

Chapter 6

Data and Analysis

Student Learning Outcomes:

By the end of this chapter, students will be able to :

- grasp the interdisciplinary nature of data science, understanding various data types, collection methods, and storage techniques. They will also demonstrate proficiency in data science, machine learning, and data visualization, effectively applying these skills to address real-world business challenges and communicating findings through advanced data visuals.

Introduction

In this chapter, students will learn the fundamentals of data and analysis and their importance in modern computing. They will explore the basics of data science and understand different types of data, including structured and unstructured data, along with methods of data collection and storage. The chapter introduces machine learning as a way for systems to learn from data and make decisions. The chapter explains how to evaluate model performance using measures like accuracy, precision, recall, and F1-score, and how to improve models through validation and tuning. Finally, the students will be introduced to common tools such as Excel, R, and Python used in data science and machine learning.

6.1 Introduction to Data Science

Data science is a field that focuses on collecting, analyzing, and using data to find useful information. It helps convert raw data into meaningful knowledge. Data science uses tools and methods to analyze large amounts of data. It supports decision-making in many areas of life. Data science also helps predict future trends. It is widely used in business, education, and healthcare. Modern technology depends heavily on data science.



TIDBIT

Many organizations use data science to support decision making. It helps understand trends and patterns hidden in data.

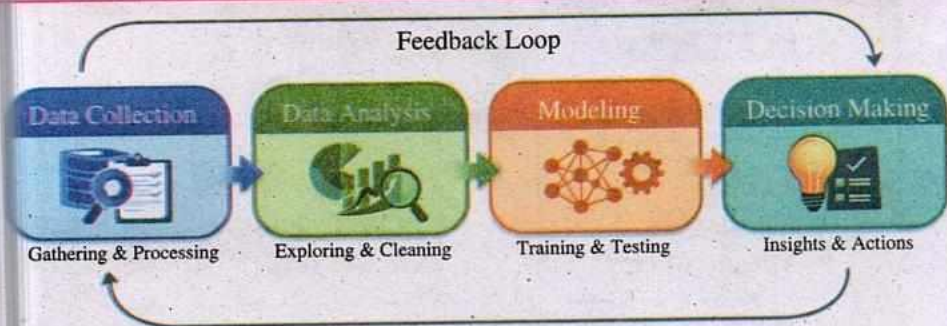


Figure 6.1: Data Science Work Flow

Data science means the study of data to discover patterns and useful information. It involves collecting data from different sources. The data is cleaned and organized before analysis. After analysis, the results are interpreted carefully. Data science helps answer questions using facts and data. It reduces guesswork in decision-making. The complete workflow of Data Science is shown in Figure 6.1. Data science turns data into knowledge. This knowledge helps solve real-world problems.

Data science combines knowledge from different subjects. It uses mathematics to work with numbers and calculations. Statistics helps in analyzing and summarizing data. Computer science provides tools and programming skills. Domain knowledge is needed to understand real problems. Communication skills help explain results clearly. Because of this combination, data science is called interdisciplinary. Each field supports the other. Together, they make data analysis more effective.

6.2 Understanding Data

Data is the foundation of Data Science. Data is a collection of raw facts and values to produce information. It can be in the form of numbers, text, images, audio or video. Data is produced every day through different activities. Understanding data is important before analysis. Different types of data need different handling methods. Proper data understanding helps in obtaining correct results. Poor data understanding can lead to errors. Data must be accurate and reliable to produce accurate results



TIDBIT

Data quality strongly affects final results. Poor data can lead to incorrect conclusions.

Types of Data: Structured and Unstructured

Structured data is organized in a clear format, such as tables or spreadsheets. It has rows and columns that are easy to understand. Examples include students' mark sheets and databases. Unstructured data does not follow a fixed format (as shown in Figure 6.2). Examples include emails, images, audio, and videos. Unstructured data is difficult to analyze directly. Special tools are needed to process it. Both types of data are important. Modern systems use both structured and unstructured data.

