VAUTECH SOLUTIONS — ML INTERNSHIP REPORT

Project Title:
Machine Learning Problem Scoping & Learning Lifecycle

Intern Name:

Vishal Ramesh Taduka

Intern ID:

VT26ML004

Domain:

Machine Learning

Mentor:
Vishal Ramkumar Rajbhar

Company:
Vautech Solutions IT Solutions

ABSTRACT

Machine Learning is a rapidly growing field of Artificial Intelligence that enables machines to learn from data and improve performance without explicit programming. This report focuses on understanding the **Machine Learning lifecycle** and the importance of **problem scoping** in real-world Machine Learning applications.

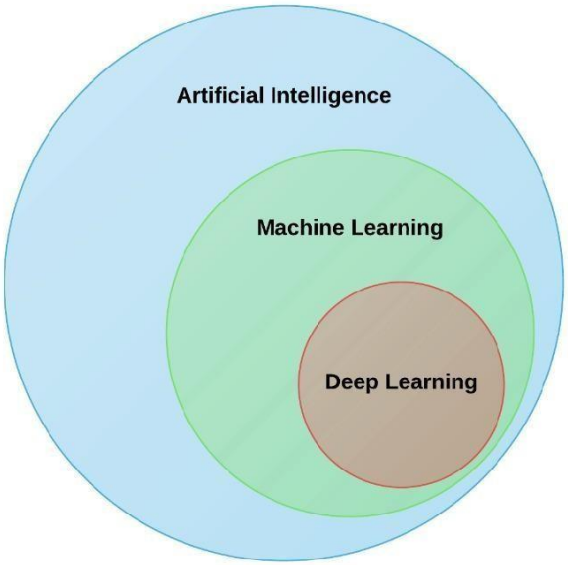
The main objective of this task is to understand how Machine Learning problems are defined, how data is identified, and how the correct learning approach such as classification or regression is chosen. This report explains the complete ML lifecycle starting from problem definition to model improvement. It also highlights real-world use cases and explains how input features and target variables are defined.

This task provides a strong conceptual foundation required for future practical Machine Learning implementations.

1. INTRODUCTION

Machine Learning is a part of Artificial Intelligence where computers learn from data and use that learning to make predictions or decisions. Today, Machine Learning is used in many areas like suggesting movies or products, detecting fraud, helping doctors find diseases, recognizing images, and predicting prices.

In simple terms, problem scoping helps us choose the right Machine Learning method so that we get accurate and useful results.



2. OBJECTIVES OF THE TASK

The main objectives of this task are:

- To understand the complete Machine Learning lifecycle
- To learn how ML problems are framed in real-world scenarios
- To identify real-world ML use cases
- To decide whether a problem is classification or regression
- To define input features and target variables clearly

3. MACHINE LEARNING OVERVIEW

Machine Learning is a part of Artificial Intelligence where computers learn from data and use that learning to make predictions or decisions.

Machine Learning allows systems to automatically learn patterns from data. Instead of manually writing rules.

Machine Learning is mainly used when:

- The problem is complex
- Large amounts of data are available
- Rules cannot be step-by-step defined

4. DATASET DESCRIPTION (KAGGLE – VEGETABLES DATASET)

The **Vegetables Dataset** contains detailed information about different vegetables.

Dataset Size

- **Rows:** 150
- **Columns:** 15

Important Columns Used

Column Name	Description
Category	Type of vegetable (Root, Leafy, Fruit, etc.)
Color	Color of vegetable
Season	Best growing season
Shelf Life (days)	Storage life
Price (per kg)	Cost of vegetable

5.MACHINE LEARNING LIFECYCLE

The Machine Learning lifecycle represents the steps followed while developing an ML solution.

5.1 Problem Definition

In this step, the problem is clearly defined by understanding:

- What is the objective of problem?
- Why ML is required
- What will be the expected output

Example: Predicting house prices based on features.

5.2 Data Collection

Data is collected from multiple sources such as:

- Databases
- CSV files

The success of an ML model depends heavily on the quality of data.

5.3 Data Preprocessing

Raw data is cleaned and prepared by:

- Handling missing values
- Removing duplicate data

5.4 Model Selection and Training

- Based on the problem type, an algorithm is selected.
- The model training is a process of teaching the model using existing data.
- Once the training is completed, the model can be used to make predictions on new and unseen data.

5.5 Model Evaluation

Model evaluation is the step where we check how a machine learning model is working after the model is trained it is test using new data. This helps us understand how accurate and reliable the models predictions are :

- Accuracy
- Prediction

5.6 Model Improvement

If the performance is not satisfactory:

- More data is collected
- Features are modified

This cycle continues until desired performance is achieved.

6. Implementation

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
# Load dataset
df = pd.read_csv("vegetables Dataset.csv")
# Select features and target
features = ['Color', 'Season', 'Shelf Life (days)']
target = 'Category'
# Encode categorical columns
encoder = LabelEncoder()
for col in features + [target]:
    df[col] = encoder.fit_transform(df[col])
X = df[features]
y = df[target]
# Split dataset
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42
)
# Train model
model = RandomForestClassifier()
model.fit(X_train, y_train)
# Prediction
y_pred = model.predict(X_test)
```

```
# Accuracy

accuracy = accuracy_score(y_test, y_pred)

print("Model Accuracy:", accuracy)
```

8. Output

```
Model Accuracy: 0.6333333333333333
```

9. PROBLEM SCOPING IN MACHINE LEARNING

Problem scoping is the process of understanding the real-world problem and converting it into a Machine Learning problem.

It involves:

- Understanding the problem requirements
- Identifying available data
- Defining inputs and outputs
- Selecting ML type

Proper problem scoping prevents incorrect model selection and poor results.

10. CLASSIFICATION AND REGRESSION

10.1 Classification

Classification problems involve predicting **categorical outputs**.

Examples:

- Email spam detection
- Disease diagnosis

Target Variable: Category or label

10.2 Regression

Regression problems involve predicting **continuous numerical values**.

Examples:

- House price prediction
- Salary prediction
- Temperature forecasting

Target Variable: Numerical value

11. REAL-WORLD USE CASE

Case Study: House Price Prediction • Problem

Type: Regression

- **Input Features:**
    - House area ○ Number of rooms ○ Location ○ Age of property
- **Target Variable:**
    - House price

The goal is to predict house prices based on given features using historical data.

12. TOOLS USED

- Documentation
- Case study analysis
- Conceptual understanding of Machine Learning

13. CONCLUSION

This task helped in understanding the complete Machine Learning lifecycle and the importance of problem scoping. Identifying the correct ML problem type and defining input features and target variables are essential steps before model development. This conceptual knowledge forms a strong base for implementing real-world Machine Learning solutions in future internship tasks.