**What is Data Science?**

* Data Science is process of getting **insights from data**.
* Data science combines **math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning** with specific subject matter expertise to **uncover actionable insights hidden** in an organization’s data. These insights can be used to guide **decision making and strategic planning**.
* Data Science is about data gathering, analysis and decision-making.
* Data Science is about finding patterns in data, through analysis, and make future predictions.

**Where is Data Science Needed?**

* Data Science is used in many industries in the world today, e.g. banking, consultancy, healthcare, and manufacturing.

**By using Data Science, companies are able to make:**

* Better decisions (should we choose A or B)
* Predictive analysis (what will happen next?)
* Pattern discoveries (find pattern, or maybe hidden information in the data)

**Healthcare:**

* + Predictive analytics to identify disease outbreaks.
  + Medical image analysis for diagnosing diseases like cancer and tumors.
  + Patient data analysis to improve treatment plans and personalized medicine.

**Finance:**

* + Fraud detection using anomaly detection algorithms.
  + Stock market prediction and portfolio optimization.
  + Credit scoring and risk assessment for loans.

**E-commerce:**

* + Recommendation systems for personalized product suggestions.
  + Customer segmentation for targeted marketing campaigns.
  + Price optimization and dynamic pricing strategies.

**4.Manufacturing:**

* 1. Quality control using sensors and statistical analysis.
  2. Predictive maintenance to reduce downtime and optimize production.
  3. Supply chain optimization and demand forecasting.

**5.Transportation and Logistics:**

* 1. Route optimization for delivery and transportation.
  2. Real-time tracking of vehicles and assets.
  3. Demand prediction for public transportation services.

**Version 2 :**

**Healthcare:**

* **Predictive Analytics:** Using data to predict disease outbreaks, helping healthcare systems prepare and respond effectively.
* **Medical Image Analysis:** Using data to analyze medical images, like X-rays, to diagnose diseases such as cancer and tumors.
* **Patient Data Analysis:** Using patient data to improve treatment plans and provide personalized medical care.

**Finance:**

* **Fraud Detection:** Using data to detect unusual patterns and identify fraud in financial transactions.
* **Stock Market Prediction:** Using data to make predictions about stock market trends and optimize investment portfolios.
* **Credit Scoring:** Using data to assess creditworthiness and determine risk for loans.

**E-commerce:**

* **Recommendation Systems:** Using data to suggest products based on customer preferences, enhancing personalized shopping experiences.
* **Customer Segmentation:** Using data to group customers based on similar traits for targeted marketing campaigns.
* **Price Optimization:** Using data to set optimal prices for products and employ dynamic pricing strategies.

**Manufacturing:**

* **Quality Control:** Using data from sensors and statistical analysis to maintain product quality and identify defects.
* **Predictive Maintenance:** Using data to predict when machines need maintenance, reducing downtime and improving production efficiency.
* **Supply Chain Optimization:** Using data to optimize the supply chain process and forecast demand accurately.

**Transportation and Logistics:**

* **Route Optimization:** Using data to find the best routes for delivery and transportation, saving time and resources.
* **Real-time Tracking:** Using data to track vehicles and assets in real-time, ensuring efficient operations.
* **Demand Prediction:** Using data to predict passenger demand for public transportation services, helping plan schedules effectively.

In simple words, data science involves using data to solve various problems in different industries. It helps in making predictions, analyzing images or information, improving processes, and providing personalized experiences. Whether it's preventing diseases, optimizing finances, enhancing shopping, maintaining machines, or improving transportation, data science plays a crucial role in making things better and more efficient.

**Who is a Data Scientist?**

1. Data scientists are **analytical data experts** who possess the technical skills to address complex problems.
2. They gather, analyze, and interpret vast amounts of data while working with a variety of computer science, mathematics, and statistics-related concepts.
3. They have a duty to offer perspectives that go beyond statistical analysis.
4. Data scientist positions are accessible in both the public and private sectors, including finance, consulting, manufacturing, pharmaceuticals, government, and education.

