RQF LEVEL 5

SWDML501

SOFTWARE DEVELOPMENT

Machine Learning Fundamentals

TRAINER'S MANUAL

October, 2024

AUTHOR'S NOTE PAGE (COPYRIGHT)

MACHINE LEARNING FUNDAMENTALS

2024

ACKNOWLEDGEMENTS

The publisher would like to thank the following for their assistance in the elaboration of this training manual:

Rwanda TVET Board (RTB) extends its appreciation to all parties who contributed to the development of the trainer's and trainee's manuals for the TVET Certificate V in Software Developments, specifically for the module **"SWDML501: Machine Learning Fundamentals".**

We extend our gratitude to KOICA Rwanda for its contribution to the development of these training manuals and for its ongoing support of the TVET system in Rwanda.

We extend our gratitude to the TQUM Project for its financial and technical support in the development of these training manuals.

We would also like to acknowledge the valuable contributions of all TVET trainers and industry practitioners in the development of this training manual.

The management of Rwanda TVET Board extends its appreciation to both its staff and the staff of the TQUM Project for their efforts in coordinating these activities.

This training manual was developed:

Under Rwanda TVET Board (RTB) guiding policies and directives

Under Financial and Technical support of

COORDINATION TEAM

RWAMASIRABO Aimable

MARIA Bernadette M. Ramos

MUTIJIMA Asher Emmanuel

Production Team

Authoring and Review

MUHIRWA David

NYIRANSHUTI Elizabeth

Validation

SEKABANZA Jean de la Paix

NYANDWI Rongin

NIYONSHUTI Yves

Conception, Adaptation and Editorial works

HATEGEKIMANA Olivier

GANZA Jean Francois Regis

HARELIMANA Wilson

NZABIRINDA Aimable

DUKUZIMANA Therese

NIYONKURU Sylvestre

Formatting, Graphics, Illustrations, and infographics

YEONWOO Choe

SUA Lim

SAEM Lee

SOYEON Kim

WONYEONG Jeong

NIYOMUGABO Silas

Financial and Technical support

KOICA through TQUM Project

TABLE OF CONTENT

ACRONYMS

AAA: Authentication, Authorization and Accountability

AI: Artificial intelligent

API: Application Programming Interfaces

AWS: Amazon Web services

CBT/A: Competency Based Training/ Assessment

CI/CD: Continuous Integration/Continuous Deployment

DL: Deep learning

HIPAA: Health Insurance Portability and Accountability Act

HTTPS: Hypertext Transfer Protocol Secure

IDE: Integrated Development Environment

JSON: JavaScript Object Notation

JWT: JSON Web Token

KNN: k-Nearest Neighbors

KOICA: Korea International Cooperation Agency

MaaS: Model-as-a-Service

ML: Machine Learning

NPM: Node package manager

ONNX: Open Neural Network Exchange

PC1: First Principal Component

PC2: Second Principal Component

PCA: Principal Component Analysis

RESTful: Representational state transfer

RL: Reinforcement Learning

RTB: Rwanda TVET Board

TQUM Project: TVET Quality Management Project

TVET: Technical Vocational and Education Training

INTRODUCTIO N

This trainer's manual includes all the methodologies required to effectively deliver the module titled **"Machine Learning Fundamentals."** Trainees enrolled in this module will engage in practical activities designed to develop and enhance their competencies.

The development of this training manual followed the Competency-Based Training and Assessment (CBT/A) approach, offering ample practical opportunities that mirror real-life situations.

The trainer's manual is organized into Learning Outcomes, which is broken down into indicative content that includes both theoretical and practical activities. It provides detailed information on the key competencies required for each learning outcome, along with the objectives to be achieved.

As a trainer, you will begin by asking questions related to the activities to encourage critical thinking and guide trainees toward real-world applications in the labor market. The manual also outlines essential information such as learning hours, didactic materials, and suggested methodologies.

This manual outlines the procedures and methodologies for guiding trainees through various activities as detailed in their respective trainee manuals. The activities included in this training manual are designed to offer students opportunities for both individual and group work. Upon completing all activities, you will assist trainees in conducting a formative assessment known as the end learning outcome assessment. Ensure that trainees review the key reading and the points to remember section.

MODULE CODE AND TITLE: NITML501 MACHINE LEARNING FUNDAMENTALS

Learning Outcome 1: Apply Data Pre-processing.
Learning Outcome 2: Develop Machine Learning Model.
Learning Outcome 3: Perform Model Deployment.

Learning Outcome 1: Apply Data Pre-processing

Indicative contents
1.1 Description of Machine learning concepts **1.2 Preparing Machine Learning environment** **1.3 Data Collection and Acquisition** **1.4 Interpret Data Visualization** **1.5 Perform Data cleaning**

Key Competencies for Learning Outcome 1: Apply Data Pre-processing

Knowledge	Skills	Attitudes
• Description of machine learning concepts • Description of machine learning tools • Description of Data Collection and Acquisition • Description of data Visualization • Description of data cleaning	• Preparing Machine Learning environment • Gathering Machine Learning dataset • Using Types of Data Visualization • Applying data Visualization Best Practices • Interpreting visualizations results • Performing Data cleaning	• Having Teamwork spirit in preparing environment • Being critical thinker in gathering the dataset • Being Innovative interpreting data • Being creative in cleaning data • Having Problem-Solving mindset in visualizing the dataset • Being Attentive analyzing dataset

Duration: 30 hours

Learning outcome 1 objectives:

By the end of the learning outcome, the trainees will be able to:

1. Describe correctly basic concepts as used in machine learning
2. Describe correctly machine learning tools as used in model development
3. Prepare properly machine learning environment as used in developing model
4. Describe clearly data collection and acquisition as used in machine learning
5. Gather correctly machine learning dataset based on project
6. Interpret properly data Visualization based on dataset
7. Use properly the types of data visualization as used in machine learning
8. Describe correctly data cleaning based on dataset
9. Perform correctly data cleaning based on required dataset.

Resources

Equipment	Tools	Materials
● Computer ● Projector ● Storage Devices	● Python Distribution (CPython, Anaconda) ● Python IDEs (Jupyter Notebook, PyCharm, Visual Studio Code, or Spyder) ● Cloud jupyter	● Internet ● Datasets

	notebook platform	

Advance Preparation:

Before delivering this learning outcome, you are recommended to:

- Provide computers capable for running machine learning environment with internet access.

Indicative content 1.1: Description of Machine Learning Concepts

Duration: 6 hours

Theoretical Activity 1.1.1: Description of machine learning concepts.

Notes to the trainer:

- While delivering this content, a small group can be used**.**
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, trainer may pass through the following steps:

Step 1: Introduce the activity and request trainees to answer the following questions:

i. Define machine learning.

ii. Explain machine learning life cycle.

iii. State and explain the applications of Machine Learning.

iv. Explain advantages and disadvantages of machine learning

v. Differentiate between machine learning, artificial intelligence and deep learning

Step 2: Ask trainees to present their findings to trainer and whole class.

Step 3: Provide expert view and clarification on asked questions if any.

Step 4: Ask trainees to read the key readings 1.1.1 in their trainee manual.

Points to Remember

- Ensure careful handling of machine learning tasks to achieve accurate outcomes.
- Follow the Machine Learning Life Cycle steps: Data Collection, Data Preparation, and Model Development.

Theoretical Activity 1.1.2: Description of machine learning types and tools.

Notes to the trainer:

- While delivering this content, a small group can be used.
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer the following questions:

i. Explain the following types of machine learning

a) Supervised

b) Unsupervised

c) Semi-supervised

d) Reinforcement

ii. Explain Machine Learning tools

Step 2: Ask trainees to present their findings to trainer and whole class.

Step 3: Provide expert view and clarification on asked questions if any.

Step 4: Ask trainees to read the key readings 1.1.2 in their trainee manual.

Points to Remember

- Choosing best machine learning types should be based on features and problem to be solved.
- Preparation of machine learning environment should be based on machine learning tools.

Indicative content 1.2: Preparing Machine Learning Environment

Duration: 6 hours

Practical Activity 1.2.1: Installing Python

Notes to the trainer

- While delivering this content, you are required to:
  - ✓ Provide computer lab with internet access.
  - ✓ Provide videos showing how to install python as didactic materials if necessary.
  - ✓ Provide Python setup.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and ask trainees to read the task described below:

As a machine learning trainee, you are asked to install python in computers.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to install python in computers.

Step 4: Asks trainees to perform the activity and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed.

Step 6: Ask trainees to read key readings 1.2.1

Points to Remember

- Set up a machine learning environment properly to ensure smooth functionality.

Practical Activity 1.2.2: Installing tools and test environment

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:

  ✓ Provide computers with internet connectivity and python installed.
  ✓ Provide video as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are asked to go to the computer lab to Install Tools and Test Environment.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to Install Tools and Test Environment.

Step 4: Ask trainees to perform the activity and monitor the procedures.

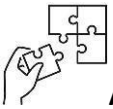
Step 5: Verify whether the tasks are clearly performed

Step 6: Ask trainees to read key reading 1.2.2.

Step 7: Ask trainees to perform the activity in the application of learning 1.2

Points to Remember

- Set up a machine learning environment properly to ensure smooth functionality.
- Avoid skipping essential steps like verifying Python installation or neglecting to install key libraries, as this can lead to issues in running machine learning models later.



Application of learning 1.2.

As trainee in machine learning, you are requested to install Python setup, Tools and Testing Python Environment in your computer lab for Preparing Machine Learning environment.

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Python is well installed	1.1 Python setup is downloaded		
		1.2 Python Version is selected		
		1.3 Python is installed		
		1.4 Python Installation is verified		
		1.5 Python idle is opened		
2	Installation of tools and Testing Environment is well performed	2.1 Tools are selected		
		2.2 Tools are installed		
		2.3 Environment is tested		

Indicative content 1.3: Data Collection and Acquisition

Duration: 6 hours

Theoretical Activity 1.3.1: Description of data collecting and acquisition

Notes to the trainer:

- While delivering this content, a small group can be used for Describing Data Collection and Acquisition.
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

i. Describe the following terms used in data acquisition

a) data
b) Information
c) dataset
d) Data warehouse
e) Big data

ii. Explain the following terms as used in Source of data

a) IoT Sensors
b) Camera
c) Computer
d) Smartphone
e) Social data
f) Transactional data

iii. Describe 6 V's of Big Data

iv. Explain the types of data

Step 2: Ask trainees to present their findings to trainer and whole class.

