



CHAPTER-6

DATA AND ANALYSIS

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.

6.1 Introduction to Data Science

Ex.Q.1. Explain the concept of data science and discuss its importance in modern applications.

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 different tools and methods to analyze large amounts of data. It supports decision-making in many areas of life. It also helps predict future trends. Data science is widely used in business, education, and healthcare. Modern technology depends heavily on data science.

Data science means the study of data to discover patterns and useful information. Data science involves collecting data from different sources. The collected data is cleaned and organized before analysis. After analysis, the results are carefully interpreted. Data science helps answer questions by using facts and data instead of guesswork. It turns data into useful knowledge that helps solve real-world problems.

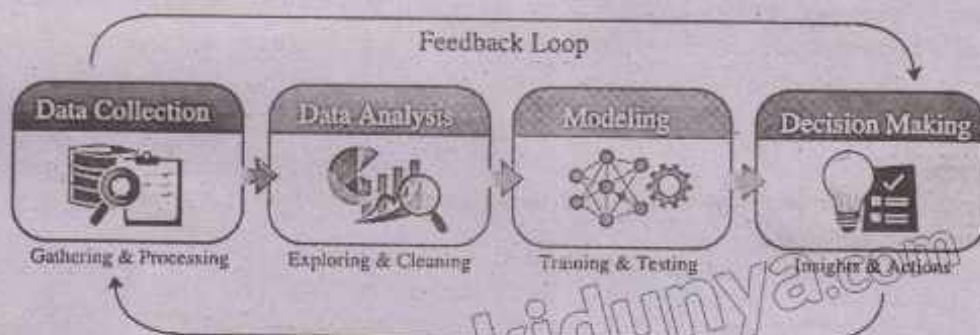


Figure 6.1 Data Science Work Flow

Data Science as an Interdisciplinary Field

Data science combines knowledge from different subjects. It uses mathematics to perform calculations and work with numbers. Statistics helps analyze and summarize data. Computer science provides programming skills and tools for data analysis. Domain knowledge helps understand real-world problems. Communication skills help explain the results clearly.

Short Question

Why is Data Science called an Interdisciplinary Field?

Because data science combines different fields, it is called an interdisciplinary field. Each subject supports the others and makes data analysis more effective.

Importance of Data Science in Modern Applications

i. Supports Decision-Making

Data science helps people and organizations make better decisions using data.

ii. Predicts Future Trends

It helps predict future events and trends from historical data.

iii. Solves Real-World Problems

It turns data into useful knowledge that helps solve real-world problems.

iv. Used in Many Fields

Data science is widely used in business, education, and healthcare.

v. Improves Data Analysis

It uses different tools and methods to analyze large amounts of data.

vi. Supports Modern Technology

Modern technology depends heavily on data science.

Short Question

Why is Data Science important in modern applications?



TIDBIT

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

6.2 Understanding Data

Ex.Q.2 What is Data? Describe different types of data and explain various data sources and collection methods.

Data

Data is a collection of **raw facts and values** that are used to produce information. Data is the foundation of Data Science. Data can be in the form of numbers, text, images, audio, or video. It is produced every day through different activities.

Understanding data is important before analysis. Different types of data need different handling methods. Proper understanding of data helps produce correct results. Poor understanding of data can lead to errors. Data must be accurate and reliable to produce accurate results.

Types of Data

1. Structured Data

Structured data is organized in a clear format, such as **tables or spreadsheets**. It is arranged in rows and columns. It is easy to understand and analyze.

Examples: Students' mark sheets and databases.

2. Unstructured Data

Unstructured data does **not follow a fixed format**. It includes emails, images, audio, and videos. This type of data is difficult to analyze directly. Special tools are needed to process unstructured data.

Examples: Emails, images and videos. Both structured and unstructured data are important. Modern systems use both types of data for analysis and decision-making.

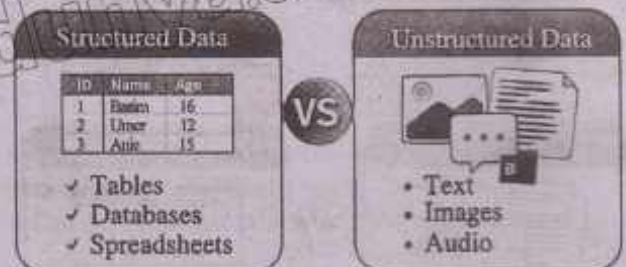


Figure 6.2: Types of Data

Data Sources

Data can be collected from different sources depending on the purpose of the analysis. Common data sources include:

- Surveys
- Interviews
- Questionnaires and forms
- Sensors and IoT devices
- Online platforms such as blogs and social media websites

Using reliable data sources improves the accuracy of analysis.

Data Collection Methods

Data collection methods are the ways used to gather data. Common methods include:

- **Manual Data Collection:** Data is collected by people using surveys, interviews, or forms.
- **Automatic Data Collection:** Data is collected automatically by machines, sensors, and software.
- **Real-Time Data Collection:** Data is collected instantly as events happen.
- **Stored Data Collection:** Data is collected and stored over a period of time for future analysis.

The method of data collection depends on the purpose of the analysis. Accurate data collection helps produce reliable and meaningful results.

Short Question
Describe two data collection methods?

Q. What are data storage and data management?

Data Storage and Management

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



TIDBIT

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

? DO YOU KNOW?

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



Additional Multiple Choice Questions

1. Data science is a field that focuses on collecting, analyzing, and using data to find:
(a) Useful information (b) Computer viruses (c) Programming errors (d) Network cables
2. Data is a collection of raw facts and values used to produce:
(a) Database (b) Software (c) Information (d) Algorithm
3. Why is data science called an interdisciplinary field?
(a) It combines knowledge from different subjects. (b) It uses only mathematics.
(c) It works only with computers. (d) It stores only databases.
4. Which of the following is an importance of data science?
(a) Predicts future trends (b) Increases file size (c) Removes hardware (d) Installs software
5. Structured data is organized in a clear format, such as:
(a) Images (b) Audio files (c) Tables or spreadsheets (d) Videos

6. Data that does not follow a fixed format is called:
 (a) Structured data (b) Numerical data (c) Database (d) Unstructured data
7. Which of the following is a common source of data collection?
 (a) Surveys (b) Compiler (c) Firewall (d) Router
8. Why is data quality important in data science?
 (a) Poor data can lead to incorrect conclusions. (c) It removes programming errors.
 (b) It increases internet speed. (d) It reduces computer memory.