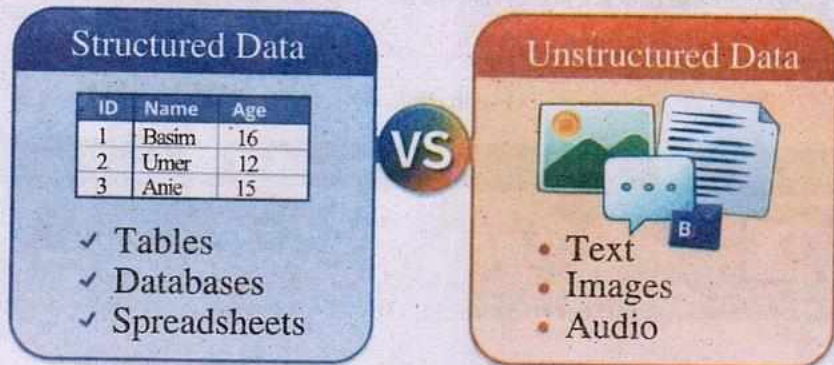


Figure 6.2: Types of Data



Data Sources and Data Collection Methods

Data can be collected from many sources. Common sources include surveys, interviews, forms, sensors, and online platforms (such as blogs, social websites, and others). Manual data collection involves human effort, whereas automatic data collection uses machines and software. Data can be collected in real time or stored over time. The method depends on the purpose of the analysis. Good data sources and accurate data collections improve analysis quality.

Data Storage and Management Concepts

Data storage refers to saving data for future use. Data can be stored in files, databases, or cloud systems. Proper storage helps protect data from loss. Data management involves organizing and maintaining data. It ensures data is easy to find and use. Security is an important part of data management. Backup systems help prevent data loss. Well-managed data improves efficiency. Data management supports long-term data use.

DO YOU KNOW?

Proper understanding of data type is very important before analysis of data.

6.3 Overview of Machine Learning

Machine learning is a subfield of Data Science that enables systems to improve from experience without explicit programming. It involves analyzing patterns in data to make predictions or decisions. Large datasets are used to train models, which adapt and improve over time as new data is introduced. This technique is widely applied in modern applications, supporting automation and intelligent systems.

Instead of predefined rules, machine learning allows systems to learn from examples, identifying patterns that guide predictions. As the system processes more data, it becomes more accurate and efficient, powering smart applications and reducing manual work.

Learning From Data

Learning from data means using past information to improve decisions. In this process, data is given as input to a learning model, which studies relationships in it. Errors are reduced through repeated training, and data becomes more accurate. Learning from data is continuous because the models update when new data is available. This process helps systems become smarter.

Difference Between Traditional Programming and Machine Learning

Traditional programming uses fixed rules written by programmers. The output is produced by following these rules. Machine learning uses data instead of fixed rules. The system learns rules automatically from data. In traditional programming, changes require new code. In machine learning, changes happen through new data. Machine learning handles complex problems better. It adapts to new situations. Figure 6.3 shows traditional programming vs machine learning.

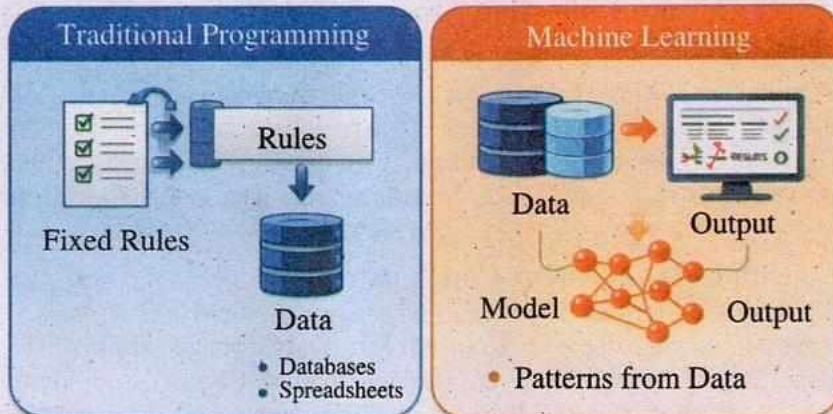


Figure 6.3: Traditional Programming vs Machine Learning

6.4 Machine Learning Mechanisms

Machine learning mechanisms describe how systems learn from data. Different mechanisms are used for different problems. Some methods use labeled data, while others do not (as shown in Figure 6.4).

