Applied Data Science with Python

Introduction to Data Science

Learning Objectives

By the end of this lesson, you will be able to:

- Explain the basics of data science and its application
- List the five steps of the data science process
- Explore the data preparation, model building, and evaluation processes
- Describe the optimization and prediction processes

Business Scenario

ABC is a FinTech company that is a payment gateway for most online websites. The organization encounters many payment failures through its gateway, which have been reported by most merchants.

ABC decides to find all payment details, understand the reasons for failure, and predict the grounds for future failures. This will also help the company optimize the system.

However, to do so, ABC will have to explore data science, optimization, and the prediction of new data.

Introduction to Data Science

Discussion: Data Science

Duration: 10 minutes

What is data science?

- What are the various applications of data science?
- What are the primary approaches of data science?

Data Science

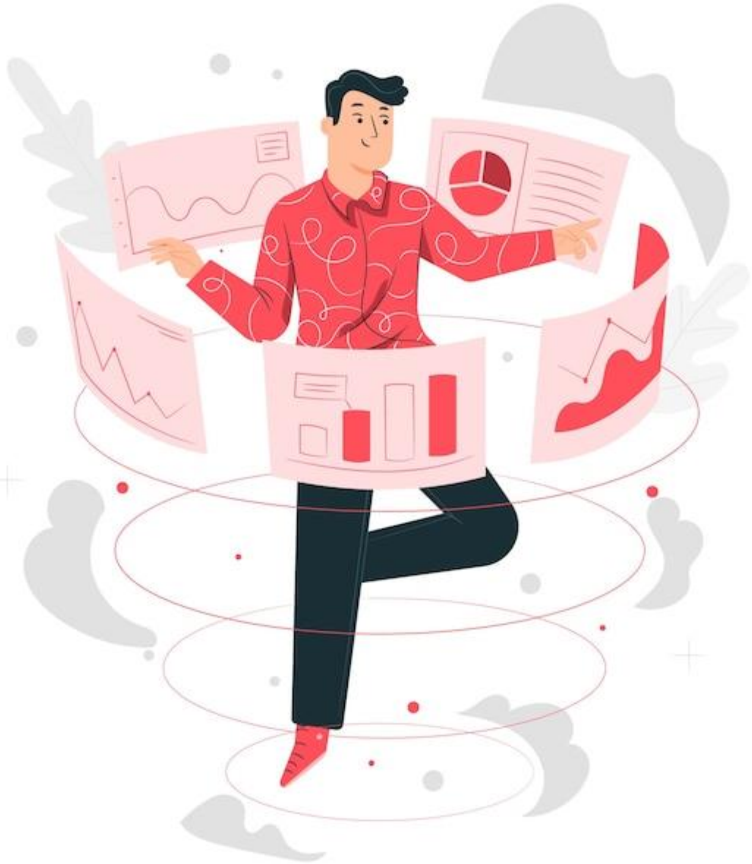
A world without data would be like a world in darkness, resulting in:

- Uninformed beliefs
- Decisions made by human intuition or gut feeling

Having access to large amounts of data and applying statistical inferences opens a plethora of possibilities

Data Science

Data Science is a multidisciplinary approach that combines the practices of mathematics, statistics, probability, and programming to analyze large amounts of data.



