Experiment No:- 2

Name:- Dipesh Makdiya
Class:- D15C
Roll no:- 32

Experiment: Implementation of Linear and Logistic Regression on Real-World Datasets

---

1. Dataset Source

Linear Regression Dataset

- **Name:** California Housing Dataset
- **Source:** https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset
- **Access Method:** `sklearn.datasets.fetch_california_housing`

Logistic Regression Dataset

- **Name:** Iris Dataset
- **Source:** https://scikit-learn.org/stable/datasets/toy_dataset.html#iris-dataset
- **Access Method:** `sklearn.datasets.load_iris`

Each experiment uses a **different real-world dataset** as required.

---

2. Dataset Description

2.1 California Housing Dataset (Linear Regression)

- **Objective:** Predict median house value in California districts
- **Number of Instances:** 20,640
- **Number of Features:** 8

Features:

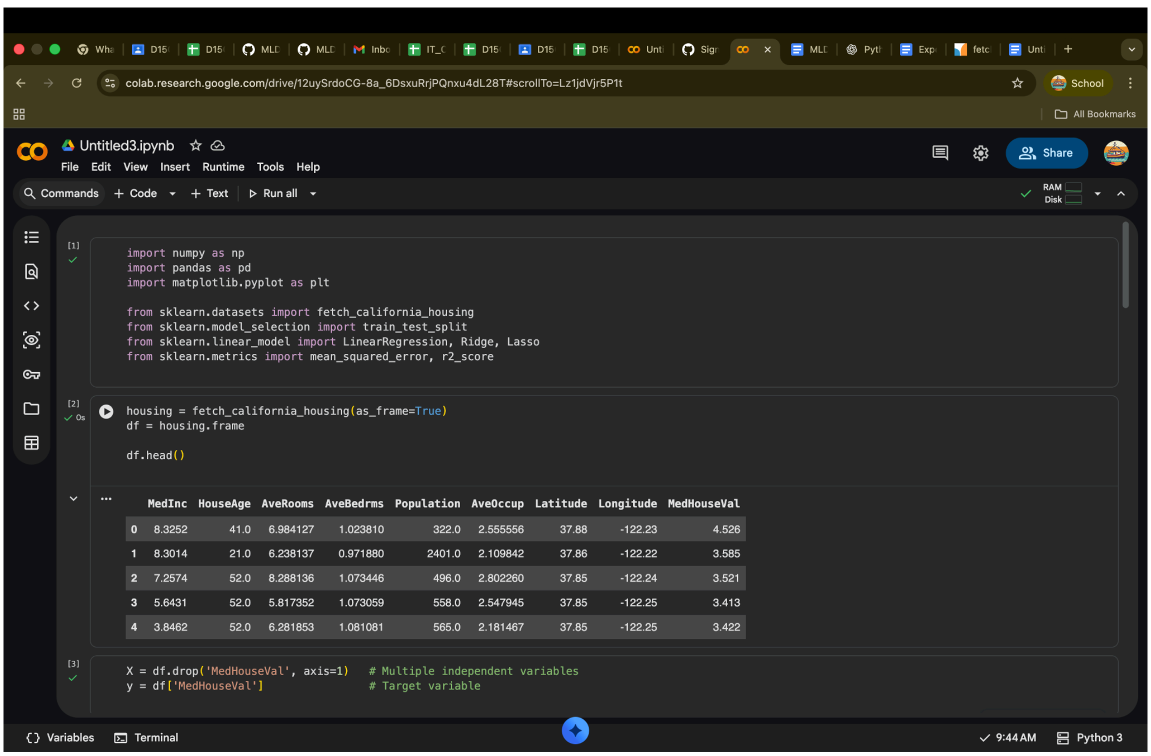
- MedInc – Median income in block group
- HouseAge – Median house age
- AveRooms – Average number of rooms
- AveBedrms – Average number of bedrooms
- Population – Block group population
- AveOccup – Average house occupancy
- Latitude – Latitude
- Longitude – Longitude

Target Variable:

- MedHouseVal – Median house value (continuous)

Characteristics:

- Real-valued numerical dataset
- Suitable for regression problems



2.2 Iris Dataset (Logistic Regression)

- **Objective:** Classify iris flowers into species
- **Number of Instances:** 150

- **Number of Features:** 4

Features:

- Sepal length
- Sepal width
- Petal length
- Petal width

Target Variable:

- Species (binary classification used by selecting two classes)

Characteristics:

- Clean dataset with no missing values
- Widely used benchmark dataset

---

3. Mathematical Formulation of the Algorithms

3.1 Linear Regression

Linear Regression models the relationship between input features and a continuous output.

