Machine Learning Project

Team Members

1.	Mahmoud Mohamed Ahmed Sayed (Team Leader)	20210876
2.	Adel Ali Ibrahim Ali	20210481
3.	Merna Ayman Louis Soares	20220455
4.	Yara Ahmed Fathy Habib	20220550
5.	Marwan Mohamed Salah El-Din Mohamed	20210899
6.	Rami Bbawy Ayad Gabriel	20210317

Project Overview

Instructor: Dr. Salwa Emam.

Project description: The project focuses on implementing and evaluating machine learning models for classification and regression tasks on two datasets:

- Oxford 102 Flowers Dataset (Image Classification)
- Credit Risk Dataset (Numerical Regression)

KNN & Logistic and Linear Regression

KNN:

- Nature: Non-parametric model (no assumption on data distribution).
- Training: No explicit training phase; stores the dataset and computes distances during prediction.
- Interpretability: Less interpretable due to reliance on nearest neighbors without a clear decision boundary.
- Performance: Can capture more complex decision boundaries and is more flexible for high-dimensional image data.
- Scalability: Computationally expensive during inference, especially with large datasets, since it requires comparing every test point to all training points.

Logistic Regression:

- Nature: Parametric model (requires assumption about the underlying distribution of data).
- Training: Faster since it involves finding optimal weights for features.
- Interpretability: Good, as the model provides the influence of each feature on the prediction.
- Performance: Works well for simpler patterns, but struggles with highly complex or non-linear image data.

- Scalability: More scalable to large datasets due to lower computational cost during inference.

Linear Regression:

- Nature: Parametric model (assumes a linear relationship between features and target).
- Training: Faster to train, as it directly calculates the optimal weights using closed-form solutions (or gradient descent).
- Interpretability: High, as it shows the linear relationship between the input features and target variable.
- Performance: Works well when the relationship between features and target is truly linear. Struggles with non-linear relationships.
- Scalability: Scalable to large datasets, as it has low computational cost after training.

Numerical Dataset Overview

1. Dataset name: Credit risk dataset.
2. Total Rows: 32,581.
3. Total Columns: 12.
4. Columns:
    - person_age: Age of the person.
    - person_income: Annual income of the person.
    - person_home_ownership: Type of home ownership (e.g., RENT, OWN, MORTGAGE).
    - person_emp_length: Length of employment (in years), with some missing values (895).
    - loan_intent: Purpose of the loan (e.g., PERSONAL, EDUCATION, MEDICAL).
    - loan_grade: Loan grade (categorical, e.g., A, B, C, D).
    - loan_amnt: Loan amount.
    - loan_int_rate: Interest rate of the loan, with missing values (3,116).
    - loan_status: Loan status (0 or 1).
    - loan_percent_income: Loan amount as a percentage of income.
    - cb_person_default_on_file: Whether the person has a history of default (Y/N).
    - cb_person_cred_hist_length: Credit history length.

5. Missing Values:
    - person_emp_length: 895 missing values.

- loan_int_rate: 3,116 missing values.

6. Sample Of Data (First five rows):

Person_age	Person_income	Person_home_o	Person_emp_le	Loan_intent	Loan_grade	Loan_amnt	Loan_int_rate	Loan_status	Loan_percent_i	Cb_person_defa	Cb_person_cred
22	59000	RENT	123	PERSONAL	D	35000	16.02	1	0.59	Y	3
21	9600	OWN	5	EDUCATION	B	1000	11.14	0	0.1	N	2
25	9600	MORTGAGE	1	MEDICAL	C	5500	12.87	1	0.57	N	3
23	65500	RENT	4	MEDICAL	C	35000	15.23	1	0.53	N	2
24	54400	RENT	8	MEDICAL	C	35000	14.27	1	0.55	Y	4

KNN & Linear Regression On The Numerical Dataset

Metric	Linear Regression	KNN
Mean Squared Error (MSE)	1.8691	1.9021
Mean Absolute Error (MAE)	1.0395	1.0399
R^2 Score	0.8051	0.8017

1. Mean Squared Error (MSE):

- Linear Regression (1.8691) has a slightly lower MSE than KNN (1.9021), meaning it makes smaller squared prediction errors on average.
- Lower MSE is better, so Linear Regression performs better here.