LABELED DATA

Data with labels (outputs) that tell us what the data is.

Image	Label
	Apple
	Banana
	Orange
	Grapes

Used in supervised learning
(e.g., classification, regression)

UNLABELED DATA

Data without labels.
We don't know what the data is.

Image		
		
		
		
		

Used in unsupervised learning
(e.g., clustering, association)

Figure 6.4: Labeled vs Unlabeled Data

Certain methods are learned through feedback and experience. Each mechanism follows a different learning process. Choosing the correct mechanism is important. The quality of learning depends on the method used. These mechanisms support intelligent decision-making. Machine learning systems rely on proper mechanisms to work effectively. Some kinds of machine learning are discussed below:



Supervised Machine Learning

Supervised machine learning uses labeled data for training. Each data item has an input and a correct output. The system learns by comparing its output with the correct answer. Supervised learning needs large labelled datasets, such as emails labelled as spam or not spam, images labelled as cat or dog, or medical scans labelled as normal or abnormal. Its accuracy usually improves when the model is trained on more high-quality examples. It is widely used in real applications, such as face recognition, disease detection, speech recognition, fraud detection, and recommendation systems. This method is used for prediction tasks. Examples include email spam detection and exam result prediction.

Unsupervised Machine Learning

Unsupervised machine learning works with unlabeled data. The system finds patterns without guidance. It groups similar data together and makes clusters. This method is useful for discovering hidden structures. No correct output is provided. Learning depends on data similarity. Unsupervised learning helps understand data behavior. It is useful for data exploration. Examples include customer grouping and data clustering.

Reinforcement Machine Learning

Reinforcement machine learning learns through experience and mimics the human learning. The system interacts with an environment. It receives rewards or penalties for actions. Positive rewards encourage correct behavior. The system improves decision making over time. This method is used in robotics and games. Learning continues through trial and error. The goal is to maximize rewards. Examples of reinforcement learning include a robot learning to navigate a maze through rewards and penalties, AlphaGo improving its strategy by playing Go, an autonomous car optimizing its driving decisions, and a movie recommendation system adjusting its suggestions based on user feedback.

6.5 Applications of Machine Learning

Machine learning is used in many areas of modern life. It helps organizations analyze large amounts of data quickly, including Healthcare, Business, Education, and Daily Life (as shown in Figure 6.5). Many systems work automatically using learned patterns. It reduces human effort and saves time.

Use of Machine Learning in Healthcare

In healthcare, machine learning helps analyze medical data and supports disease detection and diagnosis. For example, it can help detect cancerous cells in medical images, identify early signs of Alzheimer's disease or dementia, and assist in the diagnosis of cardiovascular diseases. Doctors can use these results to plan better and more timely treatments.



CLASS ACTIVITY

Provide a simple dataset example, such as marks and grades. Students decide which data can be used for training and testing.

