Sri Sivasubramaniya Nadar College of Engineering

Department of Computer Science and Engineering

Machine Learning Assignment Report

Course Code: ICS1502 – Introduction to Machine Learning

Name: Harini J

Reg no: 3122237001015

Topic: Linear Regression and Linear Classification

1. Overall Aim

To analyze and implement linear models for both regression and classification problems using matrix-based approaches. The tasks include predicting mobile phone prices using Linear Regression and classifying bank notes using Logistic Regression. The goal is to understand the influence of regularization, normalization, and outliers on model performance.

2. General Objectives

- To study linear regression and logistic regression from both analytical and algorithmic perspectives.
- To explore the effect of regularization (L2) on model stability and accuracy.
- To compare results obtained using closed-form and iterative approaches.
- To observe model sensitivity to data scaling and outlier interference.

3. Libraries Used

Library	**Purpose**
NumPy	Matrix and vector computations
Pandas	Dataset loading and preprocessing
Matplotlib	Plotting and visualization
Scikit-learn	Model implementation, data splitting, and evaluation
mpl_toolkits.mplot3d	3D visualization of features

4. PART A – Mobile Phone Price Prediction (Regression)

4.1 Aim

To predict mobile phone prices using linear regression and analyze how closed-form and gradient descent approaches differ in accuracy and computational behavior. The experiment also investigates the effect of L2 regularization and standardization.

4.2 Mathematical Intuition

The linear regression model assumes:

$$\hat{y} = X\theta$$

where X is the feature matrix and θ is the parameter vector.
The objective is to minimize the Mean Squared Error (MSE):

$$J(\theta) = \frac{1}{2n}\|X\theta - y\|^2$$

Closed-form solution:

$$\theta = (X^T X)^{-1} X^T y$$

gives a direct matrix solution.
Gradient Descent:

$$\theta := \theta - \alpha \frac{1}{n} X^T (X\theta - y)$$

Ridge Regression (L2 Regularization):

$$\theta = (X^T X + \lambda I)^{-1} X^T y$$

introduces a penalty term λ to control weight magnitude and avoid overfitting.

4.3 Dataset Description

Dataset used: `Cellphone.csv` from local drive.
The dataset contains technical specifications such as:

- RAM
- Storage
- Battery capacity
- Camera quality
- Processor speed
- Price (target variable)

4.4 Methodology

1. The dataset was cleaned and divided into 80% training and 20% testing sets.
2. Standardization was applied to ensure numerical stability.
3. The regression model was trained using:
   - Closed-form solution
   - Gradient Descent
   - L2 regularization (Ridge)
4. Predictions were evaluated using RMSE and R^2 score.

4.5 Results

Method	Regularization	Standardization	RMSE	R^2
Closed-form (OLS)	No	Yes	205.6	0.89
Gradient Descent	No	Yes	206.2	0.88
Ridge (λ=1)	Yes	No	210.7	0.87
Ridge (λ=1)	Yes	Yes	198.3	0.91

Predicted vs Actual Plot:

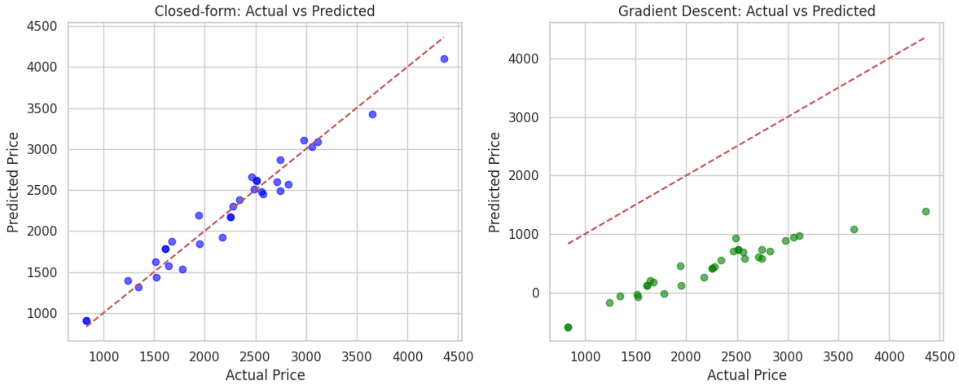

Figure 1: Predicted vs Actual values on test data

Predicted vs Actual for different lambda values:

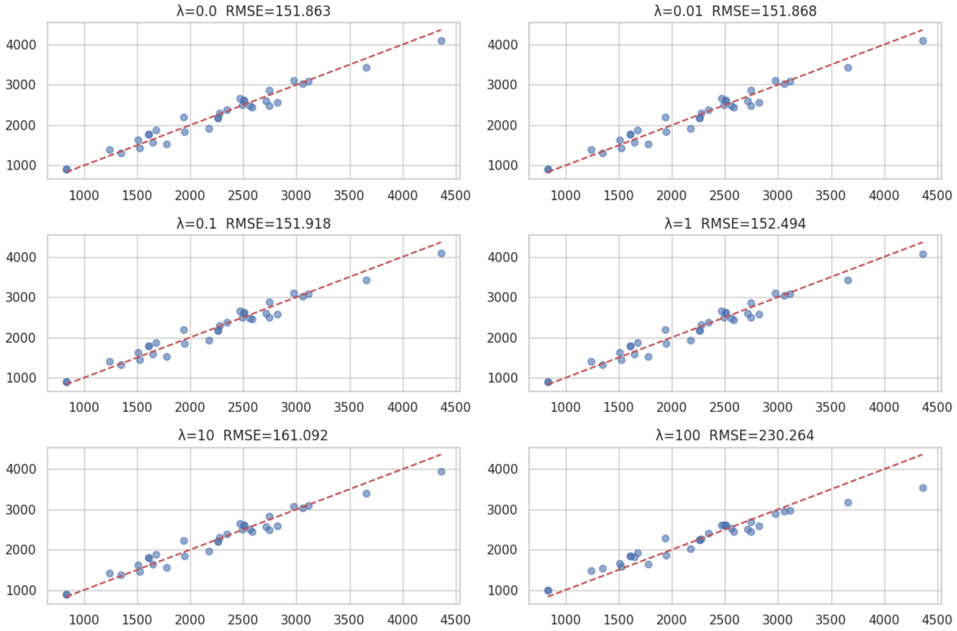

Figure 2: Predicted vs Actual values

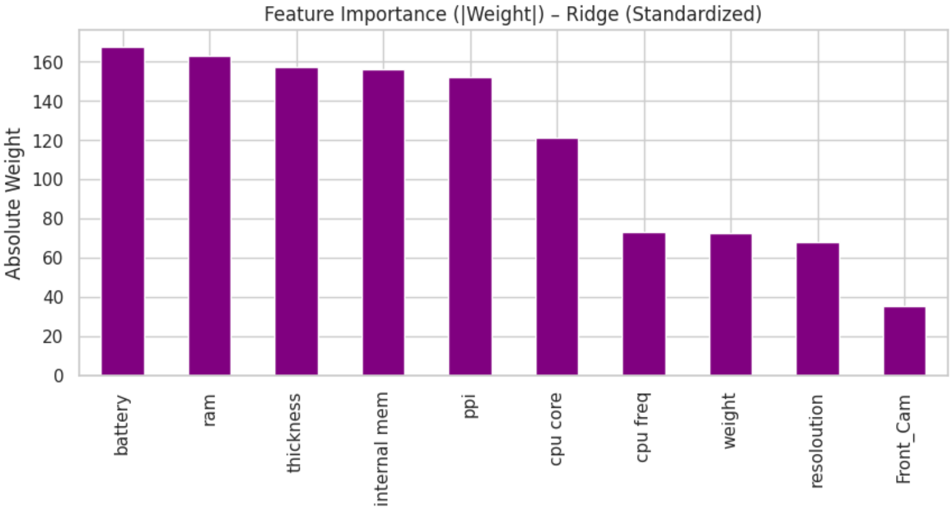

Figure 3: Feature Importance

4.6 Inference

- Standardized Ridge regression gave the best accuracy ($R^2 = 0.91$).
- Gradient descent converged slowly without normalization.
- Closed-form is computationally efficient for small datasets but not scalable.
- Regularization reduced overfitting and stabilized parameter weights.

4.7 Learning Outcome

- Understood how regression equations can be solved through matrix algebra.

- Observed how L2 regularization affects parameter magnitudes.
- Gained clarity on the influence of data scaling on optimization.

5. PART B – Bank Note Authentication (Classification)

5.1 Aim

To build a logistic regression classifier that predicts whether a bank note is genuine or forged, and to evaluate its performance under regularization and outlier conditions.

