📘 Lesson 4 Assignment— Regression

1. What is Regression?

Regression is a type of supervised learning technique in Machine Learning that predicts continuous numerical outcomes from given input features. It is different from classification, which is used to predict labels or categories.

Example (Regression): Estimating the selling price of a car based on its age, mileage, and brand.
Example (Classification): Predicting whether a car is new, used, or refurbished based on the same features.

📌 Main Difference:

Regression answers: “How much?” or “How many?”

Classification answers: “Which group or class?”

2. Major Types of Regression

a) Simple Linear Regression

Concept: Models the relationship between one independent variable (X) and one dependent variable (y) using a straight line.

Example: Predicting student exam score from hours studied.

Advantages: Very easy to build and interpret.

Drawbacks: Only accurate when the data follows a straight-line pattern.

b) Multiple Linear Regression

Concept: Extends linear regression to include multiple input variables to predict one output.

Example: Predicting crop yield using rainfall, fertilizer used, and soil quality.

Advantages: Works well when many factors affect the target.

Drawbacks: Performance drops if input variables are highly correlated (multicollinearity).

c) Polynomial Regression

Concept: Adds higher powers of input variables (like x^2 x^3) to capture curved relationships between features and target.

Example: Predicting the growth of a plant over time using both time and time² as features.

Advantages: Good for modeling non-linear trends.

Drawbacks: Higher polynomial degrees can lead to overfitting.

📊 Comparison of Regression Types

Regression Type	Relationship Shape	Example Features	Best Used When
Simple Linear Regression	Straight line	Hours studied Exam score	One main influencing factor
Multiple Linear Regression	Flat multi-dimensional plane	Rainfall, Fertilizer, Soil → Yield	Several features influence outcome
Polynomial Regression	Curved line	Time, Time² → Plant growth	Trend is non-linear

3. Measuring Regression Performance

To know how good a regression model is we use evaluation metrics that compare predicted values to actual values:

Metric Name	What It Measures	Sensitive to Large Errors	Units
MAE (Mean Absolute Error)	Average size of prediction errors	❌ No	Same as target
MSE (Mean Squared Error)	Average of squared errors	✅ Yes	Squared units
RMSE (Root Mean Squared Error)	Square root of MSE	✅ Yes	Same as target
R^2 (Coefficient of Determination)	Percentage of variance explained	❌ No	0–1 (or %)

Example:

RMSE = 3.5 → model's predictions are off by about 3.5 units on average.

R^2 = 0.92 → model explains 92% of the variation in the target values.

4. Underfitting vs Overfitting

Underfitting: The model is too basic and fails to detect patterns → low accuracy on both training and test sets.

Overfitting: The model fits the training data too closely and performs poorly on unseen data.

Reasons for Overfitting:

- Overly complex models
- Very small training dataset

- Including irrelevant features

Ways to Prevent It:

- Simplify the model
- Use regularization methods (like Lasso or Ridge)
- Apply cross-validation
- Increase dataset size
- Stop training when validation accuracy stops improving

5. Case Study — Regression in Healthcare

Title: “Predicting Patient Hospital Stay Duration Using Multiple Linear Regression”

(Published in Health Informatics Journal, 2022)

Objective: Estimate how many days a patient will stay in the hospital based on their age, illness severity, and treatment type.

Dataset: Records from 40,000 patients, including age, diagnosis category, treatment plan, and previous medical history.

Model Used: Multiple Linear Regression

Key Outcomes:

$R^2 \approx 0.80$ and RMSE ≈ 1.2 days

Illness severity and age were the strongest predictors

Helped hospitals plan bed availability and staffing

Takeaway: Regression can help improve resource planning and efficiency in healthcare services.

References

Alpaydin, E. (2020). Introduction to Machine Learning. MIT Press.

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly.

"Predicting Patient Hospital Stay Duration Using Multiple Linear Regression." Health Informatics Journal, 2022.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.