**How Does a Data Scientist Work?**

1. **Ask the right questions** - To understand the business problem.
2. **Explore and collect data** - From database, web logs, customer feedback, etc.
3. **Extract the data** - Transform the data to a standardized format.
4. **Clean the data** - Remove erroneous values from the data.
5. **Find and replace missing values** - Check for missing values and replace them with a suitable value (e.g. an average value).
6. **Normalize data** - Scale the values in a practical range
7. **Analyze data**, find patterns and make future predictions.
8. **Represent the result** - Present the result with useful insights in a way the "company" can understand.
   1. Business Analyst
   2. Database Engineer
   3. Data Analyst
   4. Data Engineer
   5. Data Scientist
   6. Research Scientist
   7. Software Engineer
   8. Statistician
   9. Product Manager
   10. Project Manager

**Business Analyst:**

* A person who uses data to understand business needs, analyze trends, and make informed decisions to improve operations and strategies.

**Database Engineer:**

* Someone who designs, develops, and maintains databases that store and organize data securely, ensuring efficient data retrieval and management.

**Data Analyst:**

* A professional who examines data to discover insights, create reports, and present findings that help businesses make informed decisions.

**Data Engineer:**

* An expert who designs and builds the infrastructure and pipelines to collect, process, and store data for analysis and modeling.

**Data Scientist:**

* A role focused on exploring complex data, building predictive models, and generating insights to solve intricate problems and make predictions.

**Research Scientist:**

* Someone who conducts research using data, applying scientific methods to explore new findings, validate hypotheses, and contribute to advancements in various fields.

**Software Engineer:**

* A developer who creates applications, tools, and systems to process, visualize, and analyze data efficiently and effectively.

**Statistician:**

* An expert in statistics who uses mathematical techniques to analyze data, draw conclusions, and make predictions based on patterns.

**Product Manager:**

* A person responsible for developing data-driven products and features, aligning them with user needs and business goals.

**Project Manager:**

* Someone who oversees data-related projects, ensuring they are completed on time, within budget, and meet the specified objectives.

**Version 2 :**

**Business Analyst (BA):**

* Gathers information about viewer preferences and behaviors.
* Identifies potential features for the recommendation system.

**Database Engineer:**

* Designs the structure of the database for user profiles, viewing history, and content metadata.
* Ensures quick retrieval of personalized recommendations by optimizing the database.

**Data Analyst:**

* Analyzes user interaction data like watch history, ratings, and search queries.
* Identifies patterns and trends and creates visualizations and reports to understand popular genres, viewing habits, and user segments.

**Data Engineer:**

* Builds data pipelines to transform raw user activity data into usable formats.
* Integrates the data with the recommendation engine to ensure a smooth flow of information.

**Data Scientist:**

* Develops machine learning algorithms that power the recommendation system.
* Creates models that consider user preferences and context to provide accurate suggestions.

**Research Scientist:**

* Explores advanced recommendation techniques, such as deep learning or reinforcement learning algorithms.
* Enhances the accuracy and effectiveness of the recommendation system.

**Software Engineer:**

* Develops the user interface where users receive recommendations.
* Integrates the recommendation engine's output into the interface seamlessly.

**Statistician:**

* Conducts experiments to test different recommendation approaches.
* Uses statistical analysis to measure the impact of personalized suggestions on user engagement and retention.