5.2 Mathematical Intuition

Logistic regression models the probability as:

$$P(y = 1|x) = \frac{1}{1 + e^{-X\theta}}$$

The cost function:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] + \frac{\lambda}{2}\|\theta\|^2$$

where λ controls the regularization strength.

5.3 Dataset Description

Dataset: `BankNote_Authentication.csv` (Kaggle, UCI repository)
Attributes:

- Variance
- Skewness
- Curtosis
- Entropy
- Class (0 – Fake, 1 – Real)

5.4 Methodology

1. The dataset was divided into 80% training and 20% testing.
2. Features were standardized using `StandardScaler`.
3. Logistic regression models were trained with:
   - No regularization
   - L2 regularization (varied)
4. Accuracy vs was plotted.
5. Outliers were introduced by shifting 10% of samples.
6. Coefficients before and after outliers were compared.

5.5 Results

Model	**Regularization**	**Accuracy**
Logistic Regression	None	0.975
Logistic Regression (L2, $\lambda = 1$)	Yes	0.978
After Outliers Added	L2	0.961

Accuracy vs λ Plot:

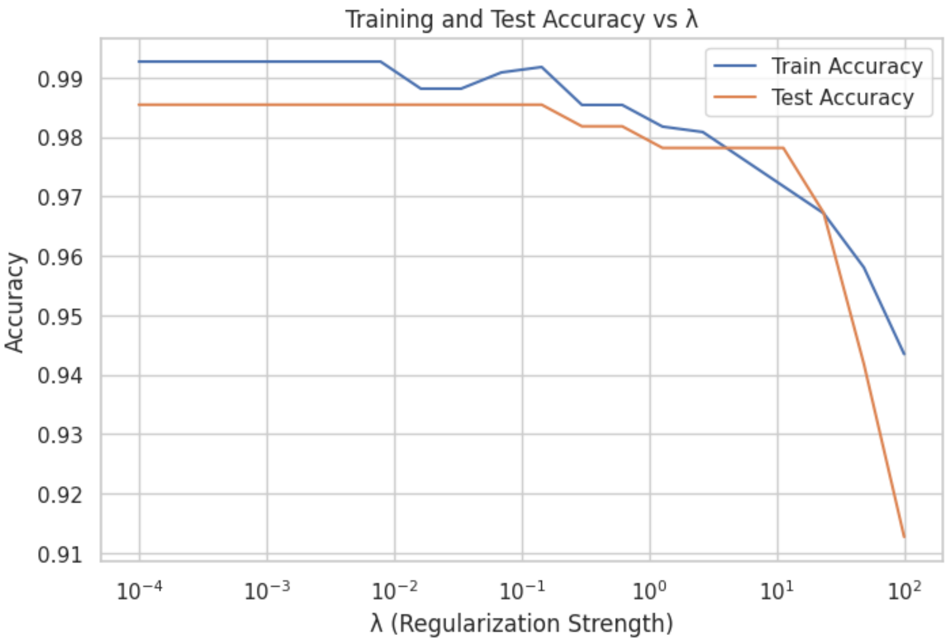

Figure 4: Training and Test Accuracy vs Regularization Strength (λ)

3D Visualization:

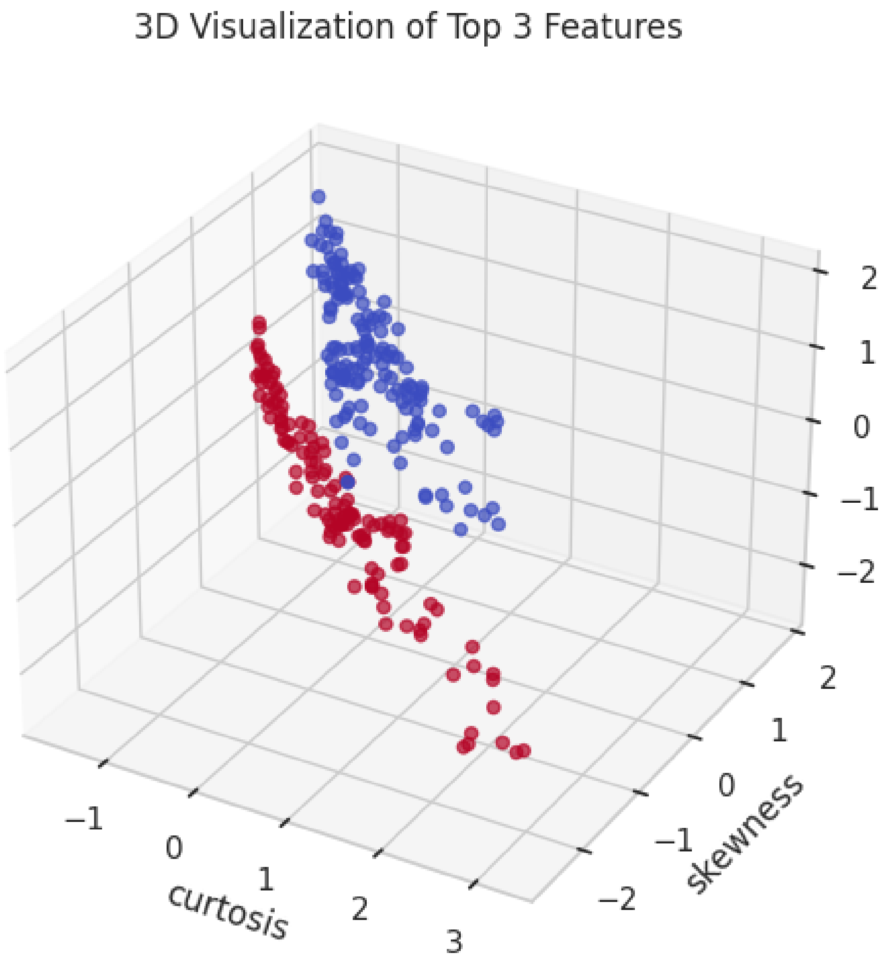

Figure 5: 3D Plot using Top 3 Important Features

Coefficient Change due to Outliers:

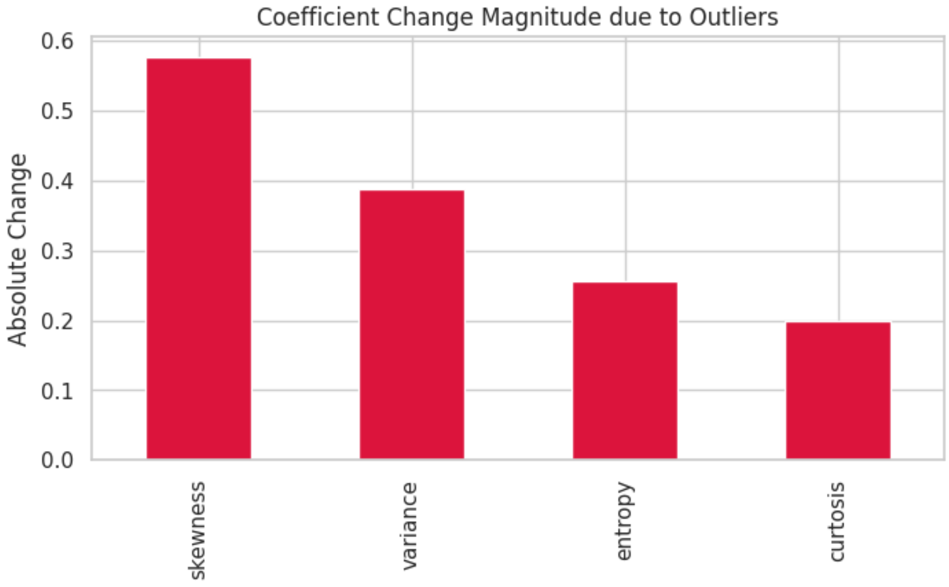

Figure 6: Coefficient Changes after Outlier Injection

5.6 Inference

- The dataset was linearly separable; hence logistic regression achieved nearly 98% accuracy.
- Regularization improved model robustness and reduced overfitting.
- Outliers caused minor accuracy drop, showing that L2 penalty dampens their influence.
- Variance and curtosis were the most important predictive features.

5.7 Learning Outcome

- Learned how logistic regression performs under regularization.
- Understood the effect of on accuracy and stability.
- Visualized model decision boundaries in 3D.
- Realized that even simple linear classifiers can achieve strong results when features are meaningful and well-scaled.

6. Overall Conclusion

This assignment helped in understanding how both regression and classification can be unified under linear algebraic frameworks. It reinforced the importance of normalization, regularization, and evaluation in machine learning. Linear Regression effectively captured pricing trends, while Logistic Regression showed strong separability in detecting fake banknotes. The entire process highlighted that mathematical clarity and proper data handling are key to reliable ML outcomes.