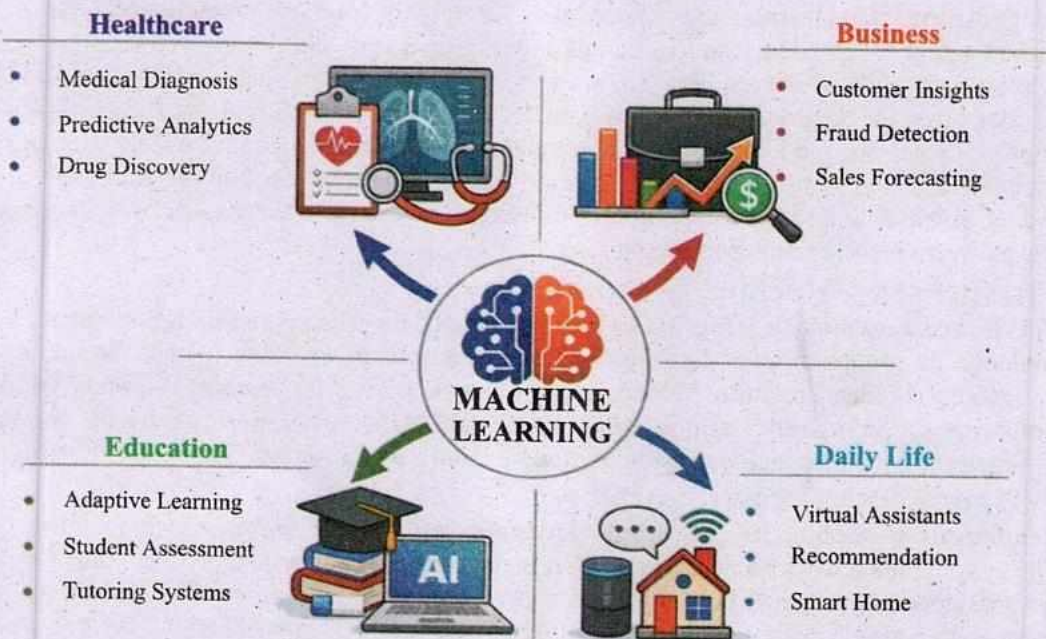


Figure 6.5: Applications of Machine Learning

Use of Machine Learning in Business

Businesses use machine learning to understand customer behavior. It helps predict sales and market trends. Businesses use it for inventory management. Examples include detecting fraud in financial systems, customer services through robots or chatbots, and suggesting products to customers. Marketing strategies become more effective using machine learning. Business decisions become data-driven. This improves profit and efficiency.

Use of Machine Learning In Education

In education, machine learning improves learning systems. It helps track students' performance to improve their grades. Personalized learning content is provided. Attendance and assessment systems become smarter. Learning outcomes improve.

Everyday Applications of Machine Learning

Machine learning is used in daily life activities. Mobile phones use it for face recognition. Voice assistants understand spoken commands. Social media shows personalized content. Email services filter spam messages. Navigation apps suggest routes. Online platforms recommend videos and music. Smart devices use machine learning and daily tasks become easier and faster.

6.6 Building a Machine Learning Model

Building a machine learning model involves several clear steps (as shown in Figure 6.6). A simple example of machine learning is student grade prediction. In this example, data about students is first collected, such as attendance, quiz marks, assignment scores, study hours, and



previous exam results. The data is then cleaned and prepared so that it can be used for learning. Important values, also called features, are selected because they help the model understand the factors that affect student performance. After this, the machine learning model is trained using the prepared data. Once trained, the model can predict a student's possible grade based on new input data. This helps teachers identify students who may need extra support before the final examination.



TIDBIT

Building a model involves selecting data and features carefully. Training data teaches the model patterns. Testing data checks performance. Feature selection improves results. A well-built model gives reliable predictions.

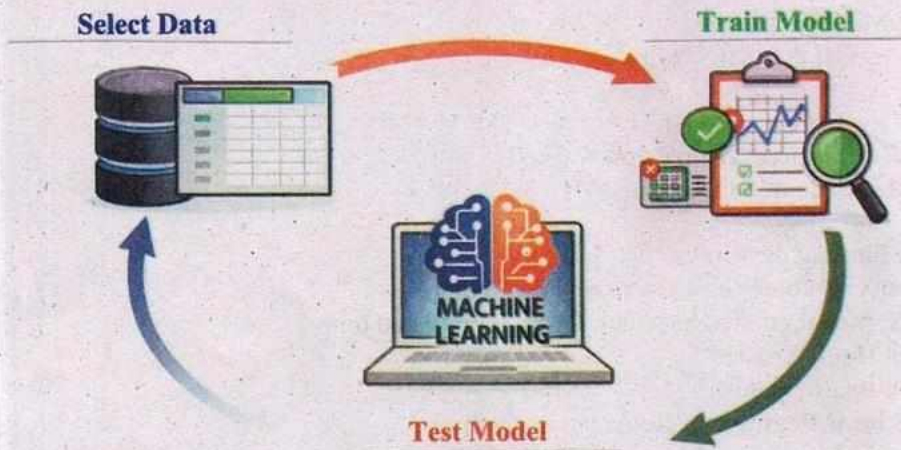


Figure 6.6: Machine Learning Model Building Process

Example:

Students' exam scores are often affected by the number of hours they study. The purpose of the following program is to build a simple linear regression model that predicts exam scores based on study hours. The dataset contains examples of hours studied and corresponding exam scores. The data is divided into training and testing sets, the model is trained using the training data, and its performance is evaluated using Mean Squared Error (MSE). Finally, the actual data points and the predicted regression line are plotted to show the relationship between study hours and exam scores.

```
# Import necessary libraries
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
```



```
# Example data (Hours studied vs. Exam scores)
X = np.array([[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]]) # Hours studied

y = np.array([1, 2, 2, 3, 3, 4, 5, 5, 6, 7]) # Exam scores

# Step 1: Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Step 2: Create and train the model
model = LinearRegression()
model.fit(X_train, y_train)

# Step 3: Make predictions
y_pred = model.predict(X_test)

# Step 4: Evaluate the model
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")

# Step 5: Plotting the results
plt.scatter(X, y, color='blue') # Actual data points
plt.plot(X, model.predict(X), color='red') # Predicted line
plt.xlabel('Hours Studied')
plt.ylabel('Exam Scores')
plt.title('Linear Regression: Hours vs. Scores')
plt.show()
```

- **Data Preparation:** The X values represent hours studied, and y values represent the corresponding exam scores (as shown in Figure 6.7).
- **Model Training:** A linear regression model is created and trained on the training data (X_train, y_train).
- **Testing:** The model is tested on the test data (X_test) and the mean squared error (MSE) is calculated to evaluate its performance. A lower MSE indicates better model performance because it means the predicted values are closer to the actual values.
- **Improvement:** As more data is gathered or adjusted, the model's accuracy improves.

Output:

Mean Squared Error: 0.1484

Model equation: Exam Score = $0.6552 \times \text{Hours Studied} + 0.1466$

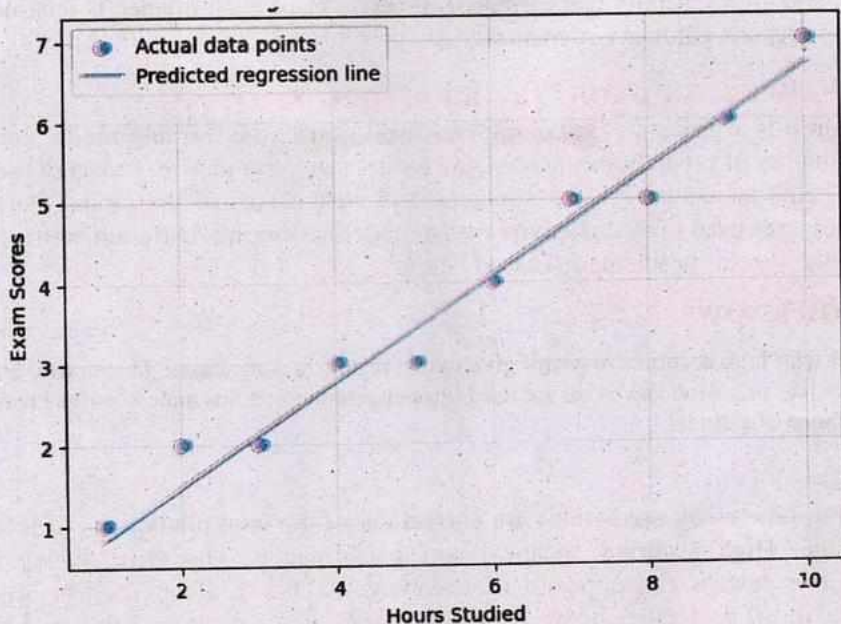


Figure 6.7: Linear Regression: Hours Studied vs Exam Scores

Selecting Features and Target Values

Features are input values used by the model. Target values are the outputs to be predicted. Correct feature selection improves learning. Unnecessary features can reduce accuracy. Features must be relevant to the problem. Target values must be clearly defined. Example: When predicting student marks, study hours would be the feature, and the target value would be the marks.

Feature Engineering Basics

Feature engineering means improving data for better learning. It includes selecting, modifying, and creating features. Raw data is often not ready for models. Data may need cleaning and scaling. Feature engineering improves model performance. It helps highlight important information. Good features produce better predictions. This step requires an understanding of data. Feature engineering strengthens the model. Example: In predicting student marks, feature engineering might involve cleaning raw data like study hours, scaling it, and creating new features like "study hours per day" to improve model performance and make better predictions.

