Machine Learning Assignments

1. Data Analysis Assignments

Dataset: Netflix Movies and TV Shows (https://www.kaggle.com/datasets/shivamb/netflix-shows)

Q1: Data Cleaning and Preprocessing

- Check and handle missing values.
- Convert date fields (like `date_added`) to datetime.
- Extract the year and month from `date_added`.

Q2: Exploratory Data Analysis (EDA)

- Count of movies vs TV shows.
- Top 5 countries by number of Netflix titles.
- Distribution of content ratings (TV-MA, PG-13, etc).

Q3: Pattern Discovery and Grouping

- Group by `release_year` to show content volume over time.
- What genres (`listed_in`) are most popular in the U.S.?
- Find the average duration of shows per genre.

2. Linear Regression Assignments

Dataset: Students Performance (https://www.kaggle.com/datasets/spscientist/students-performance-in-exams)

Q1: Simple Linear Regression

- Use `hours studied` to predict `math score`.

- Build a simple linear regression model.

- Report R^2 and Mean Squared Error.

- Visualize the regression line.

Q2: Multiple Linear Regression

- Use `reading score`, `writing score`, and `hours studied` to predict `math score`.

- Evaluate the model using RMSE and R^2.

- Interpret coefficients.

Q3: Polynomial Regression

- Fit a polynomial regression model (degree 2 or 3) to predict `math score` using `reading score`.

- Compare performance with linear regression.

3. Logistic Regression Assignments

Dataset: Heart Disease UCI (https://www.kaggle.com/datasets/ronitf/heart-disease-uci)

Q1: Binary Logistic Regression

- Predict heart disease (`target` = 0 or 1) using selected features.

- Train logistic regression and evaluate with confusion matrix and accuracy.

Q2: ROC Curve and Evaluation

- Plot the ROC curve and calculate AUC.

- Report precision, recall, and F1-score.

Q3: Feature Impact

- Print and interpret model coefficients.

- Identify which 3 features most increase the probability of heart disease.