Recitation for Assignment 4

November 13, 2025

Main Objectives

- Apply advanced EDA and feature engineering techniques.
- Train and tune multiple regression models.
- Compare model performance using Regression metrics
- Interpret feature importance.
- Understand model complexity

Question 1

Steps:

- Explain why it is crucial to first analyze data distributions and temporal patterns before implementing machine learning models
- Load the dataset and provide the structural metadata and summary statistics
- Report the data types of the variables and comment on the results
- Average the yearly life expectancy for developing and developed countries
- Plot the yearly expectancy trends from 2000 to 2015 and comment on the trends and anomalies

Question 2

Steps:

- Briefly explain how outliers can influence regression model performance and bias the results
- Determine missing value proportions per variable
- Choose one variable with significant number of missing values, fill the missing values with median, then fill the missing values again with KNN imputation
- Compare the filled variables' mean, standard deviation and data distribution
- Detect the outliers using IQR or Z-score in GDP and Percentage Expenditure
- Provide the boxplots before and after handling outliers

Question 3

Steps:

- Create at least three new features using feature engineering techniques (e.g., combining related variables, creating ratios, aggregating similar indicators)

  Example: Infant Survival Rate, Vaccination Coverage Index, Education Income Index

- Select five independent variables believed to influence life expectancy
- Compare their distributions between Developed and Developing countries using appropriate plots
- Create a correlation heatmap to visualize relationships among variables
- Identify the strongest positive and negative correlations with life expectancy
- Interpret the observed relationships and explain why correlation does not necessarily imply causation in predictive modeling

Question 4

Steps:

- Encode categorical variables appropriately (e.g., Status: Developed/Developing)
- Apply feature transformation and scaling where required using StandardScaler
- Explain why certain models (such as KNN or Gradient Boosting) are sensitive to scale differences

Question 5

Steps:

- Split the dataset into training and testing sets (use 80/20 or 70/30 ratio)
- Train and tune four models: Decision Tree Regressor, Random Forest Regressor, KNN Regressor, and Gradient Boosting Regressor
- Use GridSearchCV with appropriate hyperparameter grids for each model
- Evaluate performance using RMSE, MAE, and R^2 metrics on the test set
- Report the best hyperparameters for each model
- Explain the role of hyperparameter tuning in controlling overfitting and underfitting

Question 6

Steps:

- Compute feature importance from the best-performing model
- Select the top features (e.g., top 5-10) based on importance scores
- Retrain all four models using only the selected features with their best hyperparameters
- Evaluate their performance on the test set using RMSE, MAE, and R^2
- Compare the results with the original full-feature model in a table
- Discuss the trade-offs between model simplicity, interpretability, and predictive accuracy

Question 7

Steps:

- Compare model-derived feature importance (from Q6) with simple correlation values (from Q3)
- Explain why they may not always align
- Extract the top five most important features from the best-performing model
- Discuss their real-world relevance in explaining differences in life expectancy

Logistics

- Make sure you submit all the files required. Missing codefile, datafiles or report may attract penalties.
- The accepted format for the report is a **.pdf**
- Make sure the codefile submitted in the zip file runs without errors.
- Make sure you load the data from the correct path!
- Do **not** include **code** in the report.
- Do not create other folders in your submission folder
- Check your submission before you submit otherwise you might submit the wrong files

Report

- Cover page
- Should explain the procedure and clear inferences from the results
- Avoid adding screenshots of your codes in the report
- Should be concise and coherent