Step 3: Provide expert view and clarification on asked questions if any.

Step 4: Ask trainees to read the key readings 1.3.1 in their trainee manual.

Points to Remember

- Collect data from key sources, such as IoT sensors, cameras, computers, and social media.
- Proper pre-processing of datasets is essential for effective model training

Practical Activity 1.3.2: Gathering Machine Learning dataset.

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:

  ✓ Provide computers with internet connectivity and python installed.
  ✓ Provide video as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are asked to gather the Iris dataset.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to gather the Iris dataset.

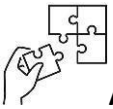
Step 4: Asks trainees to perform the activity and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed.

Step 6: Ask trainees to read key reading 1.3.2.

Points to Remember

- Ensure proper preprocessing of datasets for effective model training to avoid inaccuracies.
- You can find machine learning datasets on platforms like UCI Machine Learning Repository, Kaggle, Scikit-learn, OpenML, and others such as Google Dataset Search, Data.gov, and GitHub.



Application of learning 1.3.

As a trainee in machine learning, you are requested for gathering the dataset.

1. The dataset to be used is iris dataset
2. Visit any platform and download a dataset

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Gathering dataset is well performed	1.1 Platform is selected		
		1.2 Platform is visited		
		1.3 Iris dataset is downloaded		
		1.4 Iris dataset is interpreted		

Indicative content 1.4: Interpret Data Visualization.

Duration: 6 hours

Theoretical Activity 1.4.1: Describing data Visualization tools.

Notes to the trainer:

- While delivering this content, a small group can be used for Describing Data Visualization tools.
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

i. Describe the following popular tools used for data visualization

a) Matplotlib
b) Seaborn
c) Plotly
d) Tableau
e) Power BI

Step 2: Ask trainees to present their findings to trainer and whole class.

Step 3: Provide expert view and clarification on asked questions if any.

Step 4: Ask trainees to read the key readings 1.4.1 in their trainee manual.

Points to Remember

- Apply best practices: Simplify visuals, choose appropriate charts, label axes, and avoid misleading representations.

Practical Activity 1.4.2: Use Types of Data Visualization.

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:

  ✓ Provide computers with internet connectivity and python installed.
  ✓ Provide video as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to the computer lab to Use Types of Data Visualization as used in machine learning.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to Use Types of Data Visualization.

Step 4: Ask trainees to perform the activity and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed

Step 6: Ask trainees to read key reading 1.4.2

Points to Remember

- Apply best practices: Simplify visuals, choose appropriate charts, label axes, and avoid misleading representations.

Practical Activity 1.4.3: Applying data Visualization Best Practices and Interpreting visualizations results.

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:
  - ✓ Provide computers with internet connectivity and python installed.
  - ✓ Provide video as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to the computer lab to apply best practices data Visualization and Interpreting Visualized results in practical activity 1.4.3

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to Use Types of Data Visualization.

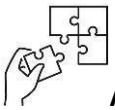
Step 4: Asks trainees to perform the activity and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed

Step 6: Ask trainees to read key reading 1.4.3

Points to Remember

- When interpreting visualizations, identify patterns, provide context, and understand relationships between data variables.



Application of learning 1.4.

As a trainee in machine learning, you are requested to develop a robust machine learning by choosing best Data Visualization tools and interpret visualization result based on iris dataset.

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Types of Data Visualization are well Used	1.1 Scatter Plot is used		
		1.2 Line Plot is used		
		1.3 Bar Chart is used		
		1.4 Histogram chart is used		
		1.5 Box Plot is used		
		1.6 Heat map plot is used		
2	Data Visualization best practices are well performed	2.1 Right Chart Type is selected		
		2.2 Appropriate Scale is used		
		2.3 Elements are clearly labelled		
		2.4 Colors are wisely selected		
3	Interpret visualization of result is well performed	3.1 Patterns and trends are used		
		3.2 Context of the data being visualized is well interpreted		
		3.3 Correlations between variables are used		

Indicative content 1.5: Perform Data Cleaning.

Duration: 6 hours

Theoretical Activity 1.5.1: Describing data cleaning.

Notes to the trainer:

- While delivering this content, a small group can be used for Describing data cleaning.
- Sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions related to data cleaning:

i. Define data cleaning and explain its purpose.
ii. Explain steps involve in data cleaning
iii. Describe characteristics of quality Data
iv. Explain the importance of data cleaning for inconsistencies rectification
v. Explain data cleaning techniques for inconsistencies rectification
vi. Discuss on importance of data normalization
vii. Describe data normalization techniques
viii. Explain the importance of data transformation
ix. Describe data transformation Techniques

Step 2: Ask trainees to present their findings to trainer and whole class.

Step 3: Provide expert view and clarification on asked questions if any.

Step 4: Ask trainees to read the key readings 1.5.1 in their trainee manual

Points to Remember

- Data cleaning is the process of detecting, correcting, or removing inaccurate, incomplete, or inconsistent data to improve its quality.
- It ensures datasets are accurate, complete, and suitable for analysis, enhancing the reliability of statistical models and machine learning algorithms.
- The key steps include handling missing values through imputation or removal, removing duplicates, standardizing data formats, detecting and treating outliers, encoding categorical variables, and applying feature scaling.
- Text data cleaning, ensuring consistent data types, validating integrity, and documenting changes are essential parts of the process.
- Tools like Pandas, NumPy, Scikit-learn, and visualization libraries like Matplotlib help streamline data cleaning workflows.
- High-quality data possesses several characteristics that improve machine learning performance.
- Accuracy ensures correctness, while completeness minimizes missing values.
- Consistency maintains uniform data representation across sources, and relevance ensures only meaningful data is included.
- Timeliness keeps data up-to-date, and uniformity enforces consistent formats.
- Granularity determines the right level of detail, and diversity ensures representation across different conditions.
- Non-redundancy avoids duplication, and scalability ensures efficient processing.
- Structured, well-documented data with clear validity enhances reliability, leading to better model performance and decision-making.
- Inconsistencies in data arise from variations in date formats, measurement units, typos, and missing values, which can compromise data quality. Cleaning these inconsistencies improves data reliability, enhances model performance, reduces costs, facilitates better decision-making, and ensures compliance with regulatory standards. Techniques include handling missing values through imputation or removal, detecting and eliminating duplicates, standardizing formats, identifying and treating outliers, encoding categorical variables, applying feature scaling, and validating data integrity.
- Common techniques include Min-Max normalization, which scales data between 0 and 1, Z-score normalization, which centers data around the mean with a standard deviation of 1, decimal scaling, which adjusts values based on the highest absolute value, and MaxAbs scaling, which normalizes by the maximum absolute value. Choosing the appropriate normalization method depends on the dataset and machine learning algorithm used.

Practical Activity 1.5.2: Performing data cleaning

Notes to the trainer

- While delivering this content, a small group can be used for performing data cleaning.
- Sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to do the task described below:
As full-stack, you are asked to go to the computer lab to performing data cleaning for inconsistencies rectification of provided car dataset and save the creaned data as car_cleaned_dataset.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to performing data cleaning for inconsistencies rectification, while demonstrating explain the steps to be followed.

Step 4: Asks trainees to perform the activity of step 3 and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed.

Step 6: Provide feedback if necessary.

Step 7: Ask trainees to read key reading 1.5.2.

Points to Remember

- **Techniques and steps for data cleaning:**

  Step 1: Load the Dataset
  Step 2: Handling Missing Values
  Step 3: Remove Duplicates
  Step 4: Correct Data Types
  Step 5: Handle Outliers
  Step 7: Encode Categorical Variables
  Step 8: Feature Engineering
  Step 9: Text Data Processing
  Step 10: Final Data Check

Practical Activity 1.5.3: Performing data normalization

Notes to the trainer

- While delivering this content, a small group can be used for performing data normalization on cleaned dataset.
- Sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to do the task described below:
As full-stack, you are asked to go to the computer lab to performing data normalisation for cleaned data on 1.5.2.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to perform data normalisation, while demonstrating explain the steps to be followed.

Step 4: Asks trainees to perform the activity of step 3 and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed.

Step 6: Provide feedback if necessary.

Step 7: Ask trainees to read key reading 1.5.3.

Points to Remember

- Min-Max Scaling: Use when you need values between 0 and 1 (e.g., neural networks).
- Z-score Standardization: Use when data is normally distributed.
- Robust Scaling: Use when there are outliers.

Practical Activity 1.5.4: Performing data transformation

Notes to the trainer

- While delivering this content, a small group can be used for performing data transformation of normalized dataset.
- Sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to do the task described below:
As full-stack, you are asked to go to the computer lab to performing data transformation of normalised data on 1.5.3 for collected dataset.

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to perform data transformation, while demonstrating explain the steps to be followed.

Step 4: Asks trainees to perform the activity of step 3 and monitor the procedures.

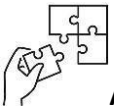
Step 5: Verify whether the tasks are clearly performed.

Step 6: Provide feedback if necessary.

Step 7: Ask trainees to read key reading 1.5.4.

Points to Remember

- Once you want to perform data transformation you follow the following steps.
  1. Load the cleaned dataset
  2. Handle missing values
  3. Detect and remove outliers
  4. Normalize and standardize numerical features
  5. Encode categorical variables
  6. Engineer new features
  7. Transform text data



Application of learning 1.5.

Sarmotor Rwanda has corrected data related to cars in year 2024, as full-stack, you are hired as machine learning expert responsible for performing the followings:

a) Performing data cleaning for inconsistencies rectification
b) Perform data normalisation
c) Perform data transformation

For each activity you have to store the file containing the information of the performed action in independent file.

