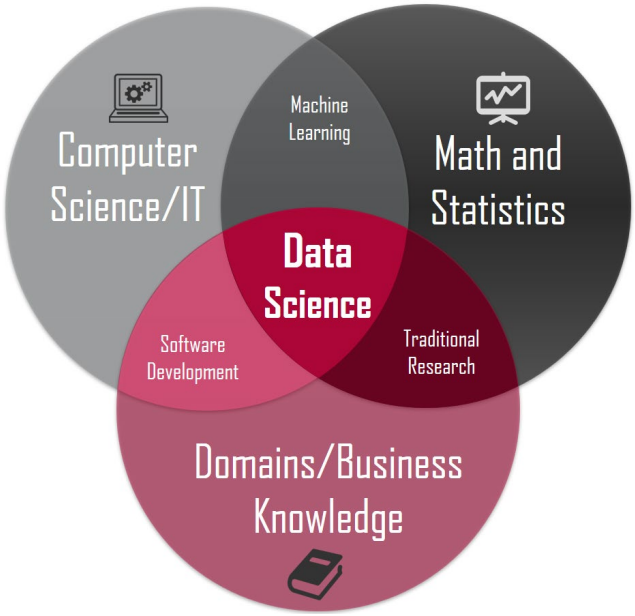
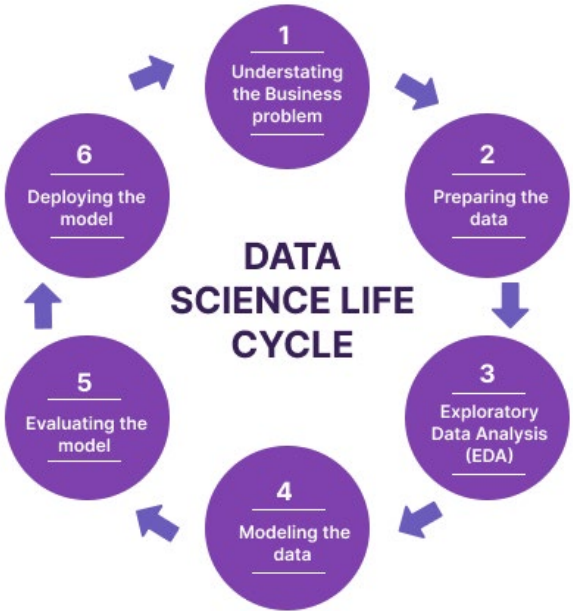
Intro to Data Science

Data Science is a combination of several subjects that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it.



1. Data Science Lifecycle

Data science follows a series of steps called the data science lifecycle. It starts with getting the data, then cleaning and analyzing it to uncover patterns and insights. Next, models are built to make predictions and achieve business goals. Finally, the models are evaluated to ensure their accuracy. This process helps businesses make informed decisions and achieve their objectives using data-driven approaches.



1.1. Understating the Business problem

Understanding the business problem starts with closely examining business trends, studying relevant data analytics, and conducting thorough market research within the industry. Stakeholders assess the internal infrastructure, available resources, project timeline, and technological requirements. To define the business problem, key considerations include identifying the issue that needs resolution, evaluating the potential value of the project, assessing potential risks, and creating a flexible, high-level project plan. This precise process ensures that the data science project aligns with business objectives and sets the foundation for a successful outcome.

1.2. Preparing the data

Preparing the data to understand the business problem and extracting information to solve the problem.

Steps for data preparations are:

- Find and collect data related to the problem.
- Combine the data sets.
- Clean the data to find the missing values.
- Handle the missing values by removing or imputing them.
- Remove any detected errors.
- Use the box plots for detecting outliers and handling them.

1.3. Exploratory Data Analysis (EDA)

EDA involves examining the data to gain insights and identify patterns. This includes checking if the data is accurate and free of duplicates, missing values, and null values. It also involves recognizing the important factors in the data set and eliminating any irrelevant noise that can reduce the accuracy of the conclusions. EDA typically takes up 70% of the data science project life cycle time and can provide valuable insights.

1.4. Modeling the data

Modeling involves selecting the right model type, that depends on whether the issue is classification, regression, or clustering. Then some algorithms must be chosen and implemented. It is also important to determine the model's ideal hyperparameter values to avoid overfitting for example. Hyperparameter tuning ensures that the model is accurate and reliable. Modeling is a crucial step in the data science process that helps ensure the success of the project.

1.5. Evaluating the model

Evaluating the model involves measuring the model's performance and determining if it is suitable for the business problem. Two techniques used widely to evaluate models are Hold-Out and Cross-Validation.

- Hold-Out evaluation is the process of testing a model with data that is distinct from the data it was trained on.
- Cross-Validation involves splitting the data into sets and using them to analyze the performance of the model.

Metrics such as Accuracy, ROC-AUC, Precision-Recall, Log-Loss, MSAE, MSPE, R Square, Adjusted R Square, Mutual Information, and Rand Index are used to evaluate the models. This step helps to ensure that the right model is chosen for the business problem.