Answers:

1 A 2 C 3 A 4 A 5 C 6 D 7 A 8 A

6.3 Overview of Machine Learning & 6.4 ML Mechanisms

Ex.Q.3: Explain the concept of machine learning and describe its different mechanisms, including supervised, unsupervised, reinforcement, and rule-based learning.

Machine Learning

Machine learning is a subfield of Data Science that enables systems to **improve from experience** without explicit programming. It works by analyzing patterns in data to make predictions or decisions. Large datasets are used to train machine learning models.

As new data is added, the models learn and improve over time. Machine learning is widely used in modern applications. It supports automation and intelligent systems.

Instead of using only predefined rules, machine learning learns from examples. It finds patterns in data and uses them to make predictions. As the system processes more data, it becomes more accurate and efficient. This reduces manual work and helps create smart applications.

Learning From Data

Learning from data means using past information to improve future decisions. Data is given as input to a learning model. The model studies the relationships in the data. During training, errors are reduced again and again. As new data becomes available, the model updates itself and improves its performance. This continuous learning process helps systems become smarter.

Machine Learning Mechanisms

Machine learning mechanisms explain how systems learn from data. Different mechanisms are used for different problems. Some methods use labeled data, while others use unlabeled data. Some methods learn through feedback and experience. Each mechanism follows a different learning process. Choosing the correct mechanism is important because the quality of learning depends on the method used. These mechanisms help systems make intelligent decisions and work effectively.

1. 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. This method requires large labeled datasets, such as emails labeled as spam or not spam, images labeled as cat or dog, and medical scans labeled as normal or abnormal. The accuracy of the model improves when it is trained with more high-quality data. Supervised machine learning is widely used in face recognition, disease detection, speech recognition, fraud detection, and recommendation systems. It is mainly used for prediction tasks.

Examples: Email spam detection and exam result prediction.

LABELED DATA		UNLABELED DATA		
Data with labels (outputs) that tell us what the data is.		Data without labels. We don't know what the data is.		
Image	Label	Image		
	Apple			
	Banana			
	Orange			
	Grapes			
Used in supervised learning (e.g., classification, regression)		Used in unsupervised learning (e.g., clustering, association)		

Figure 6.4: Labeled vs Unlabeled Data

2. Unsupervised Machine Learning

Unsupervised machine learning **works with unlabeled data**. The system finds patterns without guidance. It groups similar data together and makes clusters. No correct output is provided during training.

This method is useful for discovering hidden structures in data. Learning depends on the similarity between data items. Unsupervised machine learning helps understand data behavior and explore data.

Examples: Customer grouping and data clustering.

Short Question

What is unsupervised machine learning?

3. 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 actions while penalties discourage incorrect actions. The system improves its decisions through trial and error. The goal is to receive the highest possible rewards. This method is used in robotics and games.

Short Question

What is reinforcement machine learning?

Examples:

- A robot learning to move through a maze
- AlphaGo improving its game strategy
- A self-driving car improving its driving decisions
- A movie recommendation system changing its suggestions based on user feedback.

4. Rule-Based Learning

Rule-based learning is a machine learning approach in which a system **uses a set of rules** to make decisions or predictions. These rules are written in the form of **IF-THEN statements**. It provides clear and consistent decisions. It is useful for problems with fixed conditions.

Short Question

What is rule based learning?

Example:

- IF a student's marks are greater than 50, THEN the student is successful.
- IF an email contains suspicious words, THEN it is classified as spam.

Q. Differentiate between traditional programming and machine learning.

Difference Between Traditional Programming and Machine Learning	
Traditional Programming	Machine Learning
Traditional programming uses fixed rules written by programmers.	Machine learning uses data instead of fixed rules.
The program follows these rules to produce the required output.	The system learns the rules automatically from the data.
If the program needs to change, new code must be written.	The system improves when new data is provided without changing the program code.
Programmers manually write and update the rules.	The system automatically learns and improves from data.
It is suitable for problems with clearly defined rules.	It is better at solving complex problems.
It is less flexible when requirements change.	It can adapt to new situations.

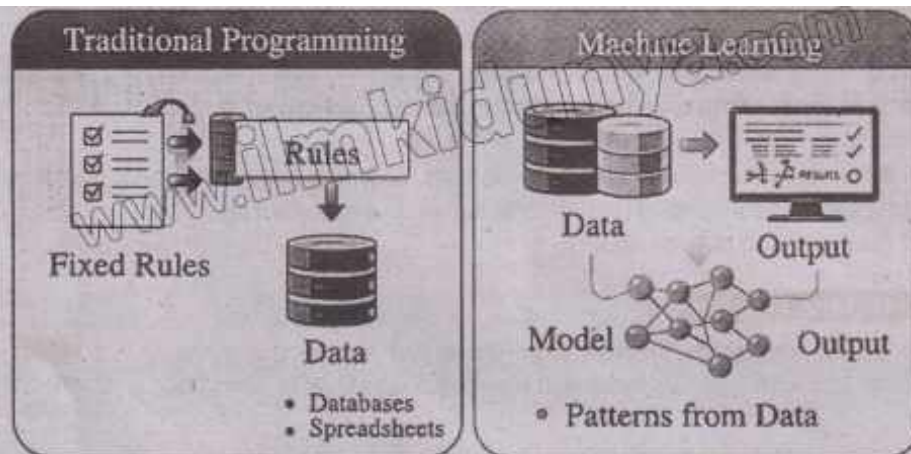


Figure 6.3: Traditional Programming vs Machine Learning

6.5 Applications of Machine Learning

Ex.Q.4. Discuss the applications of machine learning in healthcare, business, education, and daily life.

Applications of Machine Learning

Machine learning is used in many areas of modern life. It helps organizations analyze large amounts of data quickly. It is widely used in healthcare, business, education, and daily life. Many systems work automatically by learning patterns from data. This reduces human effort and saves time.