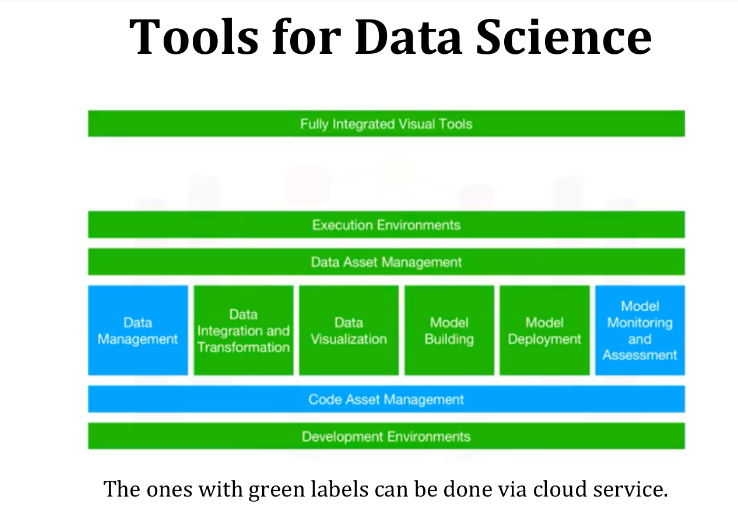
**Product Manager:**

* Oversees the recommendation system's development, working with various teams to ensure it aligns with business goals.

**Project Manager:**

* Defines project milestones, manages resources, and coordinates tasks to ensure the recommendation system is developed and integrated into the platform on time.

**Tools for Data science** :



**Data Management:**

* Process of collecting, persisting, and retrieving data securely, efficiently, and cost-effectively.
* Data is gathered from various sources like Twitter, Flipkart, Media, Sensors, etc.
* Data can be structured (customer data) or unstructured (log files, videos, IoT, social media, etc.).
* Tools include Relational databases (MySQL, PostgreSQL), NoSQL databases (MongoDB, Apache Cassandra), File-based tools (Hadoop, Ceph), Elasticsearch for text data.

**Data Integration and Transformation:**

* Involves Extracting, Transforming, and Loading (ELT) data.
* Data often distributed in multiple repositories.
* Extraction gathers data from various sources into a central repository like a Data Warehouse.
* Data Transformation modifies values, structure, and format of data.
* Transformed data is loaded into the Data Warehouse.
* Tools: Apache AirFlow, Apache Kafka, Apache Nifi, Apache SparkSQL, NodeRED.

**Data Visualization:**

* Graphical representation of data using charts, plots, maps, animations, etc.
* Conveys data effectively for decision-makers.
* Examples include bar charts, treemaps, line charts, map charts.
* Visualization tools: Pixie Dust, Hue, Kibana, Apache Superset.

**Model Building:**

* Uses machine learning algorithms to analyze patterns in data.
* The system learns and provides predictions or decisions on new data.
* Tools like IBM Watson Machine Learning offer tools and services for building models.

**Model Deployment:**

* Integrating a developed machine learning model into a production environment.
* Makes the model available to third-party applications through APIs.
* Business users make data-based decisions using these applications.
* Deployment tools: Apache PredictionIO, Seldon, Kubernetes, TensorFlow services.

**Model Monitoring:**

* Continuous quality checks to ensure model accuracy, fairness, and robustness.
* Tools like Fiddler track model performance in production.
* Monitoring tools: ModelDB, Prometheus, IBM AI Fairness 360.

**Model Assessment:**

* Uses evaluation metrics (F1 score, true positive rate) to understand model performance.

**Code Asset Management:**

* Involves versioning and collaborative features for teamwork.
* Tools: Git, GitHub, GitLab, Bitbucket.

**Development Environments:**

* IDE tools help data scientists implement, test, and deploy their work.

**Execution Environments:**

* Tools for data processing, model training, and deployment.

**API and REST API** :

**API (Application Programming Interface):**

* API is like a bridge that allows different software applications to communicate and work together.
* Imagine it as a waiter in a restaurant who takes your order and communicates it to the kitchen.

**REST API (Representational State Transfer API):**

* REST API is a specific type of API that allows software applications to communicate over the Internet.
* It's like a special kind of waiter that helps you order food from a restaurant's menu while using all the restaurant's resources.
* **Representational:** It means that data is represented in a specific format, like how a menu in a restaurant lists dishes.
* **State:** This refers to the condition of the resources (like food) that the API deals with. For instance, whether the dish is available or not.
* **Transfer:** It's about moving data between different systems. Just like how your order is transferred from your table to the kitchen.