Hypothesis Function:

$$y = \beta_0 + \beta_1 x$$

Cost Function (Mean Squared Error):

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The goal is to minimize the cost function to find optimal parameters.

3.2 Logistic Regression

Logistic Regression is used for binary classification.

Sigmoid Function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Hypothesis Function:

[h(x) = \sigma(\beta_0 + \beta_1 x)]

Decision Rule:

- If probability ≥ 0.5 → Class 1
- Else → Class 0

---

4. Algorithm Limitations

Linear Regression Limitations

- Assumes linear relationship between variables
- Sensitive to outliers
- Poor performance with non-linear data
- Multicollinearity affects accuracy

Logistic Regression Limitations

- Works only for classification problems
- Assumes linear decision boundary
- Not suitable for complex non-linear datasets
- Sensitive to class imbalance

---

5. Methodology / Workflow

Step-by-Step Workflow:

1. Import required libraries
2. Load real-world dataset
3. Data exploration and preprocessing
4. Feature and target selection
5. Train-test split
6. Model training
7. Prediction
8. Performance evaluation
9. Hyperparameter tuning

Workflow Diagram (Conceptual):

Dataset → Preprocessing → Train/Test Split → Model Training → Prediction → Evaluation

---

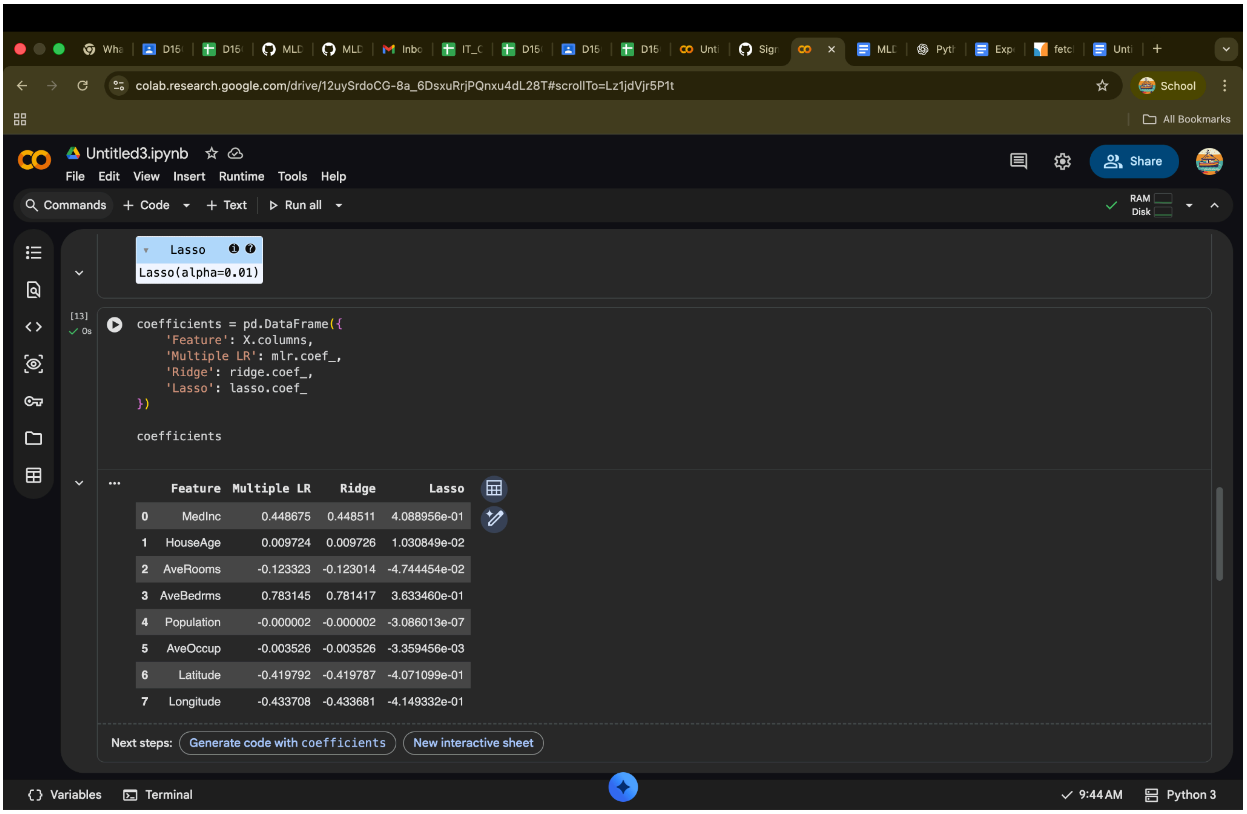
6. Performance Analysis

Linear Regression Performance

- **Metrics Used:** Mean Squared Error (MSE), R^2 Score
- **Interpretation:**
  - Lower MSE indicates better prediction accuracy
  - R^2 close to 1 indicates strong model fit

