.

Types of Machine Learning and examples of how/when/where they are applied?

1. Supervised Learning (Classification, Regression)

Email spam detection: Classifying emails as spam or non-spam based on labeled training data.

2. Unsupervised Learning (Clustering, Dimensionality Reduction)

Anomaly detection: Identifying unusual patterns or outliers in a dataset, such as fraudulent transactions.

3. Deep Learning (Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs))

Speech recognition: Converting spoken words into text, enabling voice assistants like Siri or Alexa.

4. Semi-Supervised Learning ( Combination of labeled and unlabeled data)

Document categorization: Using a small labeled dataset and a large unlabeled dataset to classify documents.

5. Ensemble Learning ( Random Forests, Gradient Boosting)

Predictive modeling: Combining multiple models to improve accuracy and make robust predictions.

Difference between dataset, dataframe and database?

1. Dataset: A collection of structured or unstructured data with a common theme or purpose.
2. DataFrame: A two-dimensional labeled data structure for tabular data analysis and manipulation.
3. Database: A structured system for efficient storage, retrieval, and management of data.

How do you notice unclean data?

1. Missing values: Look for empty or null values in the dataset. Missing values can indicate incomplete data or errors during data collection.
2. Duplicates: Identify duplicate records or instances within the dataset. Duplicate data can skew analysis results and create biases.
3. Inconsistencies: Check for inconsistent formats, inconsistent naming conventions, or conflicting data within the dataset. For example, inconsistent date formats or variations in representing categorical variables.
4. Outliers: Look for extreme values that deviate significantly from the majority of the data points. Outliers may indicate measurement errors, data entry mistakes, or abnormal observations.
5. Data validation rules: Compare the data against known business rules, constraints, or domain knowledge to identify inconsistencies or violations. For instance, age values that are negative or unrealistically high.
6. Incomplete or inconsistent data: Look for incomplete data where required fields or variables are missing. Inconsistent data can arise when different sources of data collection methods are merged without proper integration.
7. Data distribution: Analyze the distribution of data to identify potential anomalies or unexpected patterns. Skewed distributions or unusual patterns may indicate data issues.
8. Data integrity checks: Perform checks on calculated fields, relationships between variables, or aggregations to identify discrepancies or errors.