Learning outcome 1 end assessment

Theoretical assessment

1. Read the following statements concerning machine learning and circle the letter corresponding to the right answer

i. The primary goal of machine learning is:
a) To develop artificial intelligence
b) To automate complex data analysis
c) To create software applications
d) To mimic human intelligence

ii. The following are types of machine learning except.
a) Supervised Learning
b) Unsupervised Learning
c) Reinforcement Learning
d) Pattern Learning

iii. The first step in the machine learning life cycle is:
a) Data Collection
b) Model Evaluation
c) Data Visualization
d) Model Deployment

iv. Machine learning method uses labelled data to train models is:
a) Unsupervised Learning
b) Supervised Learning
c) Semi-supervised Learning
d) Reinforcement Learning

v. The main difference between Artificial Intelligence, Machine Learning, and Deep Learning is:
a) AI is a subset of ML
b) ML is a subset of AI
c) ML is superior to AI
d) None of the above

vi. Tool is for creating interactive data visualizations is:
a) Matplotlib

b) Seaborn
c) Plotly
d) Power BI

vii. Which of the following is a data cleaning technique?
a) Data Normalization
b) Data Transformation
c) Data Aggregation
d) Removing duplicates

viii. "V" in Big Data refers to the speed at which data is generated is:
a) Volume
b) Variety
c) Velocity
d) Veracity

ix. A key characteristic of structured data is:
a) No predefined schema
b) Defined schema and organization
c) Text, image, and video format
d) Generated by social media

x. the following are examples of unstructured data except:
a) Images
b) Audio recordings
c) Excel spreadsheet
d) Videos

II. Select the correct answer from the given choices.

1. _____ data refers to information that is organized in rows and columns in databases or spreadsheets.
Choices:
a) Unstructured data
b) Semi-structured data
c) Structured data
d) Transactional data

Answer: C) Structured data

2. _____ is a popular data visualization library in Python for creating plots like line charts and histograms.
Choices:

a) Matplotlib
b) Seaborn
c) Power BI
d) Tableau

Answer: A) Matplotlib

3. _____ learning requires labeled data to make accurate predictions.
   Choices:
   a) Supervised Learning
   b) Unsupervised Learning
   c) Reinforcement Learning
   d) Semi-supervised Learning

Answer: **A) Supervised Learning**

4. The _____ plot is ideal for showing the distribution of numerical data by representing their quartiles and potential outliers.
   Choices:
   a) Scatter Plot
   b) Box Plot
   c) Bar Chart
   d) Line Plot

Answer: B) Box Plot

5. Data _____ is essential to ensure that machine learning models do not give too much importance to certain features with larger scales.
   Choices:
   a) Normalization
   b) Aggregation
   c) Transformation
   d) Cleaning

Answer: A) Normalization

III. Read the following statements concerning machine learning and State whether the following statements are True or False.

1. Supervised learning involves training a model using labelled data.
   Answer: True
2. In unsupervised learning, the system tries to find patterns in data without being explicitly told what to look for.
   Answer: True
3. Machine learning is a subset of artificial intelligence.
   Answer: True

4. Data visualization is not necessary for interpreting machine learning models.
   Answer: False.
5. The 6 V's of Big Data are Volume, Variety, Velocity, Veracity, Value, and Variability.
   Answer: True

IV. Match the terms related to data collection and acquisition in column A with their corresponding meaning in the column B and write the answer in the appropriate place.

Answers	**Terms**	**Definitions**
1…..	1. Big Data	**C)** Large datasets that cannot be processed using traditional data processing techniques.
2…….	2. Data visualization	
3…….	3. Semi-Structured Data	**E)** Data that has some organizational properties but not as structured as databases.
4….	4. Data Cleaning	**B)** Ensuring that data is accurate, consistent, and free of errors.
5….	5. Data Transformation	**A)** The process of converting raw data into a machine-readable format.
6……	6. Data Normalization	**D)** A process used to scale data to a standard range, usually between 0 and 1.
7…….	Data interpretation	

Answer:

Answers	**Terms**	**Definitions**
1…..C	1. Big Data	**C)** Large datasets that cannot be processed using traditional data processing techniques.
2…….	2. Data visualization	
3…….E	3. Semi-Structured Data	**E)** Data that has some organizational properties but not as structured as databases.
4….B	4. Data Cleaning	**B)** Ensuring that data is accurate, consistent, and free of errors.
5….A	5. Data	**A)** The process of converting raw data into a machine-readable

	Transformation	format.
6……D	6. Data Normalization	**D)** A process used to scale data to a standard range, usually between 0 and 1.
7…….	Data interpretation	

Practical assessment

As a trainee in machine learning, you are requested by your school to go in computer lab to prepare a robust machine learning environment to measure the student performance, by performing the following tasks:

1. Preparing Machine Learning environment by Installing Python, installing tools and test Environment
2. Collect the student performance dataset
3. Use Types of Data Visualization like Scatter Plots, Line Plots, Bar Charts, Histograms, Box Plots and Heat map
4. Choose the best practices for data visualization and interpretation results
5. Perform data cleaning for unnecessary data in dataset
6. Apply data normalization and transformation for preparing dataset.

Checklist

SN	Criteria	Indicators	Observation	
			Yes	No
1	Machine Learning environment is well Prepared	1.1 Python setup is selected		
		1.2 Python is Installed		
		1.3 Machine learning tools are installed		
		1.4 Machine learning environment is tested		
2	Data Collection and Acquisition is well performed.	2.1 Sources of Data are selected		
		2.2 Formats for Datasets are defined		
		2.3 Dataset is loaded		
3	Types of Data Visualization are well Used	3.1 Scatter Plot is used		
		3.3 Line Plot is used		
		3.4 Bar Chart is used		
		3.5 Histogram is used		
		3.6 Box Plot is used		
		3.7 Heat map plot is used		
4	Visualization results are well Interpreted	4.1 Patterns and Trends are interpreted		
		4.2 Context and Background are interpreted		
		4.3 Correlation and Relationship are interpreted		
5	Data cleaning is well performed	5.1 Missing Data is Handled		
		5.2 Outliers are well Handled		
		5.3 Normalization is performed		
		5.4 Data transformation is performed		

END

Further information to the trainer

Brownlee, J. (2016). Master Machine Learning Algorithms - Discover how they work.

Eric, H., Sinayobye, O., Masabo, E., Ufitinema, C., Murangira, T., Rwibasira, P., . . . Gaurav, B. (2023). An Intelligent Rwandan Arabica Dataset. Technologies, 23.

Hitimana, E., Richard, M., Bajpai, G., Louis, S., & Kayalvizhi, J. (2021). Implementation of IoT Framework with Data Analysis Using Deep Learning Methods for Occupancy Prediction in a Building Future Internet.

System-Based Coffee Plant Leaf Disease Recognition Using Deep Learning Techniques.

Learning Outcome 2: Develop Machine Learning Model

Indicative contents

2.1 Description of Machine learning algorithms and applications

2.2 Selection of machine learning algorithm

2.3. Train machine learning model

2.4. Evaluation of machine learning model

2.5 Tuning Hyperparameters

Key Competencies for Learning Outcome 2: Develop Machine Learning Model

Knowledge	Skills	Attitudes
• Description Machine learning algorithms and their applications. • Selecting machine learning algorithm.	• Analyzing type of data based on dataset. • Training machine learning model based on dataset. • Evaluating machine learning model based on provided data. • Tuning Hyperparameters based on performance. • Resolving Potential Bias.	• Teamwork ability while training a model. • Being critical thinker in evaluating model. • Problem-Solving in resolving potential bias. • Having good Collaboration and Communication.

Duration: 30 hours

Learning outcome 2 objectives:

By the end of the learning outcome, the trainees will be able to:

1. Describe correctly Machine learning algorithms and their applications based on model.
2. Select correctly machine learning algorithm based on model.
3. Analyze properly the types of data based on the dataset.
4. Train properly machine learning model based on dataset.
5. Evaluate correctly machine learning model based on provided data.
6. Tune correctly Hyperparameters based on performance.
7. Resolve correctly Potential Bias based on test data.

Resources

Equipment	Tools	Materials
● Computer. ● Projector. ● Storage Devices.	● Datasets. ● Python Distribution (CPython, Anaconda), ● Python IDEs (Jupyter Notebook, PyCharm, Visual Studio Code, or Spyder). ● Machine Learning Libraries (NumPy, Pandas, Scikit-Learn, TensorFlow or PyTorch, Matplotlib, Seaborn, Keras,	● Internet.

	SciPy). ● Cloud platform.	

Advance Preparation:

Before delivering this learning outcome, you are recommended to:

- Ensure the computer is available with Python setup.
- Ensure the computer is available with internet access.
- Make the dataset available.

Indicative content 2.1: Description of Machine Learning Algorithm and Applications.

Duration: 6 hours

Theoretical Activity 2.1.1: Description of Machine learning algorithms and applications

Notes to the trainer:

- Ensure the prepared computer lab and internet access are available.
- While delivering this content, a small group can be used for description supervised machine learning algorithm.
- Provide sample videos demonstrating how the algorithm is implemented, to be used as didactic material if necessary.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to read the following questions:

i. Explain the following supervised machine learning algorithms

i. Linear Regression
ii. Logistic Regression
iii. Decision Trees
iv. Random Forest
v. Support Vector Machine (SVM)
vi. k-Nearest Neighbors (KNN)
vii. Naive Bayes

ii. List down at least 5 application machine learning algorithms

Step 2: Ask trainees in present of their findings in whole class.

Step 3: Provides expert view and ask question if any.

Step 4: Ask trainees to read the key reading 2.1.1 in the trainee manual.

Points to Remember

- Use Supervised Learning algorithms with labeled data to ensure accurate model training.