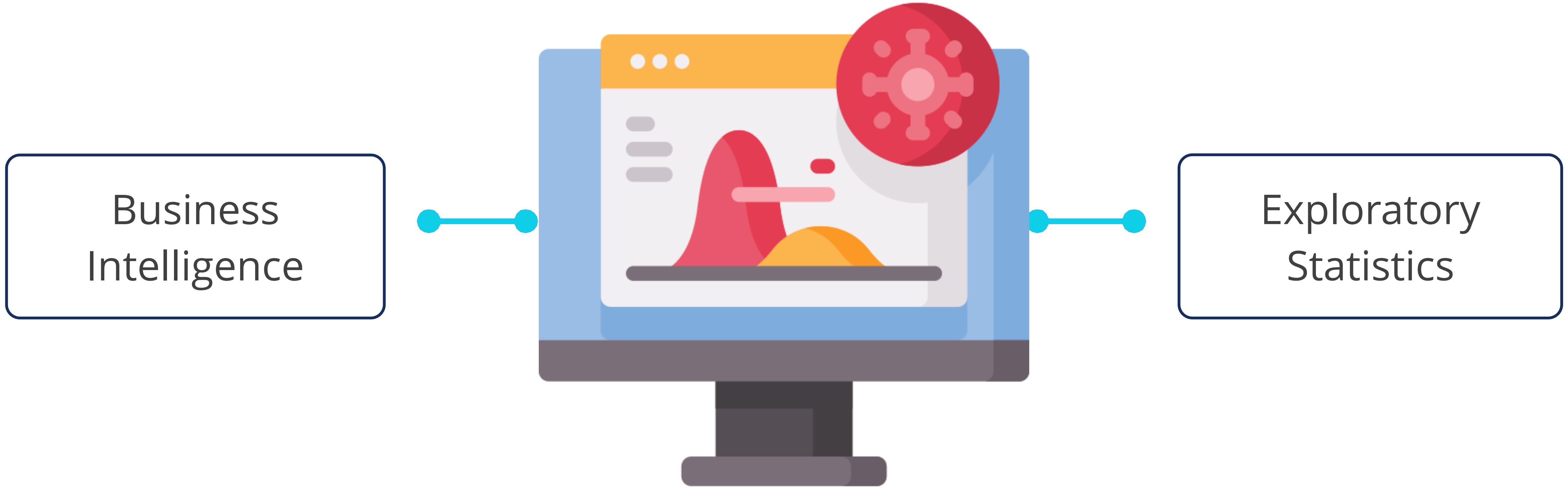
Example

Using a search engine or buying from Amazon gives valuable inputs to data-science-based software systems working in the background.

Data on interactions with online platforms is gathered to understand user preferences and suggest search results or items to buy.

Data Science

Data science is more ambitious than traditional data analysis approaches like:



Business Intelligence

Exploratory Statistics

It attempts to generate assumptions or predictions informed by data and employs them as the foundation for decision-making.

Application of Data Science

Some examples of data science applications:

Healthcare: Identify and predict diseases

Transportation: Optimize shipping logistics

E-commerce: Predict most popular products and provide recommendations.

Fintech: Evaluate credit reports and financial backgrounds for loan eligibility.

Probing the World

Data science is useful for active data gathering by probing the world.

Example: in the social media

- Goal: Social media companies rely on data science to provide the right content to their users. They can create machine-learning algorithms such as recommendation systems to achieve it.

Pattern Discovery

Data science is used to analyze large datasets of health-related information to identify patterns of disease incidence, prevalence, and mortality rates.

Predicting Future Events

Data science is used in driverless cars, which help reduce accidents.

Example

When the data is fed to the machine learning algorithm, it can find the speed limit on a given road, and how to handle different situations while driving.

Additional Applications of Data Science

Data science helps to understand:

Natural language

Computer vision

Psychology

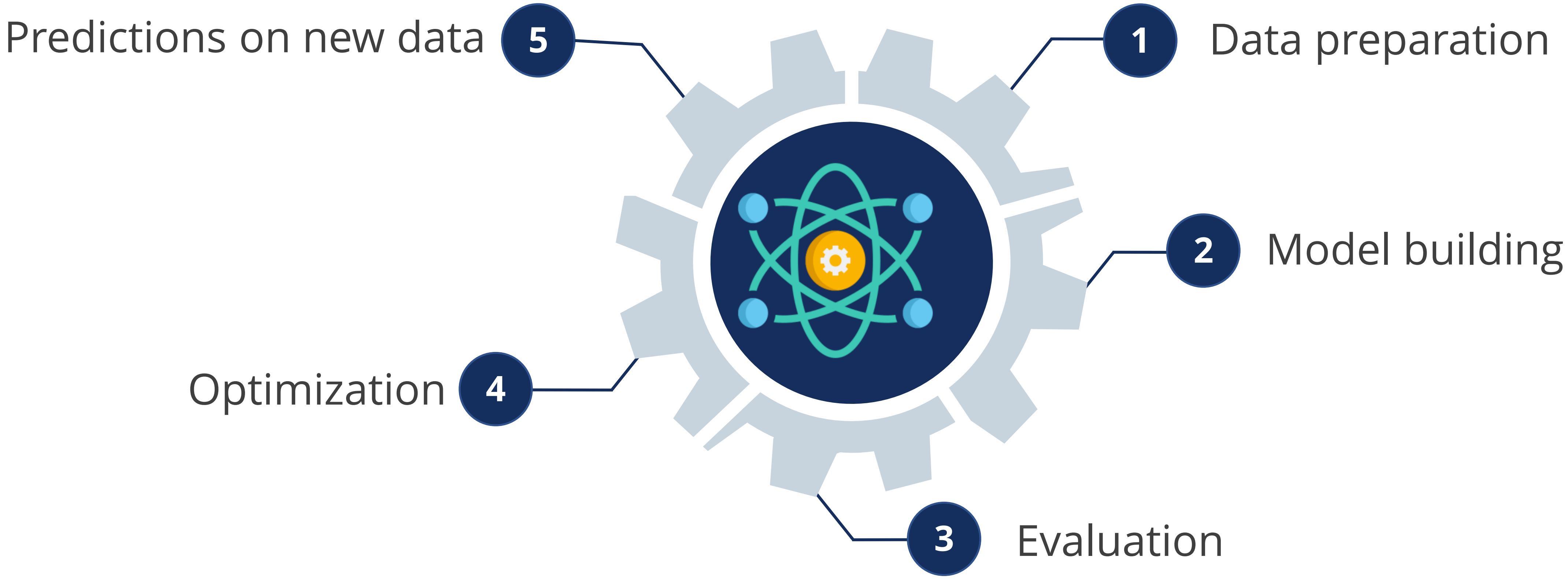
Neuroscience

This can help provide deeper insights into human behavior.

Data Science Process

Data Science Process

The data science process has five steps:



1 Data preparation

2 Model building

3 Evaluation

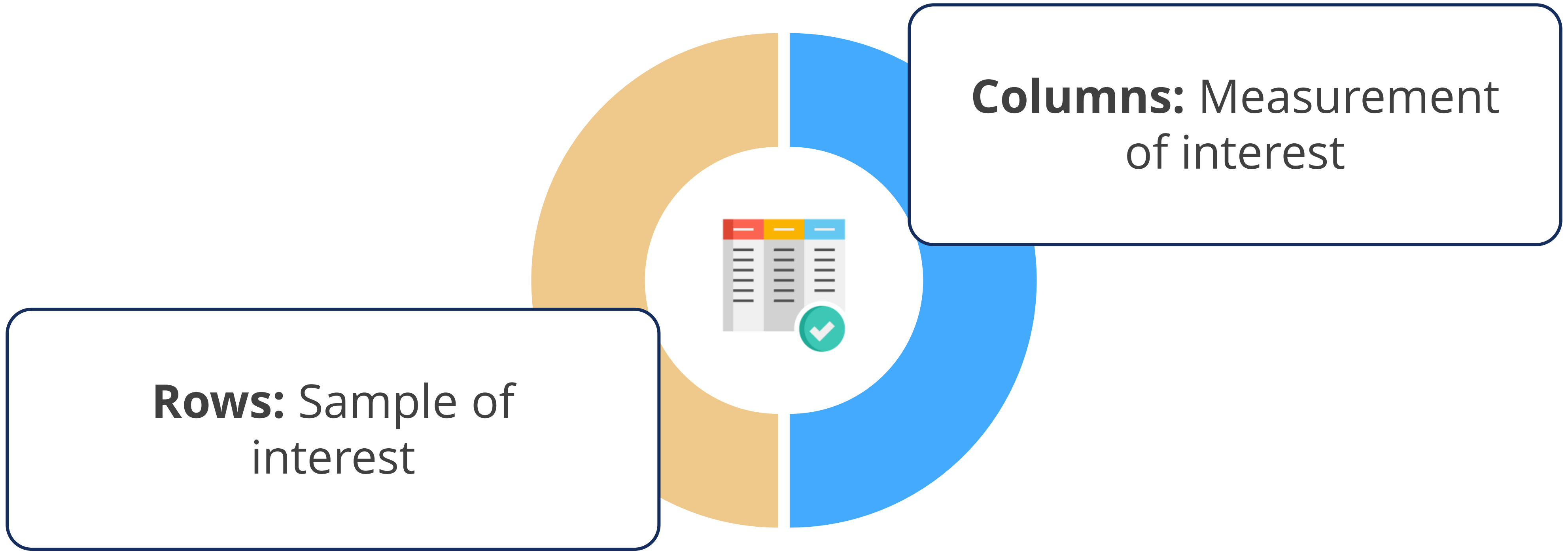
4 Optimization

5 Predictions on new data

While these steps must be followed in order, most real-world data science applications require revisiting each step multiple times, iteratively.

Step 1: Data Collection and Preparation

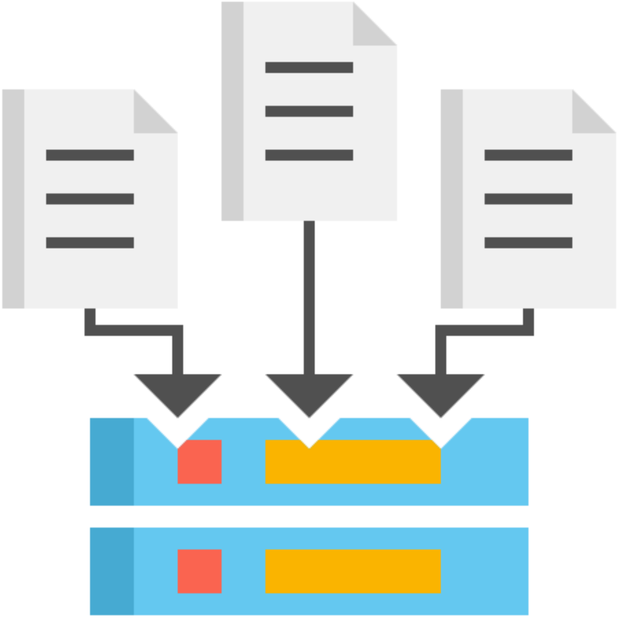
It involves gathering data from disparate sources and arranging it in easy-to-read tables with rows and columns.



Columns: Measurement of interest

Rows: Sample of interest

Step 1: Data Collection and Preparation



The columns are called **features**.

The variable that the user is trying to predict is called the **target**.

Algorithms, with a few exceptions, always require data in row and column format.

Step 1: Data Collection and Preparation

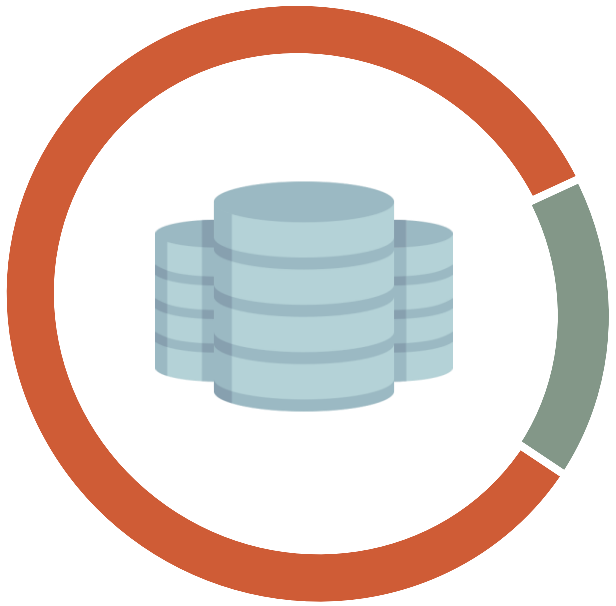
Real-world data often contains missing or noisy elements, therefore cleaning and preparing the data is critical for ensuring greater accuracy before building the model.