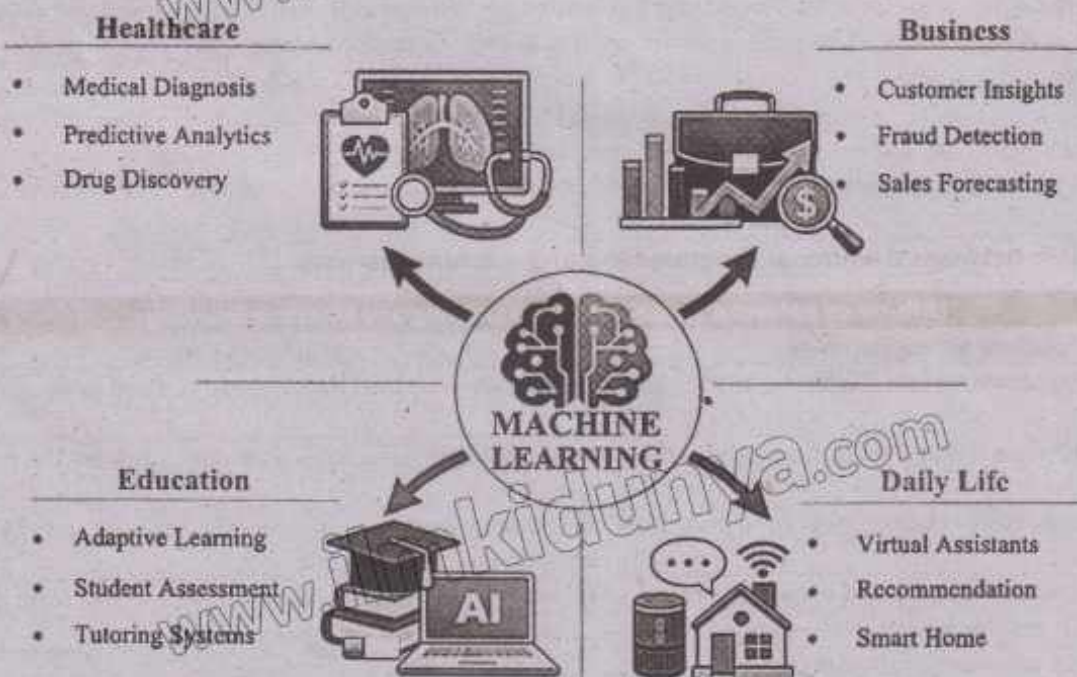


Figure 6.5: Applications of Machine Learning

1. Use of Machine Learning in Healthcare

In healthcare, machine learning helps analyze medical data. It supports disease detection and diagnosis. It can detect cancer cells in medical images. It can identify early signs of Alzheimer's disease and dementia.

It also helps diagnose cardiovascular diseases. Doctors use these results to provide better and timely treatment.

Short Question
Explain use of machine learning in healthcare.

2. Use of Machine Learning in Business

Businesses use machine learning to understand customer behavior. It helps predict sales and market trends. It is also used for inventory management. Machine learning can detect fraud in financial systems. It is used in customer service through robots and chatbots. It also recommends products to customers. This makes marketing strategies more effective. Business decisions become data-driven, which improves profit and efficiency.

3. Use of Machine Learning in Education

In education, machine learning improves learning systems. It helps track students' performance to improve their grades. It provides personalized learning content for students. It also makes attendance and assessment systems smarter. This improves learning outcomes.

4. Everyday Applications of Machine Learning

Machine learning is used in many daily life activities. Mobile phones use it for face recognition. Voice assistants understand spoken commands. Social media platforms show personalized content.

Email services filter spam messages. Navigation apps suggest the best routes. Online platforms recommend videos and music. Smart devices use machine learning to make daily tasks easier and faster.

Short Question
Describe use of machine learning in everyday applications.



CLASS ACTIVITY

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

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

Training Data		
Use the first 8 records to train the ML model.		
Student	Marks	Grade
Ali	92	A
Sara	85	B
Ahmed	78	C
Ayesha	95	A
Bilal	67	D
Fatima	88	B
Usman	73	C
Zainab	81	B

Testing Data		
Use the last 2 records to test the ML model.		
Student	Marks	Grade
Hamza	60	D
Noor	98	A

Conclusion

- **Training Data:** First 8 records are used to teach the ML model.
- **Testing Data:** Last 2 records are used to check how accurately the model predicts grades for new students.



Additional Multiple Choice Questions

1. Machine learning is a subfield of:
(a) Computer Networks (b) Data Science (c) Cyber Security (d) Database Systems
2. Machine learning enables systems to improve from experience without:
(a) Internet connection (b) User login (c) Explicit programming (d) Cloud storage
3. Instead of using only predefined rules, machine learning learns from:
(a) Examples (b) Errors (c) Hardware (d) Programs
4. In supervised machine learning, each data item has:
(a) Two databases (b) A network cable (c) An IP address (d) An input and a correct output
5. Which type of machine learning works with unlabeled data?
(a) Supervised learning (b) Rule-based learning
(c) Unsupervised learning (d) Reinforcement learning
6. Reinforcement machine learning learns through:
(a) Fixed rules (b) Experience (c) Tables (d) Images
7. Which machine learning mechanism receives rewards or penalties for actions?
(a) Rule-based learning (b) Reinforcement learning
(c) Supervised learning (d) Unsupervised learning
8. Machine learning is widely used in healthcare to:
(a) Detect diseases (b) Design websites (c) Repair hardware (d) Install software
9. Which of the following is an everyday application of machine learning?
(a) Face recognition (b) Cable testing (c) Disk formatting (d) File compression
10. Why do businesses use machine learning?
(a) To understand customer behavior and predict sales. (c) To manufacture computers.
(b) To replace all employees. (d) To increase electricity supply.

Answers:

1	B	2	C	3	A	4	D	5	C	6	B	7	B	8	A	9	A	10	A
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---

6.6 Building a Machine Learning Model

Ex.Q.5. Explain the steps involved in building a machine learning model, including feature selection, feature engineering, and train-test split.

Building a machine learning model is the process of preparing data, training a model, and using it to make predictions.