Theoretical Activity 2.1.2: Description of Unsupervised Machine Learning algorithm

Notes to the trainer:

- While delivering this content, a small group can be used for Unsupervised Machine Learning algorithm.
- Provide videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

i. Explain the following unsupervised machine learning algorithms
    i. K-Means Clustering
    ii. Principal Component Analysis (PCA)
    iii. Hierarchical Clustering
    iv. Anomaly Detection
ii. List down 5 application unsupervised machine learning algorithms

Step 2: Ask trainees in present of their findings in whole class.

Step 3: Provides expert view and clarifies ideas by using didactic materials if it's necessary.

Step 4: Address any questions or concerns.

Step 5: Ask trainees to read the key reading 2.1.2 in the trainee manual.

Points to Remember

- Use unsupervised Learning algorithms with unlabeled data to ensure accurate model training.

Theoretical Activity 2.1.3: Semi-supervised Learning and Reinforcement Learning algorithm

Notes to the trainer:

- While delivering this content, a small group can be used for Description Semi-Supervised Learning Models and Reinforcement Learning.
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

i. Explain semi-supervised and re-enforcement machine learning algorithm
ii. List down at least 5 applications of Semi-Supervised Learning and re-enforcement machine learning algorithms

Step 2: Involve trainees in presentation of their findings.

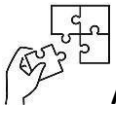
Step 3: Provides expert view and clarifies ideas by using didactic materials.

Step 4: Address any questions or concerns.

Step 5: Ask trainees to read the key reading 2.1.3 in the trainee manual.

Points to Remember

- Semi-supervised Learning uses a mix of labeled and unlabeled data, while Reinforcement Learning involves an agent interacting with an environment to learn decision-making strategies.



Application of learning 2.1.

As a trainee in machine learning, you're tasked to develop a machine learning solution for your school platform of analyzing the student performance based on each academic year. Which machine learning algorithms based on the nature of the problems you are solving should be used and why?

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Types of machine learning algorithm is well identified	1.1 Types of machine learning is selected		
		1.2 Data features is well identified		
		1.3 Algorithms is Chosen		

Indicative content 2.2: Selection of Machine Learning Algorithm

Duration: 6 hours

Theoretical Activity 2.2.1: Selecting machine learning based on algorithm machine learning based on algorithm machine learning based on algorithm

Notes to the trainer:

- Trainer may use small groups of trainees to Selecting machine learning based on algorithm.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

I. State and explain steps followed to choose the right machine learning.

Step 2: Ask trainees to present of their findings to trainer whole class.

Step 3: Provide expert view and clarification and ask question if any.

Step 4: Ask trainees to read the key reading 2.2.1 in the trainee manual.

Points to Remember

- Always follow all steps for Choosing the right machine learning algorithm based on the problem to be solved.

Application of learning 2.2.

As a trainee in machine learning, you're tasked to select the machine learning algorithm among many, which should be the best for developing a machine learning solution for your school platform of analyzing the student performance based on each academic year.

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Selecting learning algorithms and applications	1.1Problem is well identified		
		1.2 Types of data is analyzed		
		1.3 Resources is well analyzed		
		1.4 Algorithm is selected		

Indicative content 2.3: Train Machine Learning Model

Duration: 6 hours

Practical Activity 2.3.1: Loading the dataset

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:
  - ✓ Ensure computers with internet connectivity and Python installed are available.
  - ✓ Provide videos as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are asked to go to the computer lab to load the dataset by using the following library:

a) Using pandas
b) Using NumPy
c) Using Scikit-Learn
d) Using Seaborn
e) Using Requests and io

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to load the dataset.

Step 4: Ask trainees to perform the activity and monitor the procedures.

Step 5: Verify whether the tasks are clearly performed.

Step 6: Ask trainees to read key reading 2.3.1.

Points to Remember

- Loading the dataset, we use different libraries based on the data format.

Practical Activity 2.3.2: Train machine learning model

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:
- Ensure computers with internet connectivity and Python installed are available.
- Provide videos as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to the computer lab to Train machine learning model.

Step 2: Demonstrate how to Train machine learning model based on performance.

Step 3: Ask trainees to perform the activity and monitor the procedures.

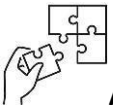
Step 4: Verify whether the tasks are clearly performed.

Step 5: Ask trainees to read key reading 2.3.2.

Step 6: Ask trainees to perform the activity in the application of learning 2.3.

Points to Remember

- While training a model for machine learning ensure that Training, testing, and validation data are set.



Application of learning 2.3.

As a trainee in machine learning, you're tasked to select the machine learning algorithm among many, which should be the best for developing a machine learning solution for your school platform of analyzing the student performance based on each academic year by performing the following tasks:

1. Load a dataset Using pandas, NumPy, Scikit-Learn, Seaborn and Requests and io.
2. Split dataset using train set, test set and validation set.
3. Initialize model.
4. Fit the training data into a model.

checklist:

SN	Criteria	Indicators	Yes	No
1	Dataset is well loaded	1.1 Pandas is used		
		1.2 NumPy is used		
		1.3 Scikit-Learn is used		
		1.4 Seaborn is used		
		1.5 Requests and io are used		
2	Splitting dataset, initializing model and fit the training data into a model is well performed	2.1 Train set is used, and		
		2.2 Test set is used		
		2.3 Validation set is used		
		2.4 Model is initialized		
		2.5 Training data are fitted into a model		

Indicative content 2.4: Evaluation of Machine Learning Model

Duration: 6 hours

Practical Activity 2.4.1: Evaluating machine learning model based on provided data

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, ensure computers with a trained model are available.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to the computer lab to evaluate machine learning model.

Step 2: Demonstrate how to evaluate machine learning model based on data.

Step 3: Ask trainees to perform the activity and monitor the procedures.

Step 4: Verify whether the tasks are clearly performed.

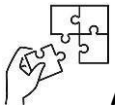
Step 5: Ask trainees to read key reading 2.4.1.

Step 6: Ask trainees to perform the activity in the application of learning 2.4.

Points to Remember

- The model's predictions are evaluated on test data to check generalization. It is also tested on new data to ensure it performs well in real-world scenarios.
- Visualization tools like Seaborn and Matplotlib help compare predicted vs. actual values. Common graphs include scatter plots for regression and confusion matrices for classification.

- Metrics like accuracy, precision measure model performance in classification tasks. Each metric provides a unique perspective on prediction quality.
- For regression, key metrics include RMSE, R Squared, and Adjusted R Squared, all of which help evaluate how closely predictions align with actual values.
- Feature importance and coefficient analysis offer insights into how a model makes decisions. Tools like SHAP and LIME explain individual predictions in greater detail.



Application of learning 2.4.

As a trainee in machine learning, you're tasked to evaluate the trained machine learning model solution for your school platform of analyzing the student performance based on each academic year by performing the following tasks:

1. Predict result on the test data and new data (unseen data).
2. Visualize the prediction.
3. Analyze evaluation metrics using accuracy, precision, Root Mean Square Error (RMSE), R Squared score and Adjusted R squared score.
4. Interpret the model.

Checklist:

SN	Criteria	Indicator	Observation	
			Yes	No
1	Predicting result and visualization are well performed	1.1 Test data is predicted		
		1.2 New data (unseen data) is predicted		
		1.3 Predictions are visualized using different plotting tools		
2	Evaluation metrics are well Analyzed	2.1 Accuracy is used		
		2.2 Precision is used		
		2.3 R Squared score is used		
		2.4 Adjusted R squared score is used		

		2.5 Model is interpreted		

Indicative content 2.5: Tuning Hyperparameters

Duration: 6 hours

Theoretical Activity 2.5.1: Tuning Hyper-parameters based on performance

Notes to the trainer:

- Ensure the computer has a running model available.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to read and answer the following questions:

i. Define Hyperparameters search space.
ii. Choose performance metric.
iii. Perform hyperparameter search.

Step 2: Ask the trainee to present their findings to trainer and the whole class.

Step 3: Provides expert view and clarification and ask question if any.

Step 4: Ask trainees to read the key reading 2.5.1 in the trainee manual.

Points to Remember

- The model's predictions are evaluated on test data to check generalization. It is also tested on new data to ensure it performs well in real-world scenarios.
- To find the optimal combination for the best model performance use the best hyperparameter.

Practical Activity 2.5.2: Evaluate performance

Notes to the trainer

- Ensure computers with internet connectivity and Python installed are available.

- Provide videos as didactic material.

 **Key steps:**

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to the computer lab to evaluate the machine learning model performance.

Step 2: Demonstrate how to evaluate the machine learning model performance.

Step 3: Ask trainees to perform the activity and monitor the procedures.

Step 4: Verify whether the tasks are clearly performed.

Step 5: Ask trainees to read key reading 2.5.2.

 **Points to Remember**

- Always evaluate performance by identifying the best parameters and model, predicting on validation data, and applying evaluation metrics on both validation and test data to ensure model effectiveness.

 **Practical Activity 2.5.3: Resolving Potential Bias**

 **Notes to the trainer**

- This activity should take place in a Computer Lab.
- While delivering this content, ensure computers with a trained model are available.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and ask trainees to read the task described below:

As a machine learning trainee, you are requested to go to Resolve Potential Bias of machine learning model.

Step 2: Demonstrate how to resolve potential bias of machine learning model.

Step 3: Ask trainees to perform the activity and monitor the procedures.

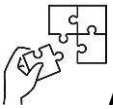
Step 4: Verify whether the tasks are clearly performed.

Step 5: Ask trainees to read key reading 2.5.3.

Step 6: Ask trainees to perform the activity in the application of learning 2.5.

Points to Remember

- Applying evaluation metrics on both validation and test data to ensure model effectiveness for resolving potential bias.



Application of learning 2.5.