Train-Test Split Of Data

Data is divided into training and testing sets. Training data teaches the model patterns. Testing data checks model performance as it uses unseen data. Results show real performance. This split prevents overfitting. Common splits include 80 percent training and 20 percent testing.

Building a Simple Predictive Model

A simple predictive model uses basic algorithms to learn from training data. It analyses the data to understand patterns and then makes predictions for new, unseen data. These models are easy



to understand and are especially helpful for beginners. Their performance is measured using test data and the results guide improvements.

6.7 Model Evaluation and Performance Metrics

Model evaluation is a process of checking how well a machine learning model performs. It measures the quality of predictions made by the model using test data that was not used during training. This helps in identifying the strengths and weaknesses of the model. Performance metrics are measures used to evaluate how well a model is working. Different metrics are used depending on the type of problem, discussed below:

? DO YOU KNOW?

A model with high accuracy may still give wrong results in some cases. That is why other measures like precision and recall are used. Proper evaluation helps understand the real performance of a model.

Accuracy and Error

Accuracy shows how many predictions are correct out of the total predictions. It is measured as a percentage. High accuracy means better performance. The error shows incorrect predictions. Error rate is the opposite of accuracy. Accuracy alone may be misleading sometimes. Balanced data gives better accuracy results. Error analysis helps find mistakes. Both measures are important for evaluation. The accuracy equation is:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- TP (True Positives): Number of correctly predicted positive instances.
- TN (True Negatives): Number of correctly predicted negative instances.
- FP (False Positives): Number of incorrectly predicted positive instances.
- FN (False Negatives): Number of incorrectly predicted negative instances.

Precision and Recall

Precision measures how many predicted positive results are actually correct. Recall measures how many actual positive cases are correctly identified. Precision focuses on correctness. Recall focuses on completeness. High precision reduces false results. High recall reduces missed cases.

Precision Equation:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- TP (True Positives): Correctly predicted positive instances.
- FP (False Positives): Incorrectly predicted positive instances.

Recall Equation:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- TP (True Positives): Correctly predicted positive instances.
- FN (False Negatives): Incorrectly predicted negative instances.



F1-Score

F1-score is the balance between precision and recall. It gives a single performance score. F-measure is useful when the data is unbalanced. It balances correctness and completeness. A higher F-measure means better performance. It avoids relying on accuracy alone. F1-score is commonly used in evaluation. It gives a fair comparison.

F1-Score Equation:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Precision: The proportion of positive predictions that are actually correct.
- Recall: The proportion of actual positive instances that are correctly identified.

Example: Suppose we are given the following values:

- TP (True Positives) = 40
- TN (True Negatives) = 50
- FP (False Positives) = 10
- FN (False Negatives) = 5

Now, let's calculate Precision, Recall, and F1-Score based on these values.

Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{40}{40 + 10} = \frac{40}{50} = 0.8$$

Recall:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{40}{40 + 5} = \frac{40}{45} = 0.8889$$

F1-Score:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.8 \times 0.8889}{0.8 + 0.8889} = 2 \times \frac{0.7111}{1.6889} = 0.8421$$

Final Results:

- Precision = 0.8
- Recall = 0.8889
- F1-Score = 0.8421

6.8 Validation and Model Improvement

Validation is used to check how well a model works on new data. It is done after training and testing to confirm model reliability. It prevents models from memorizing data. Validation supports fair evaluation. Model improvement uses feedback from validation results. Weak areas are identified and corrected. Validation is important before real use.



TIDBIT

Validation checks how a model works on new data. It prevents overfitting. Model improvement uses tuning and adjustments. Good validation increases reliability. Strong models perform well in real conditions.

Validation Data Set Preparation

A validation dataset is separate from training and testing data that is used to check model performance. Validation data must be clean and accurate to represent real conditions. The model does not learn from validation data. Validation helps detect overfitting and improves evaluation. Balanced data gives better results. This step increases model confidence.

Model Robustness and Reliability

Robustness means a model performs well in different situations. A reliable model gives consistent results. Robust models handle noise in data. They do not fail easily with small changes. Reliability is important for real-world use. Testing under different conditions checks robustness. Strong models maintain accuracy. Reliability builds trust in systems. Robust models are more useful.