Steps of Building a Machine Learning Model

Building a machine learning model involves several steps. Each step helps the model learn from data and make accurate predictions.

i. Collect the Required Data

The first step is to collect the data needed for the problem. The data should be accurate, complete, and relevant.

ii. Clean and Prepare the Data

The collected data is cleaned by removing errors, duplicate records, and missing values. The data is then prepared in a suitable format for the model.

iii. Select the Important Features

Features are the data values used by the model for learning. Only the most useful features are selected to improve the model's performance.

iv. Train the Machine Learning Model

The model is trained using the prepared data. During training, the model learns the relationship between the input data and the output.

v. Make Predictions Using New Data

After training, the model is given new data to make predictions.

vi. Evaluate the Model by Checking Its Performance

The model's performance is checked to see how accurate its predictions are. If needed, the model is improved by using better data or adjusting its settings.



Example

A school wants to predict students' exam scores. It collects data such as students' study hours and previous exam scores. The collected data is cleaned by removing errors and missing values. It is then prepared for training. The school selects **study hours** as the main feature to predict students' exam scores. The machine learning model is trained using the prepared data to learn the relationship between study hours and exam scores. The trained model predicts the expected exam score of a student based on the number of hours the student studies. The predicted exam scores are compared with the actual exam scores to check the model's accuracy. If needed, the model is improved.

Result:

This helps teachers identify students who may need extra help before the examination.

Example:

```
# 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:.4f}")
print(f"Model equation: Exam Score = {model.coef_[0]:.4f} x Hours Studied + {model.intercept_:.4f}")

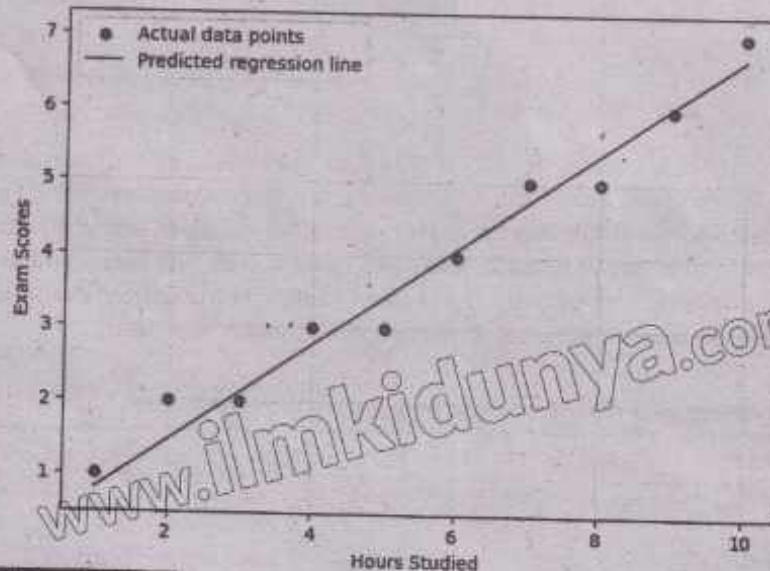
# Step 5: Plotting the results
plt.scatter(X, y, color="blue") # Actual data points
plt.plot(X, model.predict(X), color="red") # Predicted line
plt.xlabel("Hours Studied")
plt.ylabel("Exam Scores")
plt.title("Linear Regression: Hours vs. Scores")
plt.show()
```

- **Data Preparation:** The X values show the hours studied, and the y values show the exam scores.
- **Model Training:** A linear regression model is created and trained using the training data (X_{train}, y_{train}).
- **Testing:** The model is tested using the test data (X_{test}). The Mean Squared Error (MSE) is calculated to check the model's performance. A lower MSE means the predictions are closer to the actual values.
- **Improvement:** The model becomes more accurate when more data is added or the training data is improved.

Output:

Mean Squared Error: 0.1484

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



Selecting Features and Target Values

Features are the input values used by a machine learning model. Target values are the output values that the model predicts. Choosing the correct features helps the model learn better, while unnecessary features can reduce its accuracy. The features should be related to the problem, and the target value should be clearly defined. For example, when predicting student marks, study hours are the feature, and marks are the target value.

Feature Engineering Basics

Feature engineering means preparing and improving data so that a machine learning model can learn better. It includes selecting, modifying, and creating features. Raw data is often not ready for learning, so it may need cleaning to remove errors and scaling to make the values similar. Feature engineering helps the model focus on important information, improves its performance, and produces better predictions. It requires a good understanding of the data. For example, when predicting student marks, the data can be cleaned, study hours can be scaled, and a new feature such as study hours per day can be created to improve the model's predictions.

Train-Test Split of Data

In machine learning, the data is divided into training data and testing data. The training data is used to teach the model and help it learn patterns. The testing data is then used to check how well the model works on unseen data. This shows the model's real performance. Dividing the data into training and testing sets also helps prevent overfitting. A common split is 80% training data and 20% testing data.

Building a Simple Predictive Model

A simple predictive model uses a basic algorithm to learn from the training data. It studies the patterns in the data and then predicts the output for new, unseen data. These models are easy to understand and are useful for beginners. After the model is tested with the testing data, its performance is measured. The test results help improve the model and make its predictions more accurate.



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.

6.7 Model Evaluation and Performance Metrics

Ex.Q.6 Describe model evaluation techniques and explain accuracy, precision, recall, and F1-score.

Model Evaluation

Model evaluation is the process of **checking how well** a machine learning model performs. It measures the quality of the model's predictions by using **test data** that was not used during training. Model evaluation helps to identify the strengths and weaknesses of the model.

Short Question

What is model evaluation?

Performance Metrics

Performance metrics are measures used to check how well a model is working. Different performance metrics are used for different types of machine learning problems.

1. Accuracy and Error

Accuracy shows how many predictions are correct out of the total predictions. It is measured as a percentage. A high accuracy means the model is performing well.

Error shows the number of incorrect predictions made by the model. The **error rate** is the opposite of accuracy. Sometimes, accuracy alone can give a misleading result, especially when the data is not balanced. **Balanced data** gives more reliable accuracy results. **Error analysis** helps find the mistakes made by the model. Therefore, both accuracy and error are important for evaluating a model.

Accuracy Equation:

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

Where:

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

2. Precision and Recall

Precision measures how many predicted positive results are actually correct. It focuses on the correctness of positive predictions. **Recall** measures how many actual positive cases are correctly identified by the model. It focuses on finding all positive cases. High precision reduces false positive results, while high recall reduces missed positive cases.

Precision Equation:

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

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

Recall Equation:

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

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

Short Question

What are Precision and Recall performance metrics?