Data scientists spend up to 80% of their effort on data cleansing and preparation.

Step 1: Data Collection and Preparation

Data is then split into two sets:

Training set 🟧



🟩 **Testing set**

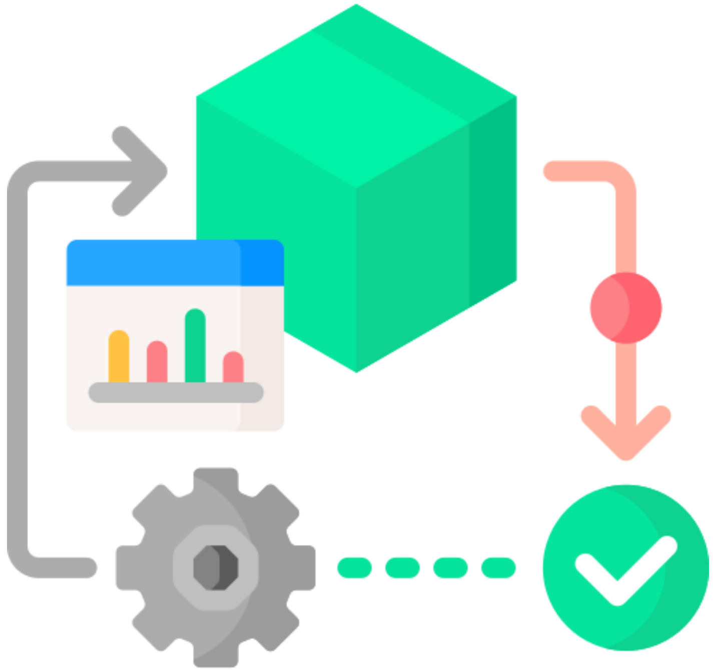
Use **80%** of the data to train the model

Use **20%** of the data to test the accuracy

The most typical ratio is 80:20. However, other ratios can be employed depending on various factors.

Step 2: Training the Model

Sophisticated mathematical modeling techniques are used to:



- Find patterns
- Make predictions

In this step, the model learns from the prepared data.

Step 2: Training the Model

Algorithms are broadly divided into two kinds:

Classification

Predicting a categorical value (class) from the given data

Regression

Predicting a numeric value from the given data

A training set is used to train the model.

Step 3: Evaluating Model Performance

Once a model has learned from the data, its performance is evaluated by using a test dataset.

It assesses the model's performance using new data.

Model evaluation provides insights into how it will perform when fed real-world data.

Models provide measures for accuracy to evaluate the success rate.

Step 4: Optimizing Model Performance

It involves making modifications to the current state of the model in order to improve its efficiency and accuracy.

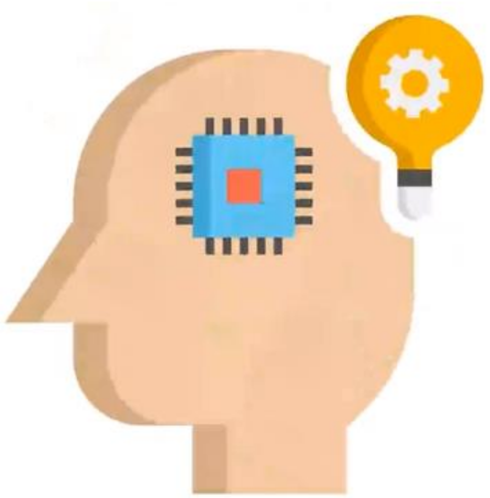
The accuracy of a model is improved in the following ways:

- Tuning the model parameters
- Selecting a subset of features

Essentially, when a data model is optimized, it performs better and gives more accurate results.

Step 4: Optimizing Model Performance

Machine learning algorithms are configured with parameters specific to the underlying algorithm.