So, a REST API is like a waiter that helps different software applications communicate using the Internet, representing data in a certain way and managing the state of resources, just as a waiter in a restaurant helps you order food and manages the status of dishes.

**Supervised Learning:**

* Think of it like teaching a computer by showing it examples with answers.
* If you're teaching it to recognize cats and dogs, you'd show it pictures of cats labeled as "cat" and dogs labeled as "dog."
* It learns from these labeled examples so it can make accurate predictions on new, unlabeled data.

**Regression:**

* Imagine you're predicting something numerical, like predicting the price of a house based on its features.
* It's like guessing the exact number of how much something will be.

**Classification:**

* It's like sorting things into categories, just like sorting emails into spam and not-spam folders.
* The computer learns from labeled examples of each category to make correct classifications in the future.

**Unsupervised Learning:**

* Imagine you're finding hidden patterns in a bunch of data without anyone telling you what the patterns are.
* It's like discovering groups of similar things in a big collection of items without being given any specific categories.

**Clustering:**

* Think of it like grouping similar things together without knowing what each group represents.
* It's like sorting different types of fruits into baskets without being told the names of the fruits.

**Anomaly Detection:**

* Imagine spotting unusual things in a crowd without anyone pointing them out.
* It's like finding unusual behavior in credit card transactions that might indicate fraud.

**Reinforcement Learning:**

* Think of it like training a dog to do tricks. The dog learns by trying different actions and getting rewards for good actions.
* The computer learns by taking actions in a simulated environment, trying to get the most rewards over time

**Explanation of each topic and sub-topic of the Data Science Methodology**

**Business Understanding:**

* **Explanation:** This phase involves obtaining a clear understanding of the problem you're aiming to solve. This requires seeking clarification, defining goals and objectives, and involving stakeholders to gain insights into business requirements.
* **Keywords/Keyphrases:** Clarification, problem to be solved, data for core question, goals, objectives, stakeholder, business requirements.

**Case Study - Healthcare Budget Allocation:**

* **Explanation:** The case study presents the problem of optimizing a limited healthcare budget to provide quality care. The focus is on reducing readmissions, particularly for patients with congestive heart failure (CHF).
* **Keywords/Keyphrases:** Healthcare budget allocation, quality care, readmissions, congestive heart failure (CHF).

**Identifying Goals and Objectives:**

* **Explanation:** Clearly define the overarching goals and specific objectives that your solution aims to achieve, ensuring alignment with stakeholder expectations.
* **Keywords/Keyphrases:** Goals, objectives, alignment, stakeholder expectations.

**Stakeholder Involvement:**

* **Explanation:** Engage stakeholders to ensure their support and guidance throughout the project. Stakeholders play a crucial role in setting the project's direction and providing necessary support.
* **Keywords/Keyphrases:** Stakeholder involvement, support, guidance, project direction.

**Identifying Business Requirements:**

* **Explanation:** Specify the requirements that your solution must fulfill to address the problem. This step outlines what you need to predict and understand from the data.
* **Keywords/Keyphrases:** Business requirements, predict, understand, data needs.

**Analytic Approach:**

* **Explanation:** Determine the most suitable analytical approach for your problem. In this case, you're looking to identify patterns effectively using Machine Learning.
* **Keywords/Keyphrases:** Analytic approach, suitable, Machine Learning, identify patterns.

**Machine Learning:**

* **Explanation:** Machine Learning involves learning patterns and relationships in data without explicit programming. It helps identify hidden trends and relationships.
* **Keywords/Keyphrases:** Machine Learning, learn patterns, hidden trends, relationships.