Referring to the trained model in application learning 2.5, As a trainee in machine learning, you're tasked to evaluate the performance metric and resolve potential bias of machine learning model solution for your school platform of analyzing the student performance based on each academic year by performing the following tasks:

1. Tuning Hyperparameter Using GridSearchCV
2. Apply Evaluation Metrics on Validation Data
3. Apply Evaluation Metrics on test Data
4. Detect the overfitting and underfitting of a model
5. Re-evaluate on Validation and Test Data

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Hyperparameter Tuning Using GridSearchCV	1.1 GridSearchCV is implemented		
		1.2 Best hyperparameters are found		
		1.3 Cross-validation is applied in GridSearchCV		
2	Applying Evaluation Metrics on Validation Data	2.1 Accuracy is calculated on validation data		
		2.2 Precision is calculated on validation data		
		2.3 Root Mean Squared Error (RMSE) is calculated		
		2.4 R-Squared and Adjusted R-Squared scores are calculated (for regression tasks)		
		2.5 Predictions are visualized using confusion matrix, ROC curve, etc.		
3	Applying Evaluation Metrics on Test Data	3.1 Accuracy is calculated on test data		
		3.2 Precision, and Recall are calculated		
		3.3 RMSE is calculated		
		3.4 R-Squared and Adjusted R-Squared are calculated (for regression tasks)		
		3.5 Test data predictions are visualized		
4	Detecting Overfitting and Underfitting of the Model	4.1 Training accuracy is much higher than validation accuracy (overfitting)		
		4.2 Training and validation errors are both high (underfitting)		

		4.3 Learning curves are generated		
5	Re-evaluating on Validation and Test Data After Adjustments	5.1 Model complexity is reduced (if overfitting)		
		5.2 Model complexity is increased (if underfitting)		
		5.3 Updated evaluation metrics are calculated on validation data		
		5.4 Updated evaluation metrics are calculated on test data		
6	Resolving Potential Bias	6.1 Data imbalance or skewed representation is detected		
		6.2 Data is resampled or synthetic data is generated to balance the dataset		
		6.3 Model performance is consistent across different demographic groups		

Learning outcome 2 end assessment

Theoretical assessment

I). Read the following statements concerning machine learning algorithm and circle letter corresponds the right answer

Q1: The primary purpose of Linear Regression in machine learning is:

A) Classifying categorical data
B) Predicting continuous values
C) Clustering data points
D) Reducing dimensionality

Answer: **B) Predicting continuous values**

Q2: Logistic Regression is mainly used for which type of problem?

A) Regression problems
B) Clustering problems
C) Classification problems
D) Dimensionality reduction problems

Answer: **C) Classification problems**

Q3: The primary goal of a Decision Tree in machine learning is

A) To perform linear regression
B) To group similar data points
C) To split data based on feature values to make decisions
D) To reduce the dimensionality of the dataset

Answer: **C) To split data based on feature values to make decisions**

Q4: The key feature of the Random Forest algorithm is:
A) It uses a single decision tree for classification
B) It builds multiple decision trees and combines their predictions
C) It only works for linear data
D) It is used for clustering tasks

Answer: **B) It builds multiple decision trees and combines their predictions**

Q5: The main goal of Support Vector Machine (SVM) algorithm is:
A) To minimize the distance between data points and the centroid
B) To find the hyperplane that best separates classes

C) To perform feature extraction from the data
D) To find clusters in the data

Answer: **B) To find the hyperplane that best separates classes**

Q6: k-Nearest Neighbors (KNN) algorithm is used for:
A) Clustering
B) Regression and classification
C) Feature extraction
D) Dimensionality reduction

Answer: **B) Regression and classification**

Q7: The primary goal of K-Means Clustering algorithm is:
A) To classify data points into predefined classes
B) To minimize the distance between data points and their corresponding cluster centroid
C) To reduce the dimensionality of the dataset
D) To detect anomalies in the data

Answer: **B) To minimize the distance between data points and their corresponding cluster centroid**

Q8: Which of the following is a limitation of K-Means Clustering?
A) It can only work with categorical data
B) It requires knowing the number of clusters (K) in advance
C) It is sensitive to the size of the dataset
D) It does not require centroids for clustering

Answer: **B) It requires knowing the number of clusters (K) in advance**

Q9: The main purpose of Principal Component Analysis (PCA) is:
A) To perform classification tasks
B) To reduce the dimensionality of the dataset by transforming it into a new coordinate system
C) To find the best hyperplane that separates the data into classes
D) To assign data points to predefined clusters

Answer: **B) to reduce the dimensionality of the dataset by transforming it into a new coordinate system**

Q10: What is the purpose of Anomaly Detection in machine learning?
A) To classify data into multiple categories
B) To identify rare items, events, or observations that deviate significantly from the majority of the data
C) To reduce the dimensionality of the dataset
D) To perform regression on the dataset

Answer: **B) To identify rare items, events, or observations that deviate significantly from the majority of the data**

Q11: Which algorithm is commonly used for Anomaly Detection?

A) Support Vector Machine (SVM)
B) Random Forest
C) Isolation Forest
D) k-Nearest Neighbors (KNN)

Answer: **C) Isolation Forest**

Q12: Semi-supervised Learning is?

A) A method where the model learns from labeled data only
B) A method where the model learns from unlabeled data only
C) A method where the model learns from both labeled and unlabeled data
D) A method that combines clustering and regression

Answer: **C) A method where the model learns from both labeled and unlabeled data**

Q13: Which of the following is an advantage of Semi-supervised Learning?

A) It requires no labeled data
B) It reduces the need for a large amount of labeled data
C) It always performs better than supervised learning
D) It does not require any data preprocessing

Answer: **B) It reduces the need for a large amount of labeled data**

Q14: The primary purpose of the Pandas library in machine learning is:

A) Performing matrix operations
B) Building and training machine learning models
C) Data manipulation and analysis, especially with tabular data
D) visualizing data with charts and graphs

Answer: **C) Data manipulation and analysis, especially with tabular data**

Q15: Which Pandas function is used to load a CSV file into a DataFrame?

A) Pd.load_csv ()
B) pd.read_csv()
C) pd.import_csv()
D) pd.open_csv()

Answer: **B) pd.read_csv()**

Q16: The primary use of Numpy library in machine learning?

A) Data visualization
B) Numerical computation and array manipulation

C) Building neural networks
D) Data collection from web APIs

Answer: **B) Numerical computation and array manipulation**

Q17. Which of the following is the correct way to create a 1-dimensional Numpy array from a Python list?
A) np.array([1, 2, 3, 4])
B) np.array([1:4])
C) np.newarray(1, 2, 3, 4)
D) np.array_list(1, 2, 3, 4)

Answer: **A) np.array([1, 2, 3, 4])**

Q18: Which function in Scikit-Learn is used to split a dataset into training and testing sets?
A) train_test_split()
B) data_split()
C) train_test()
D) split_data()

Answer: **A) train_test_split()**

Q19: The main purpose of Requests library in machine learning projects is ?
A) Loading and preprocessing datasets
B) Sending HTTP requests to access data from APIs or web servers
C) Building machine learning models
D) Performing matrix operations

Answer: **B) Sending HTTP requests to access data from APIs or web servers**

Q20: Which of the following libraries is used primarily for creating arrays and performing fast numerical operations on large datasets?
A) Pandas
B) Seaborn
C) Numpy
D) Scikit-Learn

Answer: **C) Numpy**

II) Read the following statements concerning machine learning algorithm and Match the terms in the column A with their corresponding meaning in the column B and write the answer in the appropriate place.

Answer	**Column A**	**Column B**
……	a) Feedforward Neural Networks (FNNs)	**1.** A neural network architecture used primarily for sequence data, where the

		output of one layer depends on the previous output in the sequence. This allows it to capture temporal dependencies in the data.
……	b) Random Forest	**2.** A type of neural network that processes data in one direction (from input to output) without looping back, commonly used for tasks such as classification and regression.
…….	c) Convolutional Neural Networks (CNNs)	**3.** The general term for a network of neurons connected by synapses, typically including an input layer, hidden layers, and an output layer. This structure is used in various fields such as image recognition and natural language processing
…….	d) Recurrent Neural Networks (RNNs)	**4.** A neural network architecture commonly used for image processing tasks, utilizing convolutional layers to detect spatial features in images
…….	e) k-Nearest Neighbors (KNN)	
…….	f) Neural Networks (Artificial Neural Networks - ANNs)	

Answer

Answer	**Column A**	**Column B**
……2……	a)Feedforward Neural Networks (FNNs)	**1.** A neural network architecture used primarily for sequence data, where the output of one layer depends on the previous output in the sequence. This allows it to capture temporal dependencies in the data.
……	b)Random Forest	**2.** A type of neural network that processes data in one direction (from input to output) without looping back, commonly used for tasks such as

		classification and regression.
……4…….	c)Convolutional Neural Networks (CNNs)	**3.** The general term for a network of neurons connected by synapses, typically including an input layer, hidden layers, and an output layer. This structure is used in various fields such as image recognition and natural language processing
……1…….	d)Recurrent Neural Networks (RNNs)	**4.** A neural network architecture commonly used for image processing tasks, utilizing convolutional layers to detect spatial features in images
……..	e)k-Nearest Neighbors (KNN)	
……3…….	f)Neural Networks (Artificial Neural Networks - ANNs)	

III) Read the following statements concerning machine learning algorithm and fill in the Gap-Space Questions

1. _____ is used to predict a continuous value based on one or more input features.

Choices:

A) Classification

B) Regression

C) Clustering

D) Reinforcement Learning

Answer: B) Regression

2. In _____, the goal is to group similar data points together based on their characteristics.

Choices:

A) Regression

B) Clustering

C) Classification

D) Dimensionality Reduction

Answer: B) Clustering

3. Logistic Regression is a popular algorithm used for _____ tasks.

Choices:

A) Classification

B) Regression

C) Clustering

D) Reinforcement Learning

Answer: A) Classification

4. _____ is the machine learning task where the model predicts a label from a set of predefined categories.

Choices:

A) Regression

B) Clustering

C) Classification

D) Reinforcement Learning

Answer: C) Classification

Iv). Read the following statement-based on machine learning metrics, Answer on the following question by True or False

1. Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Answer: True

2. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations**.**

Answer: **True**

3. R Squared score measures how well the independent variables explain the variance in the dependent variable, and it ranges from -∞ to 1.