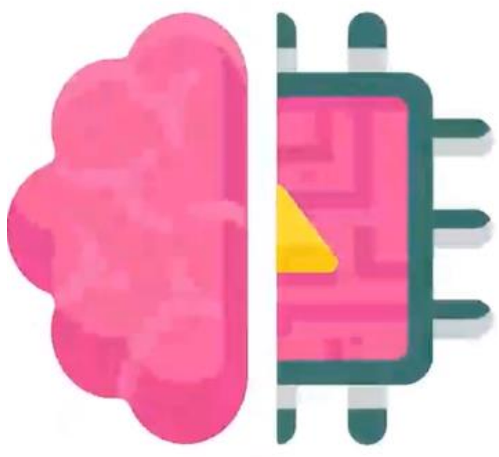
Tuning the model parameters

The optimal value of these parameters often depends on the type and structure of the data.

The value of each parameter or their combination can impact the performance of the model.

Step 4: Optimizing Model Performance

Having many features can hinder your model's performance



Selecting a subset of features

- Some features cause noise that prevents the algorithm from finding actual patterns and relationships in the data.
- Feature selection can be difficult because certain features may appear informative but add no value to the model.

Step 5: Predictions on New Data

Optimized model parameters can be used to predict the desired target variable from real-world data.

Discussion

Duration: 10 minutes

What is data science?

- What are the various applications of data science?

Answer: Data science encompasses a wider scope and ambition compared to traditional data analysis approaches, including:

- Business intelligence
- Exploratory statistics

- What are the primary approaches of data science?

Answer: The primary approaches of data science include:

- Healthcare: identifying and predicting diseases
- Transportation: optimizing shipping logistics
- Fintech: evaluating credit reports and financial backgrounds for loan eligibility
- E-commerce: predicting the most popular products and providing recommendations

Key Takeaways

- Data science involves the analysis and interpretation of data to generate actionable insights.
- The data science process has five steps: data preparation, model building, evaluation, optimization, and predictions of new data.
- The data collected from various sources is stored in a table with columns and rows.
- Optimized model parameters can be used to predict the desired target variable from the real-world data.

Knowledge Check

Knowledge Check

1

What are the five steps of the data science process?

A. Data Collection and Preparation, Model Building, Optimization, Testing, and Predictions on New Data

B. Data Collection and Preparation, Model Building and Evaluation, Optimization, Preprocessing, and Predictions on New Data

C. Data Collection and Preparation, Model Building, Evaluation, Optimization, and Predictions on New Data

D. Data Collection and Preparation, Model Building, Testing, Optimization, and Predictions on New Data

Knowledge Check

1

What are the five steps of the data science process?

A. Data Collection and Preparation, Model Building, Optimization, Testing, and Predictions on New Data

B. Data Collection and Preparation, Model Building and Evaluation, Optimization, Preprocessing, and Predictions on New Data

C. Data Collection and Preparation, Model Building, Evaluation, Optimization, and Predictions on New Data

D. Data Collection and Preparation, Model Building, Testing, Optimization, and Predictions on New Data

The correct answer is **C**

The data science process has five steps: data collection and preparation, model building, evaluation, optimization, and predictions on new data.

Knowledge Check

2

What is the purpose of data cleaning and preparation in the data science process?

A. To increase the size of the dataset

B. To decrease the accuracy of the model

C. To ensure greater accuracy while building the model

D. To reduce the amount of data required for the model

Knowledge Check

2

What is the purpose of data cleaning and preparation in the data science process?

A. To increase the size of the dataset

B. To decrease the accuracy of the model

C. To ensure greater accuracy while building the model

D. To reduce the amount of data required for the model

The correct answer is **C**

Data cleaning and preparation are important steps in the data science process to ensure greater accuracy while building the model.

Knowledge Check

3

What are the two kinds of algorithms used in the model building step of the data science process?

A. Data cleaning and data preparation

B. Classification and regression

C. Optimization and testing

D. Data gathering and data interpretation

Knowledge Check

3

What are the two kinds of algorithms used in the model building step of the data science process?

A. Data cleaning and data preparation

B. Classification and regression

C. Optimization and testing

D. Data gathering and data interpretation

The correct answer is **B**

The algorithms used in this step are of two kinds: classification, which involves dividing data into different classes, and regression, which involves the prediction of a numeric value from the given data.

Thank You