3. F1-Score

F1-Score is the balance between **precision** and **recall**. It gives a single score to measure the overall performance of a model. It is especially useful when the data is unbalanced. F1-Score balances correctness and completeness. A higher F1-Score means better model performance. It helps avoid relying only on accuracy and gives a fair comparison between models.

F1-Score Equation:

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

- **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 calculate Precision, Recall, and F1-Score based on these values.

- Precision: $Precision = \frac{TP}{TP + FP} = \frac{40}{40 + 10} = \frac{40}{50} = 0.8$
- Recall: $Recall = \frac{TP}{TP + FN} = \frac{40}{40 + 5} = \frac{40}{45} = 0.8889$
- F1-Score: $F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.8 \times 0.8889}{0.8 + 0.8889} = 0.8421$

Final Results:

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

Short Question

What is the F1-score performance metric?



TIDBIT

- 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.



Additional Multiple Choice Questions

1. Building a machine learning model is the process of preparing data, training a model, and using it to make:
 - (a) Predictions
 - (b) Programs
 - (c) Networks
 - (d) Databases
2. Features are the:
 - (a) Output values
 - (b) Input values
 - (c) Error values
 - (d) Network values
3. Which of the following is the target value when predicting student marks?
 - (a) Study hours
 - (b) Attendance
 - (c) Marks
 - (d) Quiz marks
4. Feature engineering means:
 - (a) Preparing and improving data
 - (b) Creating computer networks
 - (c) Installing software
 - (d) Writing web pages
5. In machine learning, the data is divided into:
 - (a) Files and folders
 - (b) Input and output
 - (c) Hardware and software
 - (d) Training data and testing data
6. Which data is used to teach the machine learning model?
 - (a) Testing data
 - (b) Training data
 - (c) Target values
 - (d) Error data
7. Why is testing data used?
 - (a) To teach the model
 - (b) To clean the data
 - (c) To check how well the model works on unseen data.
 - (d) To create new features.

8. **Model evaluation is the process of:**
 (a) Checking how well a machine learning model performs. (b) Creating new datasets.
 (c) Installing Python modules. (d) Drawing graphs.
9. **Which performance metric shows how many predictions are correct out of the total predictions?**
 (a) Precision (b) Recall (c) F1-Score (d) Accuracy
10. **Which performance metric is the balance between precision and recall?**
 (a) Accuracy (b) Error (c) F1-Score (d) Training data

Answers:

1	A	2	B	3	C	4	A	5	D	6	B	7	C	8	A	9	D	10	C
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	----	---

6.8 Validation and Model Improvement

Ex.Q.7. What is validation? Explain validation in machine learning and discuss how models can be improved using parameter tuning.

Validation

Validation is the process of checking **how well** a machine learning model works on **new data**. It is done after training and testing to confirm that the model is reliable. Validation makes sure that the model has learned patterns instead of only memorizing the training data.

It supports a fair evaluation of the model. The results of validation help identify weak areas in the model. These weak areas are corrected to improve performance. Model improvement is done by making adjustments and tuning the model. Good validation increases the reliability of the model and helps it perform well in real-world conditions.

Validation Data Set Preparation

A validation data set is a separate set of data used to check the performance of a machine learning model. It is different from the **training data** and **testing data**. The model does not learn from the validation data. The validation data should be clean, accurate, and should represent real-world conditions.

It helps to detect **overfitting** and improves the evaluation of the model. **Balanced data** gives better validation results. A good validation data set increases confidence in the model.

Model Robustness and Reliability

Robustness means that a machine learning model performs well in different situations.

A reliable model gives consistent results every time. Robust models handle noise in data. They do not fail easily with small changes. **Reliability** is important for real-world use. A model is tested under different conditions to check its robustness.

A strong model maintains its accuracy even when the data changes slightly. Reliable models are trusted because they produce consistent and accurate results.

Improving Models Using Hyperparameter Tuning

Hyperparameter tuning is the process of adjusting the **hyperparameters** of a machine learning model. These settings control how the model learns. The correct hyperparameter values improve the model's accuracy and reduce prediction errors. Different values are tested to find the best settings. **Validation data** is used to check which settings give the best performance. Proper tuning improves the model, while poor tuning reduces its 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.

Short Question

What is model validation in machine learning?

Short Question

What is model robustness and reliability in model validation?

6.9 Predictive Modeling and Causality

E.T.Q.8. Discuss predictive modeling and causality, and explain the importance of model interpretation in real-world use cases.

Predictive modeling

Predictive modeling is the process of using 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.

So they are not always correct. Predictive modeling tells us **what may happen**, but it does not always explain why it happens.

Therefore, understanding **causality** is also important. Both predictive modeling and causality are used together in data analysis.

Example: An online shopping website uses customers' previous purchases to predict which products they may buy next.

Predictive Outcomes in Machine Learning

Predictive outcomes are the results produced by a trained machine learning model. The model uses past data to make predictions. The accuracy of these predictions depends on the **quality of the data** used for training. Better training data gives more accurate predictions. Predictive outcomes help people to make better decisions in different fields. For example, businesses use predictions to forecast **market trends** and **product sales**. In healthcare, predictions help doctors in **diagnosis**.

Causality

Causality explains **why an event happens**. It shows the **cause-and-effect relationship** between two factors. This means that one factor directly affects another. Causality is important in scientific studies because it helps find the real reason behind an event.

Data alone may not prove causality, so experiments are often used to identify the actual cause. Understanding causality helps people make better decisions and avoid false assumptions.

Importance of Model Interpretation in Real-World Use Cases

- Model interpretation explains how a machine learning model makes predictions.
- It helps users understand and trust the model's results.
- It also helps identify errors, bias, and incorrect predictions.
- Model interpretation supports better decision making in real-world applications such as healthcare, finance, education, and business.
- Understanding model predictions helps people to use machine learning systems more effectively and responsibly.

Q. Differentiate Between Prediction and Causation

Difference Between Prediction and Causation	
Prediction	Causation
Prediction estimates what may happen by using historical data. It predicts future values or events.	Causation explains why something happens. It identifies the cause-and-effect relationship between different factors.
It does not explain the reason behind the prediction.	It explains how one factor directly affects another.
It uses past data to make predictions.	It studies the cause and effect of events.
It is used in sales forecasting, stock trend analysis, disease risk prediction, and student performance estimation.	It helps people make correct decisions and avoid false assumptions.
Example: Predicting a student's final grade from previous marks, attendance, and study hours.	Example: Regular study hours improve exam performance, so study time is the cause of better grades.

