Project Description Document

This project consists of two main parts:

1. Numerical Healthcare

2. Flower Species Recognition

Part 1: Healthcare Dataset (Numerical Data)

a. General Information

- Dataset name: Healthcare Dataset
- Problem type: Supervised Learning – Regression
- Target variable: `Test Results`
- Total number of samples: `55500`
- Features used:
    - Categorical features: Gender, Blood Type, Medical Condition, Admission Type, Medication
    - Numerical features: Age, Days Admitted

Data Cleaning & Preparation

- No missing values were found; therefore, no imputation was required.
- No duplicated columns were detected.
- Outlier detection was performed on Age using the IQR method.
- New feature engineered:
    - Days Admitted = Discharge Date − Date of Admission
- Dropped noise/irrelevant columns:
    - Name, Doctor, Hospital, Insurance Provider, Billing Amount, Room Number, Date of Admission, Discharge Date

Data Split

- Training set: 80%
- Testing set: 20%

---

b. Implementation Details

Feature Extraction & Encoding

- **Categorical encoding:** One-Hot Encoding
- **Numerical scaling:** Min-Max Scaling
- **Final feature matrix:**
    - Sparse matrix (due to one-hot encoding)
    - Converted to dense format for Gradient Descent implementation

Label Encoding

- Target variable (`Test Results`) encoded using LabelEncoder

Models Implemented

1. Linear Regression (Gradient Descent)

- Learning rate: 0.001
- Number of epochs: 500
- Loss function: Mean Squared Error (MSE)

2. K-Nearest Neighbors Regressor (KNN)

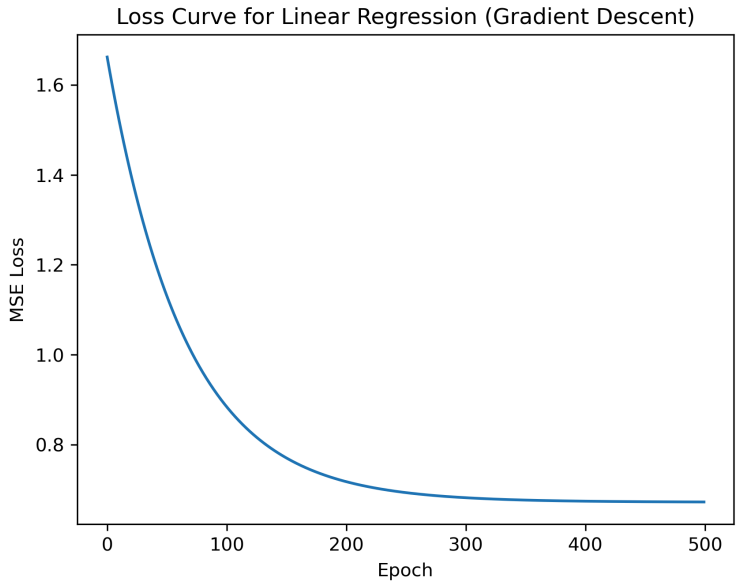
- Cross-validation: 5-Fold GridSearchCV
- Hyperparameters tuned:
  - Number of neighbors: [3, 5, 7, 9]
  - Distance metric: Euclidean, Manhattan
  - Weights: Uniform, Distance

---

c. Results Details (Testing Data)

Linear Regression

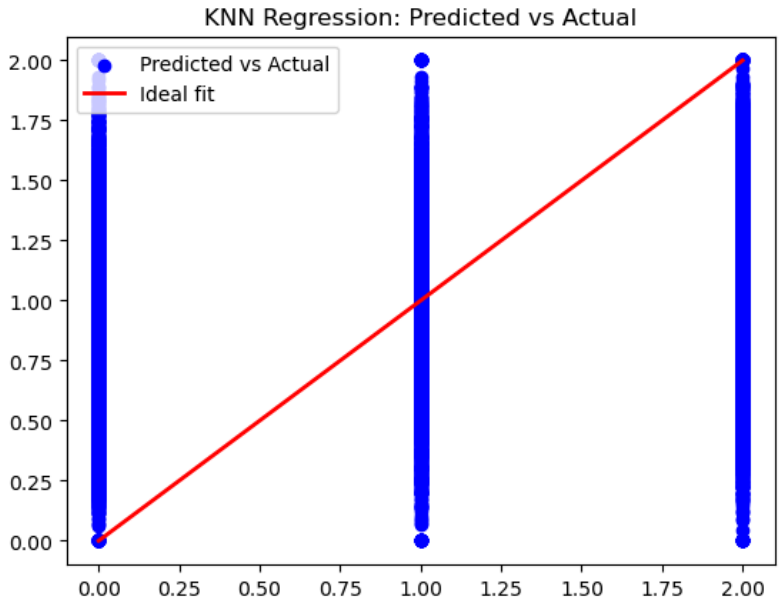
- Metric: Mean Squared Error (MSE) = 0.67
- Loss curve plotted for Gradient Descent model



