Week 3 Report: Supervised Learning - Regression Analysis

1. Project Overview:

The objective of this project was to predict housing prices using the California Housing dataset. I implemented and compared four distinct regression models: **Linear Regression, Polynomial Regression, Ridge, and Lasso**. The models were evaluated based on their ability to minimize prediction error and maintain stability across the dataset.

2. Methodology:

1. **Data Selection:** I utilized the California Housing dataset, which includes features like median income, house age, and average rooms.
2. **Preprocessing:** The data was split into a **training set (80%)** to build the models and a **testing set (20%)** to verify their accuracy on unseen data.
3. **Model Implementation:** * **Linear Regression:** Used as a baseline to find the best-fit straight line.
    a. **Polynomial Regression:** Implemented to capture non-linear relationships between variables.
    b. **Regularization (Ridge/Lasso):** Applied to prevent "overfitting" by penalizing large coefficients, ensuring the model generalizes well to new data.

3. Performance Metrics:

I used three primary metrics to evaluate the models:

1. **MAE (Mean Absolute Error):** Represents the average dollar amount the predictions were off.
2. **MSE (Mean Squared Error):** Measures the average squared difference between estimated and actual values; lower values indicate better fit.
3. **R2 Score:** Indicates the percentage of variance in house prices explained by the features. A higher score closer to 1.0 is preferred.

4. Analytical Findings & Insights:

- **Model Accuracy:** The **Linear Regression** and **Ridge** models typically show high consistency in this dataset, often achieving an **R2 score** between **0.60 and 0.70**.
- **Overfitting Prevention: Lasso Regression** proved useful for feature selection, as it effectively ignored less important variables by setting their coefficients to zero.

- **Visualization Analysis:** The "Actual vs. Predicted" scatter plot demonstrates a strong positive correlation, though some variance exists at higher price points (the "ceiling effect" where prices are capped).

5. Conclusion:

Based on the results, **Ridge Regression** is suggested as the most stable model for this dataset because it handles multi-collinearity (when features are related) better than standard Linear Regression. For future improvements, incorporating more non-linear features or using Ensemble methods could further increase the R2 score.