Short Question

What is predictive modeling?

Short Question

What is Causality?

Short Question

Write down importance of model interpretation in real-world use cases.



CLASS ACTIVITY

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

Sol: Activity: Prediction vs. Cause

Step 1: Give Examples

Present these examples to the class:

- Weather forecast says it will rain tomorrow.
- A student who studies regularly is predicted to get good marks.
- A bank predicts that a customer may not repay a loan.
- An online store predicts products a customer may buy.

Step 2: Ask Students

Ask the following questions:

1. Does predicting rain explain why it rains?
2. Does predicting good marks explain the real reason for success?
3. Does predicting loan default explain why a customer cannot repay the loan?
4. What is the difference between predicting an event and explaining its cause?

Step 3: Discussion

- Prediction tells what is likely to happen based on past data or patterns.
- Causality explains why something happens by identifying the actual cause-and-effect relationship.
- A prediction can be accurate without explaining the real cause.

Example

Situation	Prediction	Cause
Weather	It will rain tomorrow.	Clouds, humidity, and atmospheric conditions cause rain.
Student Exam	The student will score high marks.	Regular study, practice, and preparation lead to good marks.

Conclusion

Predictive modeling answers "What is likely to happen?" while causality answers "Why did it happen?" Understanding the difference helps us make better decisions and interpret data correctly.

6.10 Machine Learning Tools and Platforms

Q. What are machine learning tools and platforms? Explain Excel, R, Python and Jupyter Notebooks

Machine Learning Tools and Platforms

Machine learning tools and platforms are software used for data analysis and modeling. These tools help to collect, process, and analyze data. Some tools are simple and easy for beginners, while others are used for advanced machine learning tasks.

Choosing the right tool is important because each tool has its own strengths. These tools improve efficiency, increase accuracy, and help build a strong foundation in machine learning.

Common machine learning tools and platforms include Excel, R, Python, and Jupyter Notebooks.

1. Excel for Data Analysis

Excel is widely used for basic data analysis. It organizes data into rows and columns, making it easy to read and manage. Excel provides features such as sorting, filtering, and charts to analyze and display data. Simple formulas can also be used to perform calculations. Excel is easy to learn and is suitable for small datasets. Many beginners use Excel to understand data before learning advanced tools.

2. R for Data Science

R is a **programming language** used for data analysis and data science. It is widely used in research because it can analyze data efficiently. It provides strong graphical tools to create charts and graphs for data visualization. It can handle large datasets and offers many packages that support data science tasks. Learning R improves analytical skills and helps build a strong foundation in machine learning.

Short Question

What do you know about machine learning tool R?

3. Python and Jupyter Notebooks for Machine Learning

Python

Python is a popular programming language used in machine learning. It provides many useful libraries, such as **Pandas** and **NumPy**, for data analysis and modeling. Python is also used to build advanced machine learning models and is widely used in industry.

Jupyter Notebooks

Jupyter Notebooks provide an **interactive environment** where both code and its output appear together. This makes it easier to learn, test, and understand programs. Jupyter Notebooks also provide an interactive Integrated Development Environment (IDE) for writing and running Python code.

Short Question

What is the purpose of Jupyter Notebooks?



Additional Multiple Choice Questions

- Validation is the process of checking how well a machine learning model works on:
(a) Old data (b) New data (c) Wrong data (d) Random data
- A validation data set is:
(a) Training data (b) Testing data (c) A separate set of data (d) Database table
- Validation data helps detect:
(a) Data storage (b) Overfitting (c) Programming syntax (d) File sharing
- Robustness means that a machine learning model:
(a) Performs well in different situations (b) Uses less memory only
(c) Runs without data (d) Creates databases
- Hyperparameter tuning is the process of:
(a) Cleaning data (b) Adjusting the hyperparameters of a machine learning model
(c) Collecting new data (d) Deleting features
- Predictive modeling is the process of using data to predict:
(a) Past events (b) Hardware faults (c) Future outcomes (d) Program errors
- Causality explains:
(a) Why an event happens (b) How data is stored
(c) Where data is saved (d) How files are copied
- Why is model interpretation important?
(a) It helps users understand and trust the model's results
(b) It increases Internet speed
(c) It creates new datasets
(d) It reduces storage space.
- Which tool is widely used for basic data analysis?
(a) Python (b) Jupyter Notebooks (c) R (d) Excel
- Which platform provides an interactive environment where both code and its output appear together?
(a) Visual Studio (b) Jupyter Notebooks (c) Excel (d) PowerPoint

Answers:

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

TEXTBOOK EXERCISE (SOLVED)



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 organized 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 labeled 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

Answers:

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



SHORT QUESTIONS

1. What is data science?

Ans: 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 different tools and methods to analyze large amounts of data. It supports decision-making in many areas of life.

2. What are the two main types of data?

Ans: i. **Structured Data**

Structured data is organized in a clear format, such as tables or spreadsheets. It is arranged in rows and columns, making it easy to understand and analyze.

ii. **Unstructured Data**

Unstructured data does not follow a fixed format. It includes emails, images, audio, and videos. This type of data is difficult to analyze directly.

3. What is the difference between structured and unstructured data?

Ans:

Structured Data	Unstructured Data
Structured data is organized in a clear format, such as tables or spreadsheets.	Unstructured data does not follow a fixed format.
It is easy to understand and analyze.	It is difficult to analyze directly.
It can be processed using standard database tools.	Special tools are needed to process it.
Examples: Students' mark sheets and databases.	Examples: Emails, images, and videos.

4. Name two common data collection methods.

Ans: Two common data collection methods are surveys and interviews. Surveys collect data from many people using questionnaires. Interviews collect information through direct conversation.

5. What is machine learning?

Ans: Machine learning is a subfield of Data Science that enables systems to improve from experience without explicit programming. It works by analyzing patterns in data to make predictions or decisions. Large datasets are used to train machine learning models.

6. What is supervised learning?

Ans: 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.

