**Q(A). Assignment Title: Predicting California Housing Prices Using Linear Regression and Regularization Techniques**

**Objective:**  
The objective of this assignment is to develop predictive models for housing prices using the California Housing Dataset. You will perform exploratory data analysis (EDA), build regression models, evaluate their performance, and apply regularization techniques (Ridge and Lasso). Additionally, you will design and engineer a custom feature to assess its impact on model performance.

**Dataset:**  
Download the dataset named QA.

**Tasks and Instructions:**

Q1. Dataset Loading (5 marks)

* Load the dataset
* Convert it to a DataFrame
* Display the first few rows.

Q2. Exploratory Data Analysis (10 marks)

* Display summary statistics.
* Visualize relationships:  
  • Scatter plots between selected features and the target (MedHouseVal)  
  • Correlation heatmap  
  • Distribution of the target variable

Q3. Data Preprocessing (10 marks)

* Normalize all numeric features
* Check for missing values or outliers
* Split the dataset into training and testing sets (80:20)

Q4. Model Development: Linear Regression (10 marks)

* Train a LinearRegression model
* Report the model coefficients and intercept
* Summarize key insights from the learned weights

Q5. Model Evaluation (15 marks)

* Evaluate the model on the test set using: R² score, MAE, and RMSE
* Plot predicted vs actual values
* Visualize residuals (prediction errors)

Q6. Model Interpretation (10 marks)

* Identify which features have the most positive and negative impact on the target
* Discuss whether the model shows signs of underfitting or overfitting

Q7. Regularization Models: Ridge & Lasso (15 marks)

* Train Ridge and Lasso regression models using cross-validation
* Compare their performance with the Linear Regression model using R², MAE, and RMSE
* Analyze the difference in model complexity (e.g., number of non-zero coefficients)

Q8. Feature Engineering (10 marks)

* Create at least one new feature derived from existing columns (e.g., population per household)
* Justify why this feature could improve the model

Q9. Model Retraining with New Feature (5 marks)

* Retrain any one model (Linear, Ridge, or Lasso) using the engineered feature
* Report the updated performance

Q10. Final Evaluation (5 marks)

* Compare model performance before and after adding the new feature
* Provide a concise explanation for the change in performance

Q11**.** (Bonus) Try Polynomial Regression, interaction terms, or recursive feature elimination (RFE), and discuss any observed performance improvement.

**Grading Breakdown:**  
Q1 – 5 marks  
Q2 – 10 marks  
Q3 – 10 marks  
Q4 – 10 marks  
Q5 – 15 marks  
Q6 – 10 marks  
Q7 – 15 marks  
Q8 – 10 marks  
Q9 – 5 marks  
Q10 – 5 marks  
Q11(Bonus) – 5 marks  
**Total – 100 marks**

**Q(B). Assignment Title: Predicting Income Levels Using Logistic Regression**  
**Objective:**  
The objective of this assignment is to develop a binary classification model to predict whether an individual earns more than $50,000 per year based on demographic and employment-related features from the Income dataset. The task involves data preprocessing (including one-hot encoding), logistic regression model training, performance evaluation, and interpretability analysis.

**Dataset:**

* Download the file QB.
* Target column: income (binary: <=50K or >50K)
* Features include both numerical (age, hours-per-week, education-num) and categorical (education, sex, occupation, etc.) variables. Several categorical features require one-hot encoding

**Assignment Tasks:**

Q1. Data Loading and Exploration (5 marks)

* Load the dataset into a DataFrame.
* Display the first few rows, structure, and summary statistics.
* Identify categorical and numerical features.
* Check for missing values or entries with ‘?’.

Q2. Data Preprocessing and One-Hot Encoding (20 marks)

* Replace all ‘?’ with NaN and handle missing values appropriately.
* Apply one-hot encoding to all categorical columns.
* Normalize numerical features.
* Encode the target column (income) into binary values: <=50K → 0, >50K → 1.
* Split the dataset into training (80%) and testing (20%) sets.

Q3. Model Building: Logistic Regression (15 marks)

* Train a logistic regression model using sklearn.linear\_model.LogisticRegression.
* Print the model’s coefficients and intercept.
* Identify top features influencing the prediction.

Q4. Model Evaluation (25 marks)

* Evaluate the model using the following metrics: Accuracy, Confusion Matrix, Precision, Recall, F1-score, ROC curve, and AUC score.
* Plot the confusion matrix and ROC curve using appropriate visualization libraries.

Q5. Interpretation and Reflection (20 marks)

* Interpret the most influential features using the model coefficients.
* Discuss if the model shows potential bias or fairness concerns across gender, race, or other sensitive attributes.
* Highlight any limitations or challenges in your modeling approach and suggest improvements.

Q6. (Bonus)

* Train and evaluate another classification algorithm (e.g., Decision Tree, Random Forest, or SVM).
* Compare its performance with logistic regression.
* Justify which model is better suited for deployment and why.

**Grading Rubric:**  
Q1: 5 marks  
Q2: 20 marks  
Q3: 15 marks  
Q4: 25 marks  
Q5: 20 marks  
Q6 (Bonus): 5 marks

**Total: 85 marks**

**Submission Guidelines:**

Each student is required to submit two files as part of this assignment in a single zipped folder named YourName\_Rollno.zip. The first file should be a Jupyter Notebook (YourName\_RollNo.ipynb) that contains all code implementations for the tasks outlined in the assignment. The notebook must include clear markdown explanations, properly labeled plots, clean outputs, and modular, well-commented code. The second file should be a written report in PDF format (YourName\_RollNo.pdf) that provides a concise and structured summary of the work. This report must describe the preprocessing steps, modeling approach, and conclusions, and include relevant plots such as the confusion matrix, ROC curve, and feature importance charts, each accompanied by a brief explanation. The report should be professionally presented and clearly communicate the key insights and results of the analysis.