KNN Regression

- Best parameters:
  - n_neighbors = 9
  - metric = Manhattan
  - weights = Distance
- Evaluation metrics:
  - Mean Squared Error (MSE) = 0.722
  - R^2 Score = -0.079

- Visualization:
    - Predicted vs Actual scatter plot



Part 2: Flower Species Recognition (Image Dataset)

a. General Information

- Dataset name: Oxford 102 Flower Dataset
- Problem types:
    - Unsupervised Learning (K-Means Clustering)
    - Supervised Learning (Logistic Regression Classification)

Original Dataset

- Total number of images: 8,189
- Image size (after resizing): 128 × 128 × 3
- Original labels: 102 flower classes

Selected Classes (Top 5 by frequency)

Class ID	Flower Name
51	Petunia
77	Passion Flower
46	Wallflower
73	Water Lily
89	Watercress

Filtered Dataset

- Number of classes: 5
- Labels: Re-mapped to [0, 1, 2, 3, 4]

Data Split (Supervised Models)

- Training set: 80%
- Testing set: 20%
- Stratified splitting used to preserve class balance

---

b. Implementation Details

Feature Extraction

- Method: Histogram of Oriented Gradients (HOG)
- Image preprocessing:
  - RGB → Grayscale
  - Resize to 128 × 128
- HOG parameters:
  - Pixels per cell: (8, 8)
  - Cells per block: (2, 2)
- Number of features per image: ~8,100 (depends on image size and HOG parameters)

Feature Scaling

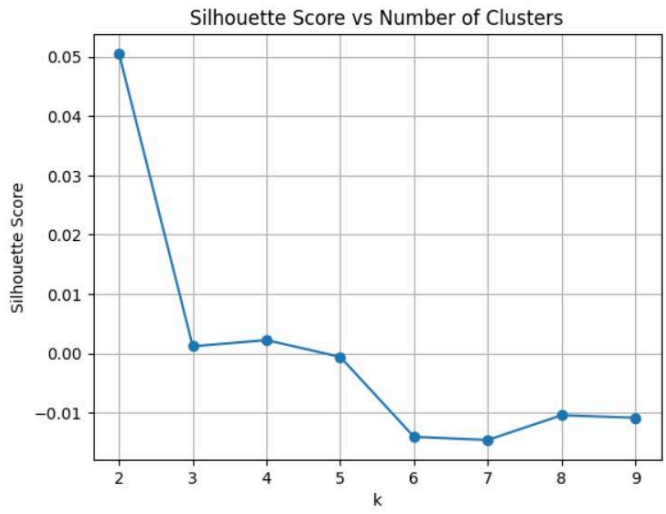
- StandardScaler (zero mean, unit variance)

Unsupervised Learning: K-Means Clustering

- Number of clusters: Tested from 2 to 9
- Final choice: 5 clusters
- Evaluation metric: Silhouette Score
- Cross-validation: Not applicable (unsupervised learning)

Results

- Silhouette score curve plotted



- Best performance observed around k = 5, matching the number of selected classes

---

Supervised Learning: Logistic Regression

Model Configuration

- Solver: LBFGS
- Max iterations per epoch: 1 (warm start enabled)
- Total epochs: 100
- Optimization: Maximum likelihood

Evaluation Metrics

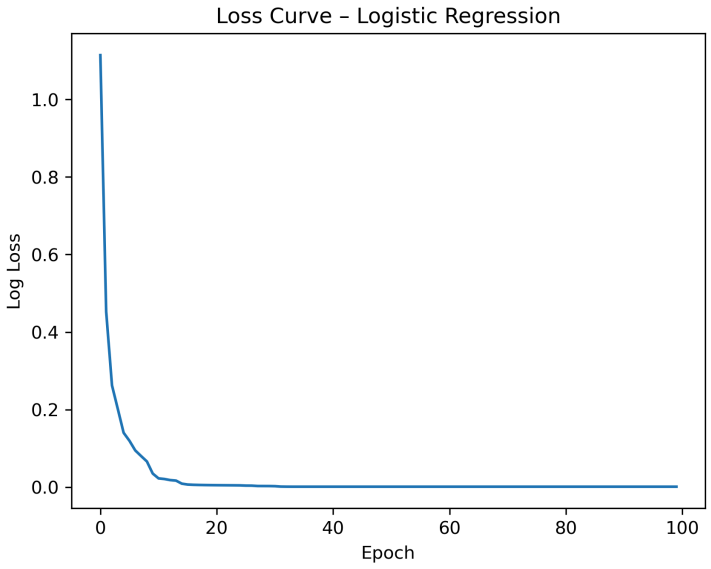
- Accuracy
- Confusion Matrix
- ROC Curve (One-vs-Rest)
- Log Loss