1.6. Deploying the model

The final step in the data science lifecycle is deploying the model. It involves selecting the suitable delivery method for the model, such as Tableau or a cloud-based platform, and updating any shortcuts used during the model phase to systems fit for production. This step is typically carried out by engineering-focused team members, such as data engineers, cloud engineers, machine learning engineers, application developers, and quality assurance engineers.

Data science is a complex process with different stages that need careful attention. If data collection is done poorly, it can lead to loss of important information and result in a subpar model. Properly cleaning the data is crucial for the model to work as expected. Additionally, thorough evaluation of the model is essential for real-world performance. It's important to give each stage, from understanding the business needs to deploying the model, the right amount of attention, time, and effort to ensure success.

2. Different Software Tools and Programming Languages Used in Data Science

Programming Language/Software	Description	Advantages	Disadvantages
R	A statistical programming language and environment for data analysis, visualization, and modeling.	• Strong support for statistical analysis and hypothesis testing. • Interactive data visualization capabilities with packages like ggplot2. • Large community of statisticians and data scientists.	• Steeper learning curve compared to Python, especially for those with no prior programming experience. • Limited support for big data processing and scalability.
SAS	A proprietary software suite for data analysis, reporting, and modeling.	• Established and widely used in industries like finance and healthcare. • Comprehensive and well-documented libraries for data analysis and reporting. • Robust data integration and data management capabilities.	• Expensive licensing and limited availability of free/open-source versions. • Limited community support compared to open-source options like Python and R. • The learning curve can be challenging for newcomers, especially those who are not familiar with SAS.
Python	A versatile and widely used programming language with extensive libraries for data manipulation, analysis, and visualization.	• Extensive libraries for machine learning and deep learning. • Strong support for big data processing with libraries like Apache Spark. • High readability and easy to understand syntax.	• Not as efficient as lower-level languages like C/C++ for computationally intensive tasks. • Limitations in multi-threading support, which could potentially affect performance in certain scenarios
MATLAB	A proprietary programming language and environment for scientific computing, including data analysis and modeling.	• Strong support for linear algebra, numerical computing, and signal processing. • Interactive environment for prototyping and testing algorithms.	• Expensive licensing, especially for commercial use. • Limited support for big data processing and scalability. • Less versatile compared to Python and R for broader data science tasks.
SQL	A domain-specific language used for managing and querying relational databases, commonly used for data extraction, transformation, and loading (ETL) tasks in data science pipelines.	• Widely used in industry for managing and querying databases, making it a valuable skill for data scientists. • Ideal for managing extensive data operations, particularly those involving structured data. • Well-suited for data integration tasks in data pipelines.	• Limited support for advanced data analysis, modeling, and visualization tasks compared to programming languages like Python and R. • Steeper learning curve for complex queries and operations.
Tableau	A popular data visualization and business intelligence tool that allows users to create interactive dashboards and visualizations from various data sources.	• User-friendly interface with drag-and-drop functionality for creating visually appealing dashboards and visualizations. • Wide range of data connectors for integrating with various data sources. • Powerful data visualization features for exploring and analyzing data.	• Limited data analysis and data processing capabilities compared to programming languages like Python and R. • Can be expensive in terms of licensing costs. • Limited customization options compared to coding-based tools.
Power BI	A business analytics tool by Microsoft that enables users to create interactive dashboards and reports for data visualization and analysis.	• Integrated with Microsoft ecosystem, making it easy to connect with various data sources such as Excel, SQL Server, and Azure. • Allows users to easily create interactive dashboards and reports using a seamless drag-and-drop functionality. • Powerful data visualization and data analysis features.	• Limited data processing capabilities compared to programming languages like Python and R. • May require additional licensing costs for advanced features and functionality. • Limited customization options compared to coding-based tools.

3. Different Types of Data Used in Data Science

In data science, there are various types of data. These types of data have different characteristics and require different techniques and tools for analysis, processing, and modeling. Understanding the different types of data is essential in data science to appropriately handle them for analysis and modeling purposes.

Data Type	Description	Advantages	Disadvantages	Examples
Structured Data	Data that is organized and stored in a specific format, such as a table with rows and columns, with fixed fields and data types.	Easy to analyze using SQL queries	Limited flexibility and scalability	Sales data in a relational database
Unstructured Data	Data that does not have a specific format or structure and does not follow a predefined schema.	Huge sources of information	Hard to analyze and process	Social media posts
Semi-structured Data	Data that has some structure, but not fully organized or formatted like structured data.	Mix of structured and unstructured data	May require specialized techniques to handle	Emails with subject, sender, and body
Image Data	Data that represents visual information, such as photos or images.	Image recognition, computer vision	Large data size, complex processing	Digital images, medical scans
Audio Data	Data that represents audio or sound signals.	Speech recognition, audio analysis	Requires audio processing expertise	Speech recordings, music files
Text Data	Data that represents textual information, such as text documents or emails.	Natural language processing	Language variations, text ambiguity	News articles, customer reviews