2. Mean Absolute Error (MAE):
    - Both algorithms have almost identical MAE: Linear Regression (1.0395) vs. KNN (1.0399).
    - MAE measures average absolute errors, and in this case, both algorithms are performing nearly equally well.

3. R^2 Score (Coefficient of Determination):
    - Linear Regression: 0.8051
    - KNN: 0.8017
    - R^2 indicates how well the model explains the variability in the target variable. Higher is better.
    - Linear Regression has a slightly better R^2 score, meaning it explains a marginally higher proportion of the variance in the data compared to KNN.

- **Conclusion:**

  Linear Regression performs slightly better overall in terms of MSE and R^2 score, while MAE is nearly identical. Linear Regression is preferred for simplicity and interpretability.

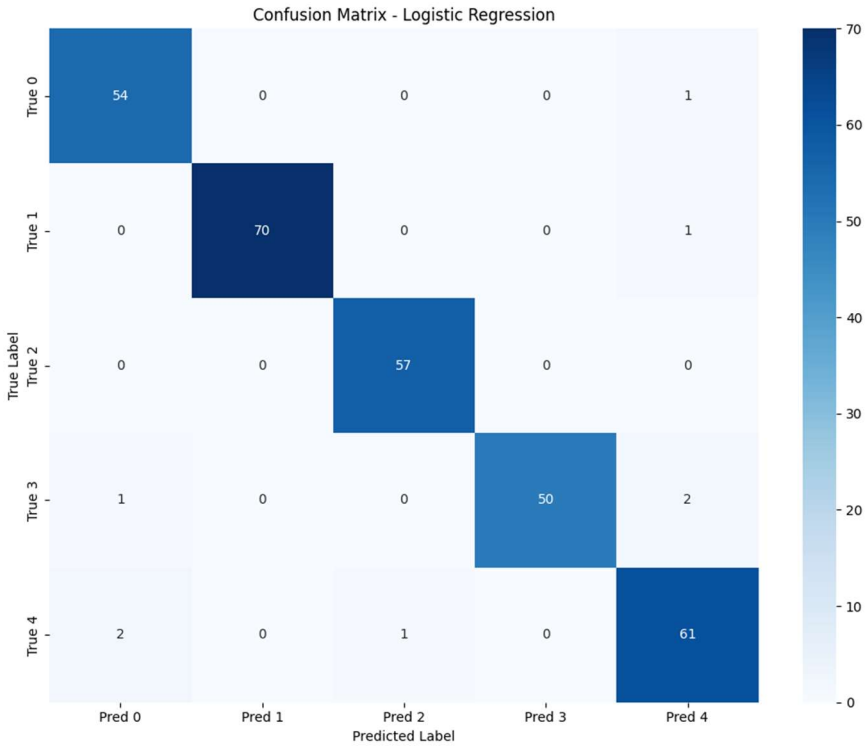
Image Dataset Overview

1. Dataset name: Oxford 102 Flowers Dataset.
2. Total Rows: 8,189 images.
3. Total Classes: 102 flower categories.
4. Image Sizes: Images are of variable sizes, but all are in JPEG format.
5. Class Distribution: Each class (flower type) has between 40 and 258 images.
6. The dataset comes with class labels (1 to 102), corresponding to the flower types.
7. Sample Of Data (First two images):