**Data Requirements and Collection:**

* **Explanation:** Compare data collection to preparing ingredients for cooking. Define the cohort of patients that will be used for analysis.
* **Keywords/Keyphrases:** Data collection, ingredients, cohort, patients.

**Cohort Selection:**

* **Explanation:** Define the specific criteria for selecting the patient cohort for analysis. This includes factors like patient location, diagnosis, enrollment duration, and more.
* **Keywords/Keyphrases:** Cohort selection, criteria, patient location, diagnosis, enrollment duration.

**Data Collection Sources:**

* **Explanation:** Gather data from various sources, including corporate data warehouses and hospital systems. Assess the available data and identify any discrepancies.
* **Keywords/Keyphrases:** Data collection sources, corporate data warehouses, hospital systems, assess data.

**Data Understanding:**

* **Explanation:** Understand the collected data by applying descriptive statistics and visualization. Check for data quality issues and iterate to refine your problem-solving approach.
* **Keywords/Keyphrases:** Data understanding, descriptive statistics, visualization, data quality, refine approach.

**Data Preparation:**

* **Explanation:** Clean, transform, and engineer features in the data to make it suitable for analysis. Feature engineering involves selecting and transforming variables for modelling.
* **Keywords/Keyphrases:** Data preparation, clean, transform, feature engineering, suitable for analysis.

**Defining Readmission and CHF Admission:**

* **Explanation:** Define what constitutes a "readmission" and a "CHF admission" based on time frames and events. Aggregate and create variables to represent patient history.
* **Keywords/Keyphrases:** Defining readmission, CHF admission, time frames, events, aggregate, variables, patient history.

**Data Collection and Availability:**

* **Explanation:** Review available data sources, such as corporate data warehouses and patient record systems. Consider data that is wanted but currently unavailable.
* **Keywords/Keyphrases:** Data collection, available sources, corporate data warehouses, patient record systems, wanted data, unavailable data.

**Modelling and Evaluation:**

* **Explanation:** Build predictive or descriptive models to address the problem. Evaluate model performance using training and test data sets.
* **Keywords/Keyphrases:** Modelling, predictive models, descriptive models, evaluate performance, training and test data sets.

**Initial Model Selection:**

* **Explanation:** Choose an initial model, like a decision tree classification model, to achieve your goal. Assess the model's accuracy and consider improvements if needed.
* **Keywords/Keyphrases:** Initial model selection, decision tree classification model, goal, accuracy, improvements.

**Model Evaluation and Improvement:**

* **Explanation:** Analyze alternative models and assess their performance using measures like sensitivity and specificity. Improve the model to better answer the problem.
* **Keywords/Keyphrases:** Model evaluation, alternative models, performance measures, sensitivity, specificity, improve model.

**Deployment and Feedback:**

* **Explanation:** Deploy the model and ensure stakeholders understand its results. Gather feedback to refine the model and assess its real-world impact.
* **Keywords/Keyphrases:** Deployment, stakeholders, model results, gather feedback, refine model, assess impact.

**Refinement and Redeploy:**

* **Explanation:** Continuously refine the model based on feedback and new data. Redeploy the improved model throughout the life of the intervention program.
* **Keywords/Keyphrases:** Refinement, continuous improvement, feedback, new data, redeploy, intervention program.

import requests

from bs4 import BeautifulSoup

def get\_geeksforgeeks\_python\_content():

"""Get the content of the Python programming language page on Geeksforgeeks."""

url = 'https://www.geeksforgeeks.org/python-programming-language/'

response = requests.get(url)

soup = BeautifulSoup(response.content, 'html.parser')

title = soup.title

print(f'Title: {title}')

content = soup.find('div', class\_='entry-content').find\_all('p')

print(content)

leftbar = soup.find('div', id='main').find('ul', class\_='leftBarList').find\_all('li')

print(leftbar)

images = soup.select('img')

images\_list = [{'src': image['src'], 'alt': image['alt']} for image in images]

print(images\_list)

get\_geeksforgeeks\_python\_content()z