ERICK WAMBUGU
KODECAMP5X ASSIGNMENT 1

MACHINE LEARNING LIFECYCLE STEPS

1. Problem Definition - Define the business problem: “Predict house prices based on various features like number of rooms, area, and location.” Identify input features (X) and target variable (y).
2. Data Collection - Gather data from reliable sources (e.g., Kaggle’s House Prices dataset, CSV files, or APIs). Example features: location, size, number_of_rooms, bathrooms, year_built, etc.
3. Data Preprocessing & Cleaning - Handle missing values. Convert categorical data (e.g., location) to numerical form (using Label Encoding or One-Hot Encoding). Normalize or scale numerical features. Remove outliers.
4. Exploratory Data Analysis (EDA) - Use visualization tools (Matplotlib, Seaborn, or Pandas profiling). Analyze relationships and correlations between variables (e.g., area vs. price). Detect trends and feature importance.
5. Feature Engineering - Create new features (e.g., price per square meter). Select the most relevant features using correlation or feature importance analysis.
6. Model Selection - Choose appropriate algorithms for regression (e.g., Linear Regression, Decision Tree Regressor, Random Forest Regressor, XGBoost). Split the data into training and testing sets (e.g., 80/20).
7. Model Training - Train the model using the training dataset. Use an appropriate loss function and optimizer.
8. Model Evaluation - Evaluate model performance using metrics such as MAE, MSE, RMSE, and R^2 Score. Fine-tune hyperparameters to improve performance.
9. Model Deployment - Save the model using pickle or joblib. Deploy the model via: Flask or FastAPI web app, or cloud platform like AWS, Google Cloud, or Streamlit.
10. Model Monitoring & Maintenance - Monitor performance on new data. Retrain if accuracy drops or data distribution changes.

ANSWERS TO QUIZ

1. Type of Prediction Problem:

Supervised Learning → Regression Problem: We are predicting continuous values (house prices) and the most likely algorithm:

- Linear Regression (for a simple baseline)
- Random Forest Regressor or XGBoost (for better accuracy)

ERICK WAMBUGU
KODECAMP5X ASSIGNMENT 1

2. Loss Function:

For regression tasks, typical loss functions include:

- Mean Squared Error (MSE)

3. Optimization Algorithm:

Used to minimize the loss function during training. Common optimizers include:

- Gradient Descent (standard for Linear Regression)
- Stochastic Gradient Descent (SGD) (for large datasets)
- Adam Optimizer (advanced, adaptive version)

4. Evaluation Metrics:

After training, evaluate performance using:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 Score (Coefficient of Determination)

5. Deployment Approach:

Deploy using Flask or FastAPI API where users can input house features and get predicted price in real time.
Other deployment options:

- Streamlit Web App (quick interactive app)
- Cloud-based deployment using AWS, Google Cloud, or Heroku.