Answer: **False**

4. Adjusted R Squared score adjusts for the number of predictors in the model and prevents overfitting**.**

Answer: **True**

5. In model interpretation, it is important to understand how each independent variable contributes to the model's predictions.

Answer: **True**

6. Precision is a more important metric than Recall when it is critical to avoid false negatives.

Answer: **False**
Explanation: **Recall** is more important than Precision when it is critical to avoid false negatives (e.g., in medical diagnoses where missing a positive case is risky).

Practical assessment

As a trainee in machine learning, you're tasked to develop a machine learning solution for your school platform of analyzing the student performance based on each academic year from 2000 to 2020. They have their dataset based on student performance and you will perform the following tasks:

1. Select different machine learning model to be used
2. Load a dataset Using pandas, NumPy, Scikit-Learn, Seaborn and Requests and io
3. Split dataset using train set, test set and validation set
4. Initialize model
5. Fit the training data into a model
6. Predict result on the test data and new data (unseen data)
7. Visualize the prediction
8. Analyze evaluation metrics using accuracy, precision, Root Mean Square Error (RMSE), R Squared score and Adjusted R squared score
9. Interpret the model
10. Tuning Hyperparameter Using GridSearchCV
11. Apply Evaluation Metrics on Validation Data
12. Apply Evaluation Metrics on test Data
13. Detect the overfitting and underfitting of a model
14. Re-evaluate on Validation and Test Data

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Type of machine learning algorithm to be used is well identified	1.1 Type of machine learning is selected		
		1.2 Data features is well identified		
		1.3 Algorithms is Chosen		
2	Dataset is well loaded	2.1 Pandas is used		
		2.2 NumPy is used		
		2.3 Scikit-Learn is used		
		2.4 Seaborn is used		
		2.5 Requests and io are used		
3	Splitting dataset, initializing model and fit the training data into a model is well performed	3.1 Train set is used		
		3.2 Test set is used		
		3.3 Validation set is used		
		3.4 Model is initialized		
		3.5 Training data are fitted into a model		
4	Hyperparameter is well Tuned Using GridSearchCV	4.1 GridSearchCV is implemented		
		4.2 Best hyperparameters are found		
		4.3 Cross-validation is applied in GridSearchCV		
5	Applying Evaluation Metrics on Validation Data	5.1 Accuracy is calculated on validation data		
		5.2 Precision is calculated on validation data		
		5.3 Root Mean Squared Error (RMSE) is calculated		
		5.4 R-Squared and Adjusted R-Squared scores are calculated (for regression		

		tasks)		
		5.5 Predictions are visualized using confusion matrix, ROC curve, etc.		
6	Applying Evaluation Metrics on Test Data	6.1 Accuracy is calculated on test data		
		6.2 Precision, Recall are calculated		
		6.3 RMSE is calculated		
		6.4 R-Squared and Adjusted R-Squared are calculated (for regression tasks)		
		6.5 Test data predictions are visualized		
7	Detecting Overfitting and Underfitting of the Model	7.1 Training accuracy is much higher than validation accuracy (overfitting)		
		7.2 Training and validation errors are both high (underfitting)		
		7.3 Learning curves are generated		
8	Re-evaluating on Validation and Test Data After Adjustments	8.1 Model complexity is reduced (if overfitting)		
		8.2 Model complexity is increased (if underfitting)		
		8.3 Updated evaluation metrics are calculated on validation data		
		8.4 Updated evaluation metrics are calculated on test data		
9	Resolving Potential Bias	9.1 Data imbalance or skewed representation is detected		
		9.2 Data is resampled or synthetic data is generated to balance the dataset		
		9.3 Model performance is consistent across different demographic groups		

END

Further information to the trainer

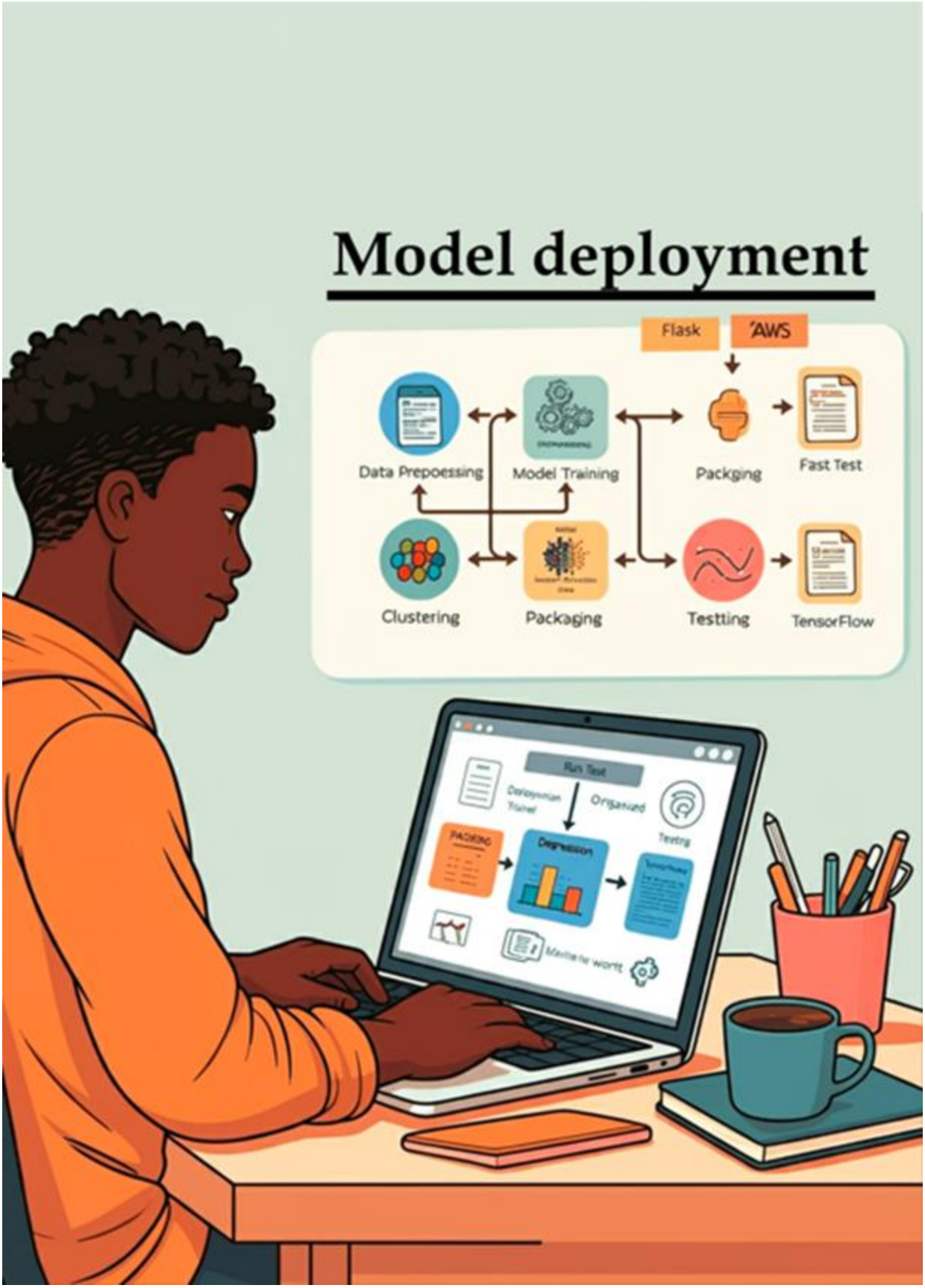
(Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.)

(James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R. Springer.)

(Müller, A. C., & Guido, S. (2016). Introduction to Machine Learning with Python: A Guide for Data Scientists. O'Reilly Media.)

(Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing.)

(Sarkar, D., Raghav, B., & Tushar, S. (2017). Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems.)


Model deployment
Flask
AWS
Data Prepoessing
Model Training
Packging
Fast Test
Clustering
Packaging
Testting
TensorFlow
Run Test
Organized
Testng

Indicative contents
3.1. Selection of model deployment method **3.2 Integration of model file** **3.3 Delivering Prediction to the clients**

Key Competencies for Learning Outcome 3: Perform Model Deployment

Knowledge	Skills	Attitudes
• Description of model deployment methods • identifying system specifications • identifying model specification • Description of Integration of model file • Description of API endpoint and data format. • Description of Prediction to the clients	• Analyzing different deployment methods • choosing the best application type and technology stack. • Evaluating the model's requirements and compatibility. • Integrating model with existing systems • Manage data formats and API usage. • Implementing model communication. • Deploying the model and continuously monitor its performance.	• Critical thinker in choosing the best technology • Being Innovative • Being creative in • Having Problem-Solving • Being collaborator and Communicator • Be attentive to Security in deploying the model.

Duration: 20 hours

Learning outcome 3 objectives:

By the end of the learning outcome, the trainees will be able to:

1. Describe clearly deployment method based on system requirements.
2. Identify correctly system specification based on model deployment.
3. Identify clearly model specification based on system requirement.
4. Integrate effectively model into an existing application architecture.
5. Track/ analyses effectively performance metrics of model based on system requirements.
6. Implement correctly a comprehensive testing model before deployment to ensure reliability and accuracy
7. Describe clearly compliance with relevant industry standards and regulations related to data privacy, security, and ethical AI practices.

Resources

Equipment	Tools	Materials
● Computer ● Projector ● Storage Devices	● Browser ● Postman ● Internet ● Python Distribution (CPython, Anaconda), ● Python IDEs (Jupyter Notebook, PyCharm, Visual Studio Code, or	● Internet ● Datasets, ● Python Distribution (CPython, Anaconda), ● Python IDEs (Jupyter Notebook, PyCharm, Visual Studio Code, or Spyder), ● Machine Learning

	Spyder), ● Machine Learning Libraries (NumPy, Pandas, Scikit-Learn, TensorFlow or PyTorch, Matplotlib, Seaborn, Keras, SciPy), ● Cloud platform, web application framework, web server	Libraries (NumPy, Pandas, Scikit-Learn, TensorFlow or PyTorch, Matplotlib, Seaborn, Keras, SciPy), ● Cloud platform, web application framework, web server

Advance Preparation:

Before delivering this learning outcome, you are recommended to:

- Have computers with installed web browser, Anaconda.
- Have sample model trained to be used for implementing testing.
- Have internet connection accessibility for all computers.