Improving Models Using Hyperparameter Tuning

Hyperparameter tuning adjusts model settings. These settings control learning behavior. It helps improve accuracy. It reduces errors in predictions. Parameters are tested with different values. Validation data guides tuning. Proper tuning improves performance. Poor tuning reduces effectiveness.

Simple Analogy: Adjusting the Volume of a Speaker

Hyperparameter tuning is like adjusting the volume of a speaker. If the volume is too high, the sound becomes distorted and unclear. If the volume is too low, the sound cannot be heard properly. However, when the volume is set at the right level, the sound becomes clear and balanced.

Similarly, in machine learning, if hyperparameters are not set properly, the model may perform poorly. The correct hyperparameter values help the model learn better and produce more accurate results.

6.9 Predictive Modeling and Causality

Predictive modeling uses data to predict future outcomes. Machine learning models analyze past data patterns. These patterns help estimate what may happen next. Predictive models are used in many fields. They support planning and decision making. Predictions are based on probability. They do not always explain reasons. Understanding causality is also important. Both concepts are used together in analysis.



CLASS ACTIVITY

Give examples like weather prediction. Ask students whether prediction explains the cause. Discuss the difference.



Predictive Outcomes in Machine Learning

Predictive outcomes are results produced by a trained model. The model uses past data to make predictions. These outcomes estimate future values or events. Prediction accuracy depends on data quality. Good training improves predictions. For example, businesses use predictions for forecasting the market or product selling trends. Healthcare uses predictions for diagnosis. Predictions support better decisions.

Understanding Causality

Causality explains why an event happens. It shows a cause-and-effect relationship. One factor directly influences another. Causality is important in scientific studies to understand real reasons behind events. Data alone may not show causality but experiments often identify causes. Causal understanding improves decisions and avoids false assumptions.

Difference Between Prediction and Causation

Prediction estimates what may happen based on historical data. It focuses on forecasting outcomes, where the model uses past information to estimate future values or events. For example, a model may predict a student's final grade by analysing previous marks, attendance, and study hours. Similarly, a weather model may predict tomorrow's temperature by using past weather records. A model can make accurate predictions without necessarily understanding the underlying causes. This makes prediction useful in areas such as sales forecasting, stock trend analysis, disease risk prediction, and student performance estimation.

Causation, on the other hand, explains why something happens by identifying cause-and-effect relationships between variables. It seeks to uncover the reasons behind events and how one factor directly influences another. For example, if regular study hours directly improve exam performance, then study time may be considered a cause of better grades. Similarly, if smoking directly increases the risk of lung disease, this shows a causal relationship. Understanding causality is important for decision-making because it helps avoid false assumptions and provides deeper insight into real-world problems.

6.10 Machine Learning Tools and Platforms

Machine learning uses different tools for data analysis and modeling. Tools help collect, process, and analyze data. Some tools are simple and beginner-friendly, while others support advanced machine learning tasks. Choosing the right tool is important. Each tool has specific strengths. Tools improve efficiency and accuracy and build strong foundations. These platforms support modern data science work. Machine learning tools and platforms include:

- Excel for data analysis
- R for data science
- Python and Jupyter notebooks for machine learning

Using Excel for Data Analysis

Excel is widely used for basic data analysis by organizing the data in rows and columns. Excel supports sorting, filtering, and data visualization using charts. Charts help visualize data. Simple formulas perform calculations. Excel is easy to learn and is useful for small datasets. Beginners often start with Excel, and it supports early data understanding.



Introduction to R for Data Science

R is a programming language for data analysis that is widely used in research. It provides strong graphical tools to visualize data. R handles large datasets efficiently. Many packages support data science tasks. Learning R improves analytical skills that support machine learning basics.

Python and Jupyter Notebooks for Machine Learning

Python is a popular programming language for machine learning. It has many useful libraries such as Pandas, NumPy, and others. Python supports data analysis and modeling. Jupyter notebooks provide an interactive environment. Code and results appear together on Jupyter notebook. This helps learning and testing the code on an interactive Integrated Development Environment (IDE). Python supports advanced machine learning and is widely used in industry.