7. What is the purpose of feature selection in machine learning?

Ans: Feature selection is the process of choosing the most important features from a dataset. It improves model accuracy, reduces unnecessary data, and makes the machine learning model faster and more efficient.

8. What does accuracy measure in a model?

Ans: Accuracy measures how many predictions made by a model are correct. A higher accuracy means the model makes more correct predictions and performs better.

9. What is the difference between prediction and causality?

Ans:

Prediction	Causality
Prediction estimates what may happen by using historical data.	Causality explains why something happens.
It predicts future values or events.	It identifies the cause-and-effect relationship between different factors.
It does not explain the reason behind the prediction.	It explains how one factor directly affects another.
It is used to forecast future outcomes.	It helps people make correct decisions and avoid false assumptions.

10. Name any one tool used for machine learning.

Ans: Python

Python is a popular programming language used in machine learning. It provides many useful libraries for data analysis and modeling. Python is also used to build advanced machine learning models and is widely used in industry.



LONG QUESTIONS

All Qs are solved in Chapter

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.

ADDITIONAL SHORT QUESTIONS

Introduction to Data Science

1. Why is data science important for modern technology?

Ans: Data science helps analyze large amounts of data. It supports better decision-making. It helps predict future trends. Modern technology depends heavily on data science.

2. How does data science help to solve real-world problems?

Ans: Data science turns data into useful knowledge. It helps answer questions by using facts and data. It supports better decision-making. It helps solve real-world problems.

Understanding Data

3. What is data and analysis?

Ans:

- **Data** is a collection of raw facts and values that are used to produce information. Data can be in the form of numbers, text, images, audio, or video.
- **Analysis** is the process of examining, studying, and interpreting data to find useful information. It helps to understand patterns and relationships in information.

4. Why is it important to understand data before analysis?

Ans: Understanding data is important before analysis. Different types of data need different handling methods. Proper understanding of data helps produce correct results. Poor understanding of data can lead to errors.

5. Name any four common data sources.

Ans: Data can be collected from different sources. Common data sources include surveys and interviews. Questionnaires, forms, sensors, IoT devices, apps, and social media are also common sources. Reliable data sources improve the accuracy of analysis.

6. What is data storage?

Ans: Data storage means saving data for future use. Data can be stored in files, databases, or cloud systems. Proper storage protects data from loss. It helps keep data safe for future use.

7. What is data management?

Ans: Data management means organizing and maintaining data. It makes data easy to find and use. Security is an important part of data management. Well-managed data improves efficiency and supports long-term use of data.

8. Describe any two data collection methods.

Ans: Manual data collection is done by people using surveys, interviews, or forms. Automatic data collection is done by machines, sensors, and software. Real-time data is collected instantly as events happen. Stored data is collected and saved for future analysis.

Machine Learning

9. Differentiate between Traditional Programming and Machine Learning.

Ans:

Traditional Programming	Machine Learning
Traditional programming uses fixed rules written by programmers.	Machine learning uses data instead of fixed rules.
The program follows these rules to produce the required output.	The system learns the rules automatically from the data.
It is suitable for problems with clearly defined rules.	It is better at solving complex problems.
It is less flexible when requirements change.	It can adapt to new situations.

10. What is meant by learning from data?

Ans: Learning from data means using past information to improve future decisions. The model studies the relationships in the data. It updates itself when new data becomes available. This helps the system become smarter over time.

11. Differentiate between Supervised and Unsupervised Machine Learning.

Ans:

Supervised Machine Learning	Unsupervised Machine Learning
Supervised machine learning uses labeled data for training.	Unsupervised machine learning uses unlabeled data for training.
It learns by comparing its output with the correct answers.	It learns by finding patterns and relationships in the data.
It is mainly used for prediction and classification tasks.	It is mainly used for grouping similar data and clustering.
An example is predicting whether students will pass or fail.	An example is grouping customers according to their shopping habits.

12. Why is machine learning better than traditional programming for complex problems?

Ans: Machine learning uses data instead of fixed rules. It learns the rules automatically from the data. It improves when new data is provided. Therefore, it is better at solving complex problems.

13. Why is it important to choose the correct machine learning mechanism?

Ans: Different machine learning mechanisms are used for different problems. Some methods use labeled data, while others use unlabeled data. Choosing the correct mechanism improves the quality of learning. It helps the system make intelligent decisions.

14. How is machine learning used in business?

Ans: Machine learning helps businesses understand customer behavior. It predicts sales and market trends. It detects fraud in financial systems and recommends products. This improves business decisions and efficiency.

15. How is machine learning used in education?

Ans: Machine learning improves learning systems. It helps track students' performance. It provides personalized learning content. It also makes attendance and assessment systems smarter.

Building a Machine Learning Model

16. What is the first step in building a machine learning model?

Ans: The first step is data collection. The required data is collected from different sources. The quality of data affects the performance of the model. Good data helps build a better model.

17. What is model training?

Ans: Model training is the process of teaching the machine learning model. The model learns from the training data. It identifies patterns and relationships in the data. This prepares the model to make predictions.

Model evaluation and Performance Metrics

18. Why is testing data kept separate from training data?

Ans: Testing data is used to check the performance of the model. It is not used during training. It shows how well the model works on new data. This helps measure the accuracy of the model.

19. Why is model evaluation important?

Ans: Model evaluation checks the performance of the trained model. It uses test data to measure the results. It helps identify the strengths and weaknesses of the model. This helps improve the quality of predictions.

20. Differentiate between Training Data and Testing Data.

Ans:

Training Data	Testing Data
Training data is used to teach the model and help it learn patterns.	Testing data is used to check how well the model works on unseen data.
The model learns from the training data.	The model does not learn from the testing data.
It is used to build the machine learning model.	It is used to measure the model's performance.

21. What is precision in machine learning?

Ans: Precision measures how many predicted positive results are actually correct. It focuses on the correctness of positive predictions. High precision means fewer false positive results. It is important when false alarms should be reduced.

22. What is recall in machine learning?

Ans: Recall measures how many actual positive cases are correctly identified. It focuses on finding as many positive cases as possible. High recall reduces missed positive cases. It is important in medical diagnosis and disease detection.

Model Validation and Improvement

23. Why is a validation dataset used?

Ans: A validation dataset is used to test the model during development. It is different from the training dataset. It helps improve the model before final testing. It also checks whether the model performs well on unseen data.