Indicative content 3.1: Selection of model deployment method

Duration: 5 hours

Theoretical Activity 3.1.1: Description of Model Deployment and Specifications

Notes to the trainer:

- While delivering this content, small groups can be used for introducing model deployment methods.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer the following questions:

i. Define the term Model Deployment in Machine Learning

ii. Discuss on benefit of Model Deployment in Machine Learning

Step 2: Ask trainees in presentation of their findings to trainer and whole class.

Step 3: Provide expert view and clarifies ideas if any.

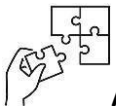
Step 4: Address any questions or concerns.

Step 5: Ask trainees to read the key reading 3.1.1 in their trainee manual.

Points to Remember

- Integrating a machine learning model into an existing production environment it's a crucial thing in machine learning deployment
- while deploy your ML model use the stipulated modes
- While delivering this content, small groups can be used for describing application and technology of model deployment.
- Both memory limitations and computation power play crucial roles in determining how data is processed, models are trained, and which models can be used.
- Applying Common model file formats allow for efficient storage and deployment of trained models.

- While training and saving model, you are always recommended to respect stipulated steps.



Application of learning 3.1.

As Student from level, five you are requested to go to computer Lab are given the task with researching and selecting appropriate deployment methods for a machine-learning model, considering factors like cost, performance, scalability, security, and maintenance**Checklist:**

SN	Criteria	Indicator	Observation	
			Yes	No
1	Data Preparation	1.1 Dataset is properly loaded and inspected.		
		1.2 Data preprocessing is completed.		
2	Model Training	2.1 Dataset is split into training and testing sets.		
		2.2 Model is trained on the training data.		
3	Model Evaluation	3.1 Model predictions are generated on the test data.		
		3.2 Model performance is evaluated and metrics reported.		
4	Model Saving and Loading	4.1 Model is saved correctly using Joblib.		
		4.2 Model is loaded correctly for future predictions.		
5	Predictions on New Data	5.1 Predictions are made on new student data.		
		5.2 Discussion of implications based on model predictions.		

Indicative content 3.2: Integration of Model File

Duration: 10 hours

Theoretical Activity 3.2.1: Description of model file integration

Notes to the trainer:

- While delivering this content, small groups can be used for understanding model integration.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to Discuss on the followings:

i. Explain the goals of machine learning model integration

ii. What is the importance of checking compatibility

iii. Why do we need to use API endpoint for model integration?

Step 2: Involve trainees in presentation of their findings.

Step 3: Provide expert view and clarifies ideas by using didactic materials.

Step 4: Address any questions or concerns.

Step 5: Ask trainees to read the key reading 3.2.1 in the trainee manual.

Points to Remember

- In Model integration combining multiple predictive models, algorithms, or systems to work cohesively as a unified entity enhanced Accuracy, Robustness, Comprehensive Insights of the model.

Practical Activity 3.2.2: Integrating of Model with existing systems

Notes to the trainer

- This activity should take place in the computer lab where trainees should use flask framework to integrate a model file.
- While delivering this content, you are required to:

  ✓ Provide computers with anaconda, visual studio and python installed.
  ✓ Provide sample model trained to be used.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to perform the tasks described below:

i. Test flask python framework successfully installation

ii. Create restful API with flask framework

iii. Avail endpoint to display your model prediction in existing system

iv. Integrate API endpoint created into existing system

v. Use gradio library interface to integrate with model saved

Step 2: Explain the task and provide clear work instruction.

Step 3: Demonstrate how to use flask to integrate a model file. While demonstrating, explain the steps to be followed.

Step 4: Ask trainees to use flask to integrate a model file and monitor the procedures.

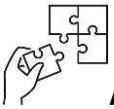
Step 5: Verify whether to use flask to integrate a model file have been performed.

Step 6: Ask trainees to read key reading 3.2.2.

Points to Remember

- Always follow high-level outline to guide you through the process of integrating a model file with Your Existing Application as required.

- Create the API structure, and test it using HTTP methods for RESTful API operations.
- Ensure performance monitoring with tools like Prometheus and Grafana, secure the API with HTTPS, authentication (e.g., OAuth2, JWT), and rate limiting. Track API health using centralized logging (e.g., ELK Stack), and maintain a feedback loop by storing predictions for retraining and tracking performance over time.
- Choose an appropriate model serving method and use Docker for containerization alongside hosting platforms like cloud services or Kubernetes for scalability.
- Set up performance monitoring with tools like Prometheus and Grafana to track API requests, response times, and resource utilization, and test load capacity with tools like JMeter.
- Secure the API using HTTPS, proper authentication (e.g., OAuth2, JWT), and implement rate limiting to prevent misuse.
- Monitor errors and log requests using centralized logging systems like the ELK Stack to track system health and troubleshoot issues.
- Maintain a feedback loop by storing predictions and actual outcomes for model retraining and performance tracking over time.



Application of learning 3.2.

You are tasked with deploying a TensorFlow model to predict customer sentiment in real-time for a customer service application that handles thousands of users daily. The model must be deployed using an appropriate method, such as real-time API serving, to meet the application's need for quick predictions, while considering memory, computing power, and system constraints. Once deployed, you must monitor system performance during peak hours, including API response times, requests, and model accuracy.

Tasks:

1. Choose a suitable deployment method (real-time API serving) and justify based on real-time requirements.
2. Integrate the TensorFlow SavedModel into the web application via an API to serve predictions in real-time.
3. Monitor and optimize system performance during peak hours, tracking API requests, response times, and model accuracy.
4. Implement performance monitoring tools and address potential slowdowns using tools like Prometheus and Grafana.

5. Secure the API using HTTPS, authentication (e.g., OAuth2, JWT), and rate limiting to ensure robust operation.

Checklist:

SN	Criteria	Indicator	Observation	
			Yes	No
1	Deployment Method Selection	1.1 Chosen method aligns with real-time prediction requirements		
		1.2 Justification provided for the selected deployment method		
		1.3 System constraints considered in the method selection		
2	System Specifications	2.1 Identified memory and computing power requirements for the model		
		2.2 Type of application (web, mobile, embedded) specified		
		2.3 Model file format compatibility assessed (TensorFlow SavedModel)		
		2.4 Considered potential memory or compute limitations		
3	Model Integration	3.1 API endpoint created to serve predictions		
		3.2 Model successfully integrated into the existing application		
		3.3 Predictions are correctly formatted for display to end-users		
		3.4 Input data format (JSON, form data) identified and handled properly		
4	Performance Monitoring	4.1 API request and response times tracked		
		4.2 Monitoring for model accuracy established		
		4.3 Higher response times during peak hours addressed		
		4.4 Optimization techniques (e.g., quantization) applied for improved performance		
5	Error Handling	5.1 Error responses handled correctly in case of prediction failures		
		5.2 Corrective measures for improving prediction accuracy implemented		

Indicative content 3.3: Delivering Prediction to the Clients

Duration: 5 hours

Theoretical Activity 3.3.1: Describing API Integration for Model Predictions

Notes to the trainer:

- While delivering this content, a small group can be used for Describing API Integration for Model Predictions.
- Provide sample videos to be used as didactic materials.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the activity and request trainees to answer to the following questions:

i. What is the role of an API in integrating machine learning models into applications?
ii. How does an API handle the input and output data when serving model predictions?
iii. What are common issues that may occur during API-based model predictions, and how can they be resolved?
iv. Why is it important to format predictions correctly when delivering them to a client application?
v. What error-handling strategies should be implemented in an API to ensure seamless delivery of predictions to clients?

Step 2: Involve trainees in presentation of their findings.

Step 3: Provides expert view and clarifies ideas by using didactic materials.

Step 4: Ask trainees to read the key reading 3.3.1 in the trainee manual.

Points to Remember

- Use APIs as bridges between machine learning models and client applications to enable seamless communication and real-time interaction by exposing model functionalities as services.
- Ensure proper data handling by structuring input data (e.g., JSON or form data) and formatting output predictions in a user-friendly way for client consumption.
- Format predictions correctly to provide clear, interpretable, and consistent data that enhances the user experience and reduces confusion

Practical Activity 3.3.2: implementing API Integration and Error Handling for a Machine Learning Model

Notes to the trainer

- This activity should take place in a Computer Lab.
- While delivering this content, you are required to:
  - ✓ Provide computer with internet connectivity.
  - ✓ Provide video as didactic material.

Key steps:

While delivering this activity, pass through the following steps:

Step 1: Introduce the topic and ask trainees to read the task described below:

As a machine learning trainee, you are asked to go to the computer lab to deploy machine learning model into a web and implement error-handling mechanisms.

Step 2: Explain the task and provide clear work instruction.

Step 3: Ask trainees to read key reading 3.3.2.

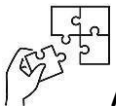
Step 4: Demonstrate how to deploy machine learning model into a web and implement error-handling mechanisms.

Step 5: Ask trainees to perform the activity of step 4 and monitor the procedures.

Step 6: Verify whether the tasks are clearly performed.

Points to Remember

- Implementing API Integration and Error Handling for a Machine Learning Model is crucial for enhancing performance of the model.
- Plan for future updates by using version control in your API
- Design the API to handle multiple requests concurrently, considering load balancing if needed for large-scale applications.



Application of learning 3.3.

Mietech LTD is a house seller company located in Gasabo district, it used to sell house through application. The task involves using an API to serve predictions based on user inputs such as location, square footage, and the number of rooms. You need to ensure that the predictions are correctly formatted for display and handle errors like invalid input or server issues, providing clear feedback to users. The APIs will facilitate interaction between the application and the model, focusing on prediction delivery and error management for a seamless user experience. You are assigned a task to integrate a machine learning model for predicting house prices into a real estate web (application).