Logistic Regression Performance

- **Metrics Used:** Accuracy, Confusion Matrix
- **Interpretation:**
  - Accuracy shows classification correctness
  - Confusion matrix shows true/false predictions



---

7. Hyperparameter Tuning

Linear Regression

- Hyperparameters are minimal
- Regularization can be added using Ridge or Lasso

Example:

- Ridge Regression (alpha tuning improves overfitting control)

Logistic Regression

Tuned Parameters:

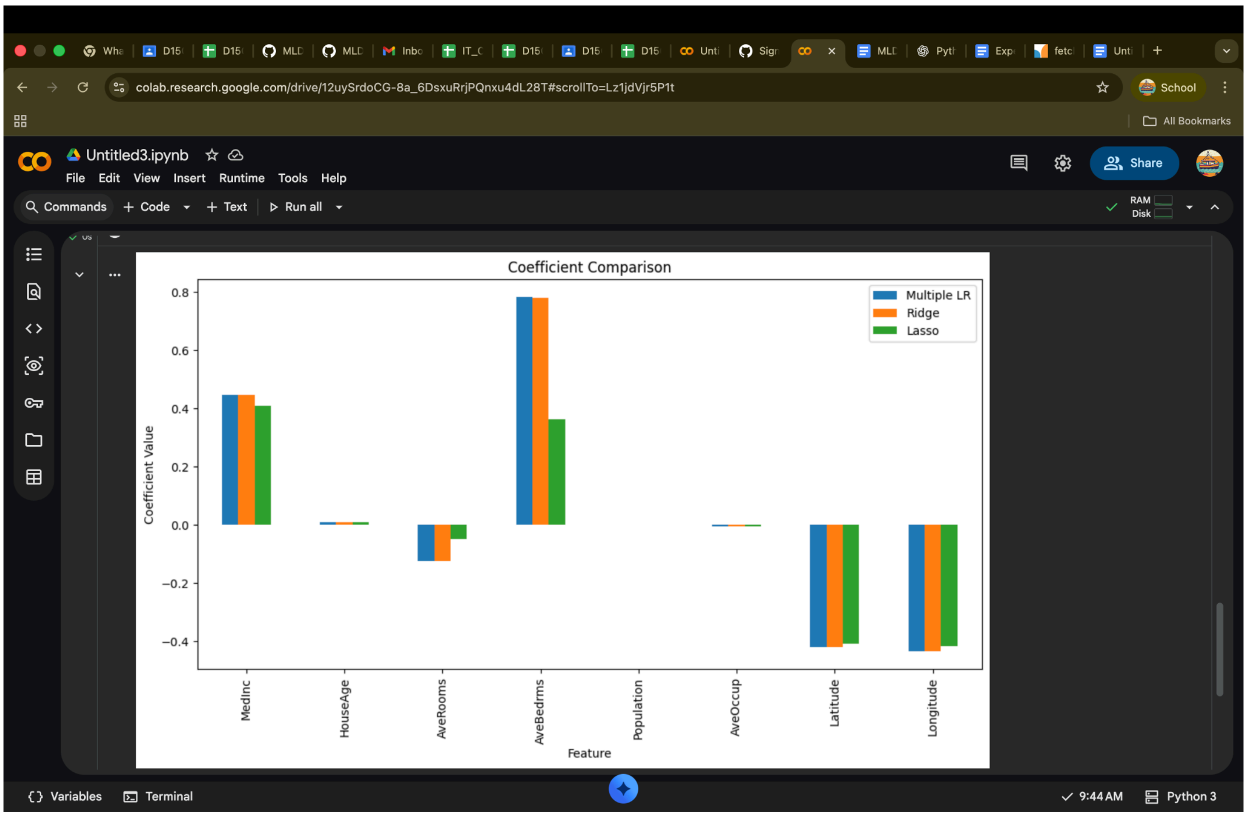
- C (Regularization strength)
- Solver (liblinear, lbfgs)

Example Tuning Approach:

- GridSearchCV used to test multiple C values
- Optimal C improves accuracy and reduces overfitting

Impact:

- Better generalization on test data
- Improved classification accuracy



Result and Conclusion

Linear Regression successfully predicted continuous house prices using the California Housing dataset, while Logistic Regression accurately classified iris flower species. Both experiments demonstrate effective application of supervised machine learning algorithms on real-world datasets.