24. Why should training data not be used as validation data?

Ans: Training data is used to teach the model. Validation data is used to check the model's performance. Using the same data for both purposes gives unreliable results. A separate validation dataset provides a fair evaluation.

25. What is overfitting?

Ans: Overfitting happens when a model learns the training data too closely. It performs well on training data but poorly on new data. It cannot make accurate predictions for unseen data. A good model should avoid overfitting.

26. What is robustness in machine learning?

Ans: Robustness is the ability of a model to perform well in different situations. A robust model gives reliable results on new data. It is not easily affected by small changes in the input. This makes the model more dependable.

27. What is hyperparameter tuning?

Ans: Hyperparameter tuning is the process of selecting the best settings for a machine learning model. Different settings affect the model's performance. Proper tuning improves accuracy and reduces errors. It helps build a better model.

Predictive Modeling and Causality, Machine Learning Tools

28. Why is causality important in data analysis?

Ans: Causality helps explain the real reason behind an event. It identifies the relationship between cause and effect. It helps people make better decisions. It also prevents incorrect conclusions.

29. What is Excel?

Ans: Excel is used to organize and manage data. It can perform simple calculations and create charts. It is suitable for small datasets. It is easy for beginners to use.

30. What is Scikit-learn used for?

Ans: Scikit-learn is a Python library for machine learning. It provides tools for building and testing models. It supports classification, regression, and clustering. It is simple and widely used for machine learning tasks.

CONCEPTUAL SHORT QUESTIONS

1. A school wants to predict whether students will pass or fail by using previous exam records. Which machine learning mechanism should be used? Why?

Ans: Supervised machine learning should be used. It uses labeled data for training. The system learns by comparing its output with the correct answer. It is mainly used for prediction tasks.

2. A supermarket wants to group customers according to their shopping habits, but no customer groups are already defined. Which machine learning mechanism should be used?

Ans: Unsupervised machine learning should be used. It works with unlabeled data. The system finds patterns and groups similar data together. It is useful for customer grouping and data clustering.

3. A robot learns to move through a maze by receiving rewards for correct actions and penalties for wrong actions. Which machine learning mechanism is being used?

Ans: Reinforcement machine learning is being used. It learns through experience. The system receives rewards or penalties for its actions. It improves its decisions through trial and error.

4. A teacher wants to identify students who may need extra help before the final examination. How can a machine learning model help?

Ans: The model is trained using students' previous data. It predicts a student's possible grade by using new input data. Teachers can identify students who may need extra support. This helps improve student performance.

5. A hospital wants to detect as many patients with a disease as possible. Which performance metric is more important?

Ans: Recall is more important in this case. It measures how many actual positive cases are correctly identified. High recall reduces missed positive cases. This helps doctors detect more patients.

6. A beginner wants to organize data, create charts, and perform simple calculations. Which machine learning tool is most suitable? Why?

Ans: Excel is the most suitable tool. It organizes data into rows and columns. It provides sorting, filtering, charts, and simple formulas. It is easy to learn and is suitable for small datasets.

7. A company wants to analyze emails, videos, and images received from customers. What type of data is this, and why are special tools needed?

Ans: This is unstructured data. It does not follow a fixed format. It is difficult to analyze directly. Therefore, special tools are needed to process it.

8. A shopping website improves its product recommendations whenever new customer data is added. Which feature of machine learning makes this possible?

Ans: Machine learning learns from new data. The model updates itself and improves its performance. It becomes more accurate over time. This helps provide better product recommendations.

9. A bank wants to detect fraudulent financial transactions automatically. How can machine learning help?

Ans: Machine learning can detect fraud in financial systems. It analyzes patterns in transaction data. It identifies unusual activities automatically. This improves security and reduces fraud.

10. A face recognition system works well in daylight but gives incorrect results at night. What does this show about the model?

Ans: This shows that the model is not robust. A robust model performs well in different situations. It should maintain its accuracy even when the data changes slightly. Reliable models give consistent and accurate results.

TEST YOUR SKILLS

Total Marks: 40

Time: 1:30 Hour

8 x 1 = 8

Q.1 MULTIPLE CHOICE QUESTIONS

1. Data that does not follow a fixed format is called:
(a) Structured data (b) Numerical data (c) Database (d) Unstructured data
2. Data organized in rows and columns is called:
(a) Unstructured data (b) Structured data (c) Raw data (d) Random data
3. Which machine learning mechanism receives rewards or penalties for actions?
(a) Rule-based learning (b) Reinforcement learning
(c) Supervised learning (d) Unsupervised learning
4. The type of machine learning that uses labeled data is:
(a) Unsupervised learning (b) Reinforcement learning
(c) Supervised learning (d) Rule-based learning
5. The metric that measures the overall correctness of a model is:
(a) Precision (b) Recall (c) Accuracy (d) F1-score
6. Which performance metric shows how many predictions are correct out of the total predictions?
(a) Precision (b) Recall (c) F1-Score (d) Accuracy
7. Why is model interpretation important?
(a) It helps users understand and trust the model's results. (b) It increases Internet speed.
(c) It creates new datasets. (d) It reduces storage space.
8. A commonly used tool for machine learning and data analysis is:
(a) MS Word (b) Python (c) PowerPoint (d) Notepad

Q.2 SHORT QUESTIONS:

Note: Attempt any 8 short questions

8 x 2 = 16

- i. What is the difference between structured and unstructured data?
- ii. Name any four common data sources.
- iii. Differentiate between Supervised and Unsupervised Machine Learning
- iv. What is supervised learning?
- v. What is the purpose of feature selection in machine learning?
- vi. What is hyperparameter tuning?
- vii. What is the difference between prediction and causality?
- viii. What is the F1-score performance metric?
- ix. Describe any one tool used for machine learning.
- x. A face recognition system works well in daylight but gives incorrect results at night. What does this show about the model?

LONG QUESTIONS:

Note: Attempt any 2 long questions

2 x 8 = 16

3. Explain the concept of machine learning and describe its different mechanisms, including supervised, unsupervised, reinforcement, and rule-based learning.
4. Explain the steps involved in building a machine learning model, including feature selection, feature engineering, and train-test split.
5. Discuss predictive modeling and causality, and explain the importance of model interpretation in real-world use cases.