Checklist:

SN	Criteria	Indicator	Observation	
			Yes	No
1	API Integration	1.1 API successfully integrated into the web application		
		1.2 API correctly communicates with the machine learning model		
		1.3 Predictions are served in real-time based on user input		
2	Prediction Formatting	2.1 Predictions are formatted correctly for display in the application		
		2.2 Data types for predictions match the expected input format of the application		
3	Error Handling	3.1 Clear error messages are returned for invalid user input		

		3.2 Server errors are handled effectively and communicated to users		
		3.3 Logging error occurrences is implemented for debugging		
4	Testing	4.1 Thorough testing of the API integration completed		
		4.2 Invalid inputs scenario is tested		
		4.3 Valid inputs scenario is tested		
		4.4 Performance under load is monitored and meets acceptable response times		

Learning outcome 3 end assessment:

Theoretical assessment

I. **Read the following statements concerned to the machine learning deployment and circle the letter corresponding to the right answer**

1) The Primary benefit of using a web application for model deployment is?

A) Limited accessibility

B) Real-time interaction with users

C) High maintenance costs

D) Requires extensive hardware

Answer: B) Real-time interaction with users

2) Common format for model files is?

A) CSV

B) XML

C) TensorFlow SavedModel

D) TXT

Answer: TensorFlow SavedModel

3) API stand for?

A) Application Programming Interface

B) Automated Programming Instruction

C) Application Process Integration

D) Advanced Programming Interface

Answer: A) Application Programming Interface

4) Integrating a machine learning model, data format that is commonly used for sending input data to the API is?

A) HTML

B) PDF

C) JSON

D) CSV

Answer: C) JSON

5) Model deployment method which is NOT used in Machine learning is?

A) Real-time inference

B) Batch processing

C) Data scraping

D) Streaming inference

Answer: C) Data scraping

6) Main purpose of tracking API requests is?

A) To increase the application size

B) To monitor user behavior

C) To assess model performance and response times

D) To improve the front-end interface

Answer: C) To assess model performance and response times

7) In which situation would you typically use a mobile app for model deployment?

A) For applications needing heavy computational resources

B) For users requiring offline access and notifications

C) For data-heavy processing tasks

D) For machine-to-machine communication

Answer: B) For users requiring offline access and notifications

8) common HTTP response code for a successful API request is?

A) 404

B) 200

C) 500

D) 400

Answer: B) 200

9) The crucial thing for ensuring the compatibility of the model with the existing system is?

A) Model size

B) Model accuracy

C) Programming language and framework

D) Data collection methods

Answer: C) Programming language and framework

10) Primary goal of integrating the model file into an application is?

A) To generate random data

B) To analyze historical data

C) To provide predictions or insights

D) To store large datasets

Answer: C) To provide predictions or insights

11. Which of the following are common methods for delivering predictions to clients in machine learning?

a) REST APIs
b) Batch Processing
c) Real-time Streaming
d) All of the above

Answer: d) All of the above

12. Which method is typically used for large volumes of data that do not require immediate predictions?

a) Real-time Streaming
b) Batch Processing
c) REST APIs
d) Cloud Functions

Answer: b) Batch Processing

13. Which method is best suited for applications where clients need to interact with the model in real-time, such as chatbots or recommendation systems?

a) Batch Processing
b) Real-time Streaming
c) REST APIs
d) Email

Answer: c) REST APIs

14. Which of the following factors should be considered when choosing a prediction delivery method?

a) Latency requirements
b) Data volume
c) Security and privacy
d) Cost
e) All of the above

Answer: e) All of the above

19. What is the role of APIs in delivering machine learning predictions?

a) APIs enable clients to interact with the model programmatically.
b) APIs are only used for batch processing.
c) APIs are not relevant for real-time predictions.
d) APIs are mainly used for internal model management.

Answer: a) APIs enable clients to interact with the model programmatically.

II. Read the following terms as used in machine learning model deployment Match the terms in Column A with their corresponding descriptions in Column B and provide the answer in appropriate place.

Answer	Column A	Column B
1.....**B**	1. What format is commonly used to save and integrate machine learning models?	A. Enables communication between the model and external applications.
2.....**A**	2. What is the role of APIs in model integration?	B. h5, .pkl, .joblib, and .onnx formats
3.....**D**	3. What library is commonly used to save Python-based ML models?	C. ONNX
4.....**C**	4. Which framework supports saving models as .onnx for cross-platform use?	D. joblib and pickle
5.....**G**	5. What is the purpose of serialization in model integration?	E. Flask or FastAPI
6.....**E**	6. What tool is commonly used to serve machine learning models as REST APIs?	F. Loading the saved model back into memory for inference.
7.....**F**	7. What does model deserialization involve?	G. Converts a model into a file or byte stream for storage and transfer.
8.....**I**	8. What environment factors must be considered during integration?	H. Docker

9.....**H**	9. How can a TensorFlow model be saved for integration?	I. Python version, library versions, and dependencies.
10.....**J**	10. What platform is widely used for deploying containerized ML models?	J. Using the model.save () function to create .h5 or SavedModel files.

III. Read the following statements related to the deployment of machine learning and answer by t true if the statement is correct or false if the statement is incorrect.

1) The main advantage of using a standalone program for model deployment is its ease of integration with existing web applications.

Answer: False

2) Scikit-Learn is a commonly used framework for deploying machine learning models.

Answer: True

3) Monitoring performance involves tracking API requests, response times, and model accuracy.

Answer: True

4) Batch processing is the only method available for deploying machine learning models.

Answer: False

5) Proper formatting of predictions is unnecessary if the model is accurate.

Answer: False

6) Cloud platforms offer high scalability and flexibility for deploying machine learning models.

Answer: True

7) On-premise deployment provides the highest level of control over the model and its environment.

Answer: True

8) Edge devices are typically used for high-latency applications where real-time processing is critical.

Answer: True

9) Cost is a significant factor to consider when choosing a deployment method.

Answer: True

10) Security is not a major concern when deploying machine learning models in the cloud.

Answer: False

Practical assessment

ACCENTURE is top data analytics consulting firm that works with clients from 40+ industries. You are hired to deploy a machine learning model to provide real-time predictions for patient health outcomes. In this application-based learning scenario, select a suitable model from Scikit-Learn, TensorFlow, or PyTorch, such as one that predicts the likelihood of hospital readmission or disease progression based on patient data.

Tasks:

1. Include the required dataset characteristics,
2. Chose the best application type and technology stack
3. Design and implement an API using a web framework of their choice,
4. Generate predictions, and return well-formatted responses,
5. Implement strategies for monitoring performance in healthcare applications.

Checklist:

SN	Criteria	Indicators	Observation	
			Yes	No
1	Model Selection is well done	1.1 Model is chosen from Scikit-Learn, TensorFlow, or PyTorch		
		1.2 Model type is identified (e.g., Logistic Regression, Neural Network)		
		1.3 Model is trained on dataset		
		1.4 Model is evaluated on dataset		
2	Dataset Characteristics are well prepared	2.1 Dataset size and format are specified		
		2.2 Data includes relevant features (e.g., demographics, health metrics)		
		2.3 Data are pre-processed		
		2.4. Data are cleaned		
		2.5 dataset are used to train a model		

		2.6 dataset are used to evaluate a model		
3	System Requirements is well chosen	3.1 Application type is identified (web app, mobile app, standalone)		
		3.2 Technology stack are specified (e.g., Flask, Django, FastAPI)		
		Infrastructure requirements are defined (server specs, cloud services)		
4	Model Specifications is well done	4.1 Model format is specified (e.g., Scikit-Learn, TensorFlow SavedModel, ONNX)		
		4.2 Model performance metrics are documented (accuracy, F1-score, etc.)		
		4.3 Memory and compute requirements are estimated		
5	API is well Designed	5.1 API framework is selected (e.g., Flask, FastAPI)		
		5.2 Endpoints are designed (e.g., /predict, /status)		
		5.3 Input data format is specified (e.g., JSON, form data)		
6	Implementation	6.1 API is implemented		
		6.2 API is tested locally		
		6.3 Error handling mechanisms are incorporated (e.g., try-except blocks)		
		6.4 Logging is implemented for error tracking		
7	Model is well Tested and Validated	7.1 Unit tests are written for API endpoints		
		7.2 End-to-end tests are conducted with sample data		
		7.3 Performance testing is conducted (latency, throughput)		
8	Model is well Deployed	8.1 Deployment environment is configured (e.g., Docker, cloud service)		

		8.2 API is deployed to production environment		
		8.3 API is accessible from external clients		
9	Model is well Monitored and Maintained	9.1 Performance metrics are tracked (response times, prediction accuracy)		
		9.2 Monitoring tools are set up (e.g., Prometheus, Grafana)		
		9.3 Error tracking is in place (e.g., Sentry, ELK stack)		
		9.4 Documentation is created for API usage and troubleshooting		
10	Reflection	10.1 Challenges faced during deployment are documented		
		10.2 Solutions to challenges are identified and recorded		
		10.3 Lessons learned and future improvement strategies are documented		

END

Further information to the trainer

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. *IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. https://doi.org/10.1109/ICSE-SEIP.2019.00042

Kraska, T., Beutel, A., Chi, E. H., & Dean, J. (2020). The case for learned index structures. *Communications of the ACM, 63*(12), 105–113. https://doi.org/10.1145/3380971

Olston, C., Fiedel, N., Gorovoy, K., Harmsen, J., Lao, L., Ramesh, S., & Soergel, D. (2017). TensorFlow-Serving: Flexible, high-performance ML serving. *arXiv preprint arXiv:1712.06139*. https://arxiv.org/abs/1712.06139

Raschka, S., & Mirjalili, V. (2017). Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2 (2nd ed.). Packt Publishing.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., & Young, M. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511. https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems

Zhang, Z., Lipton, Z. C., Li, M., & Smola, A. J. (2020). Dive into deep learning. *MIT Press*. https://d2l.ai

RTB | RWANDA TVET BOARD

October 2024