---

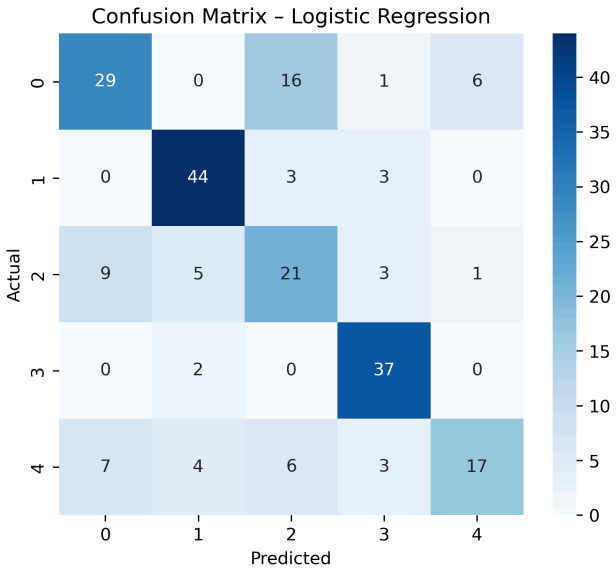
c. Results Details (Testing Data)

Logistic Regression Results

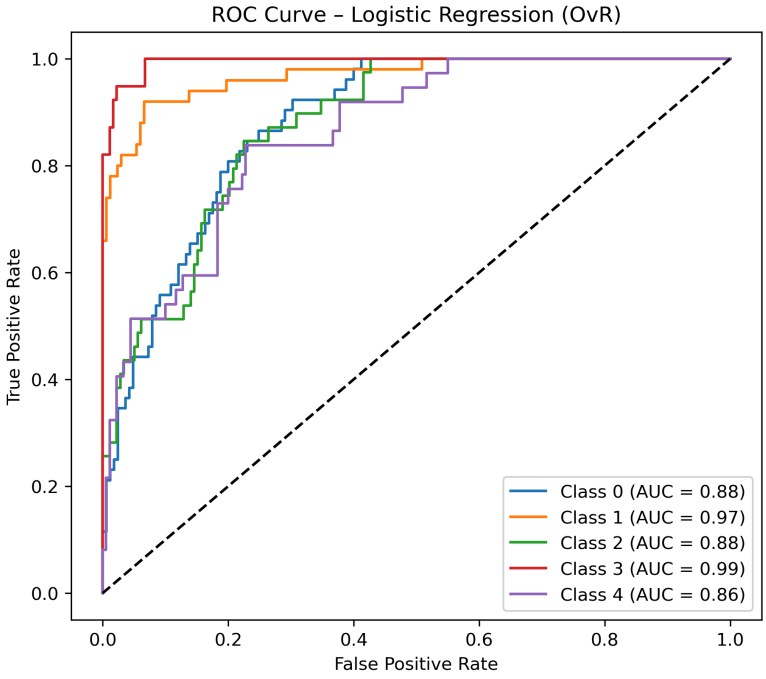
- Accuracy = 0.682
- Loss Curve:



- Confusion Matrix:

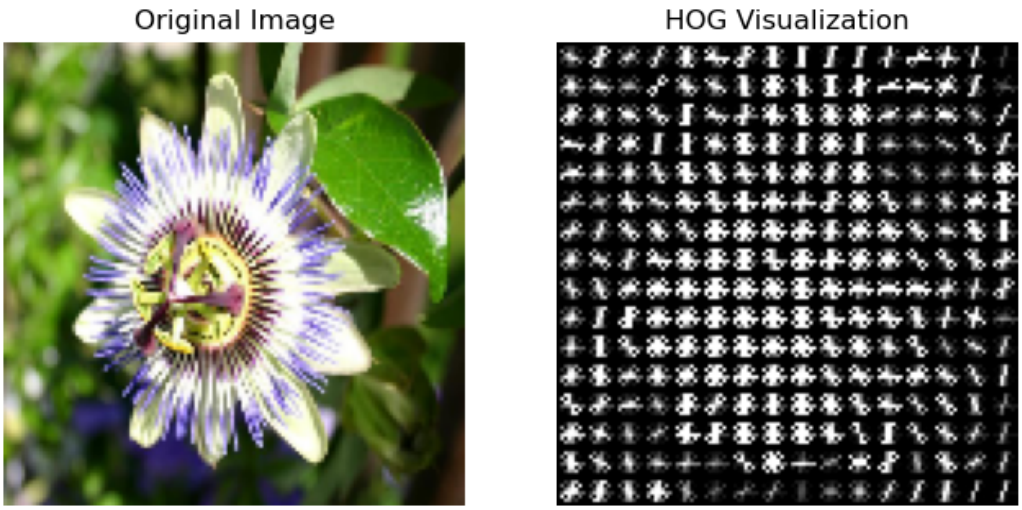


- ROC Curve:



Feature Visualization

- HOG feature visualization displayed alongside original flower image



Conclusion

- The healthcare dataset demonstrated the effectiveness of regression models and hyperparameter tuning using cross-validation.
- The flower dataset showed strong performance using handcrafted HOG features combined with classical machine learning models.
- Both supervised and unsupervised approaches were successfully implemented and evaluated using appropriate metrics.