Summary

Data Science

Data science is the process of collecting, organising, analysing, and interpreting data to find useful information and support decision-making.

Structured and Unstructured Data

Structured data is arranged in a fixed format, such as tables with rows and columns. Unstructured data does not follow a fixed format, such as images, videos, emails, and audio files.

Machine Learning

Machine learning is a branch of artificial intelligence in which computers learn patterns from data and use them to make predictions or decisions.

Supervised Learning

Supervised learning is a machine learning method in which the model learns from labelled data, such as study hours and known exam scores.

Unsupervised Learning

Unsupervised learning is a machine learning method in which the model finds hidden patterns or groups in data without labelled outputs.

Reinforcement Learning

Reinforcement learning is a learning method in which an agent learns by taking actions and receiving rewards or penalties.

Feature and Target Value

A feature is an input value used by a model, such as study hours or attendance. A target value is the output the model tries to predict, such as an exam score.

Train-Test Split

Train-test split is the process of dividing data into training data for teaching the model and testing data for checking its performance.

Predictive Model

A predictive model uses past data to estimate future values or outcomes.

Model Evaluation

Model evaluation is the process of checking how well a machine learning model performs using measures such as accuracy, error, precision, recall, and F1-score.

Parameter Tuning

Parameter tuning is the process of adjusting model settings to improve its performance.

Prediction

Prediction is the process of estimating a future value or event based on past data.

Causality

Causality explains cause-and-effect relationships between variables and helps identify why something happens.

Machine Learning Tools

Machine learning tools are software platforms used to analyse data and build models. Common examples include Excel, R, Python, and Jupyter Notebook.



EXERCISE

Multiple Choice Questions

- Data science is mainly concerned with:
(a) Designing hardware (b) Extracting useful information from data
(c) Writing only code (d) Creating websites
- Data organised in rows and columns is called:
(a) Unstructured data (b) Structured data
(c) Raw data (d) Random data
- An example of unstructured data is:
(a) Tables (b) Spreadsheets
(c) Images and videos (d) Databases
- Machine learning allows systems to:
(a) Work without data (b) Learn from data and improve performance
(c) Only store data (d) Replace databases
- The type of machine learning that uses labelled data is:
(a) Unsupervised learning (b) Reinforcement learning
(c) Supervised learning (d) Rule-based learning
- The machine learning method that learns through rewards and penalties is:
(a) Supervised learning (b) Unsupervised learning
(c) Reinforcement learning (d) Rule-based learning
- Feature engineering is used to:
(a) Delete data (b) Improve data for better model performance
(c) Store data (d) Visualise data
- The metric that measures the overall correctness of a model is:
(a) Precision (b) Recall
(c) Accuracy (d) F1 score
- Train-test split is used to:
(a) Delete data (b) Divide data for training and testing
(c) Store data (d) Visualise data
- A commonly used tool for machine learning and data analysis is:
(a) MS Word (b) Python
(c) PowerPoint (d) Notepad



Short Questions

- What is data science?
- What are the two main types of data?
- What is the difference between structured and unstructured data?
- Name two common data collection methods.
- What is machine learning?



6. What is supervised learning?
7. What is the purpose of feature selection in machine learning?
8. What does accuracy measure in a model?
9. What is the difference between prediction and causality?
10. Name any one tool used for machine learning.



Long Questions

1. Explain the concept of data science and discuss its importance in modern applications.
2. Describe different types of data and explain various data sources and collection methods.
3. Explain the concept of machine learning and describe its different mechanisms, including supervised, unsupervised, reinforcement, and rule-based learning.
4. Discuss the applications of machine learning in healthcare, business, education, and daily life.
5. Explain the steps involved in building a machine learning model, including feature selection, feature engineering, and train-test split.
6. Describe model evaluation techniques and explain accuracy, precision, recall, and F1-score.
7. Explain validation in machine learning and discuss how models can be improved using parameter tuning.
8. Discuss predictive modeling and causality, and explain the importance of model interpretation in real-world use cases.



Answer Key

1. B
2. B
3. C
4. B
5. C
6. C
7. B
8. C
9. B
10. B