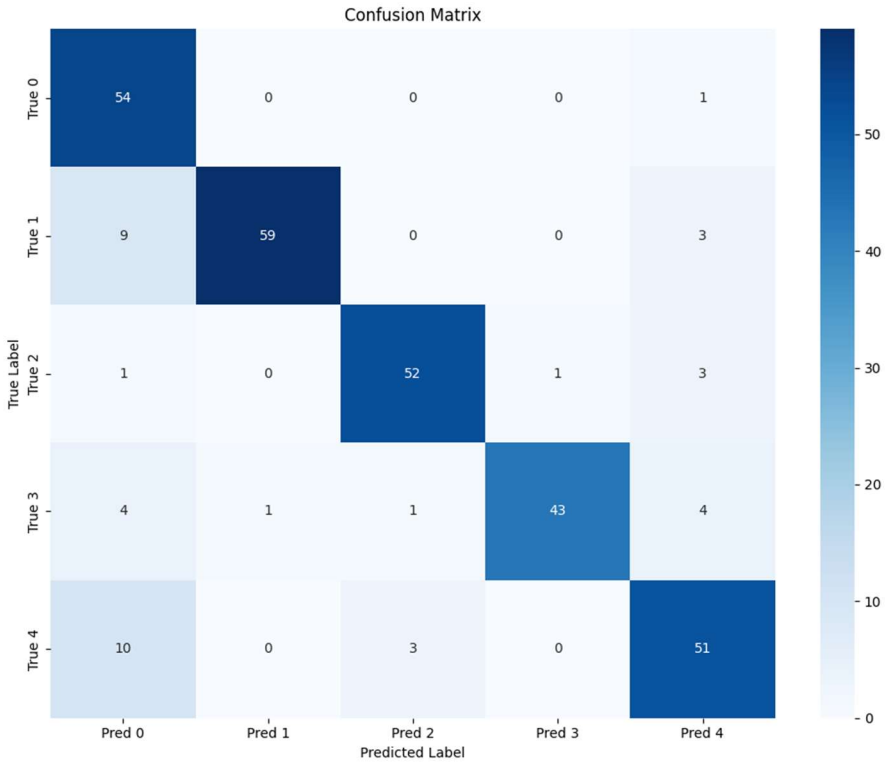
KNN &Logistic on The Image Dataset

Metric	Logistic	KNN (K = 4)
Accuracy Score	0.9733333333333334	0.8633333333333333
Loss Value	0.1570	1.0681
F1-Score	0.97	0.88
Recall	0.97	0.87
Precision	0.97	0.88

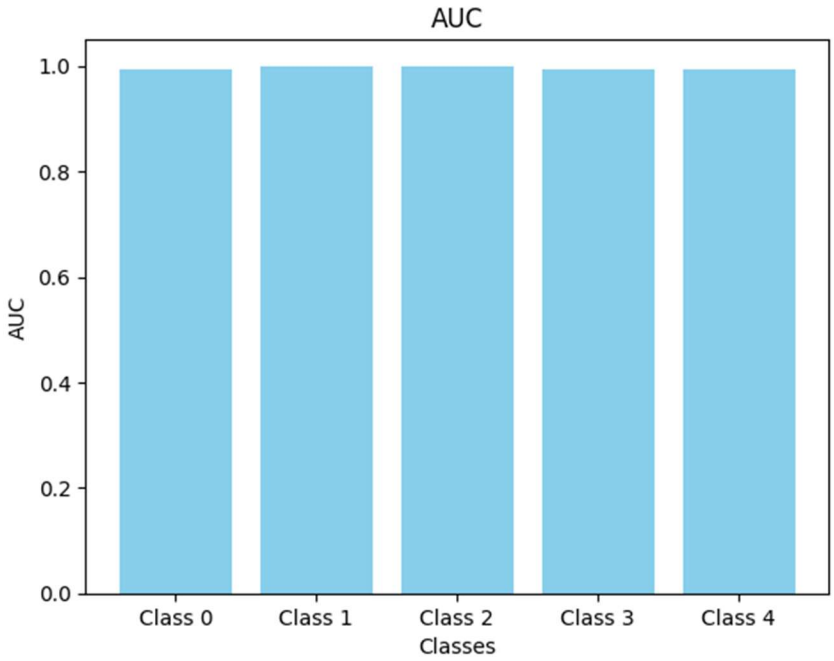
1. Confusion Matrix:
   - Logistic:



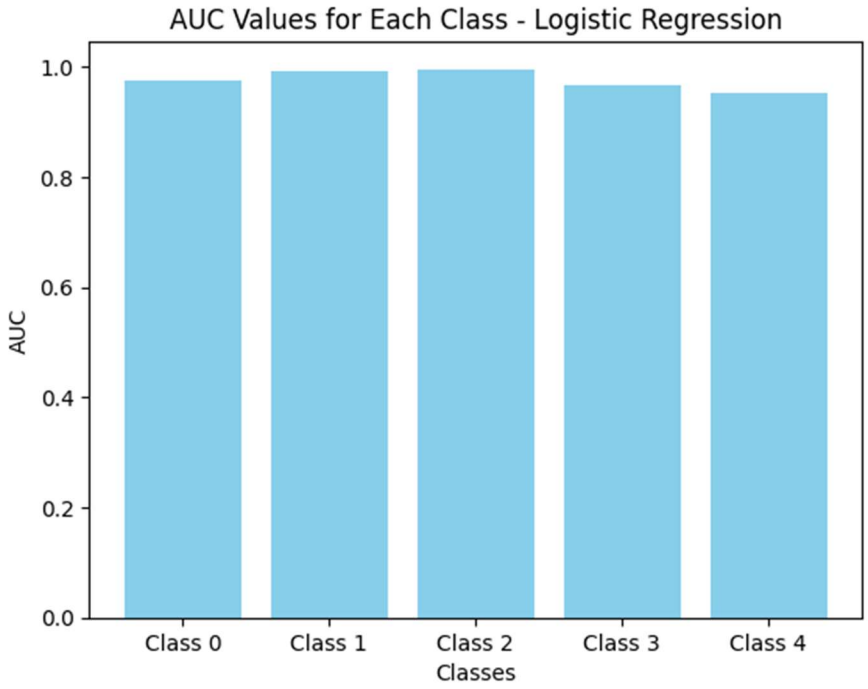
- KNN:



2. AUC:
   - Logistic

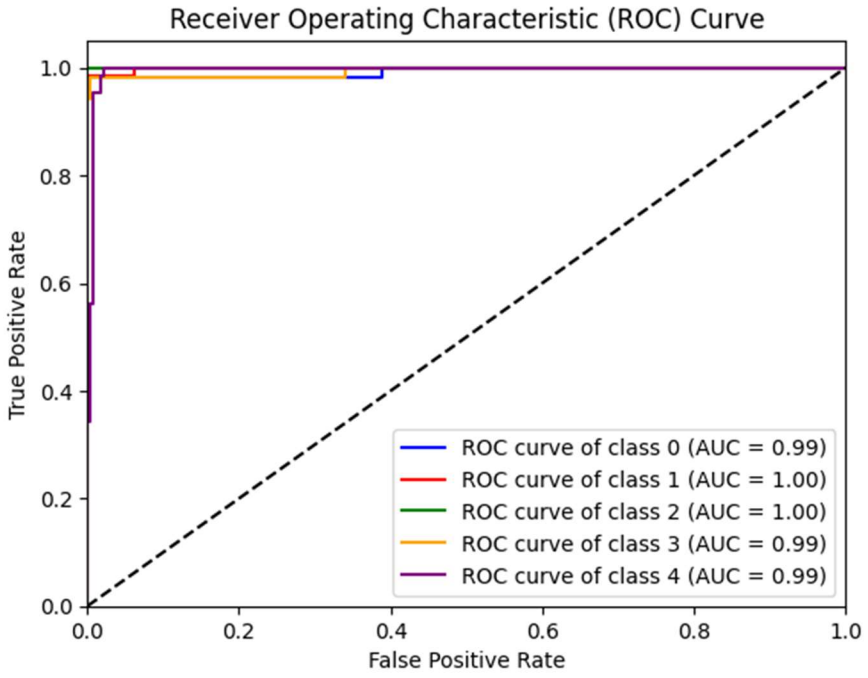


- KNN

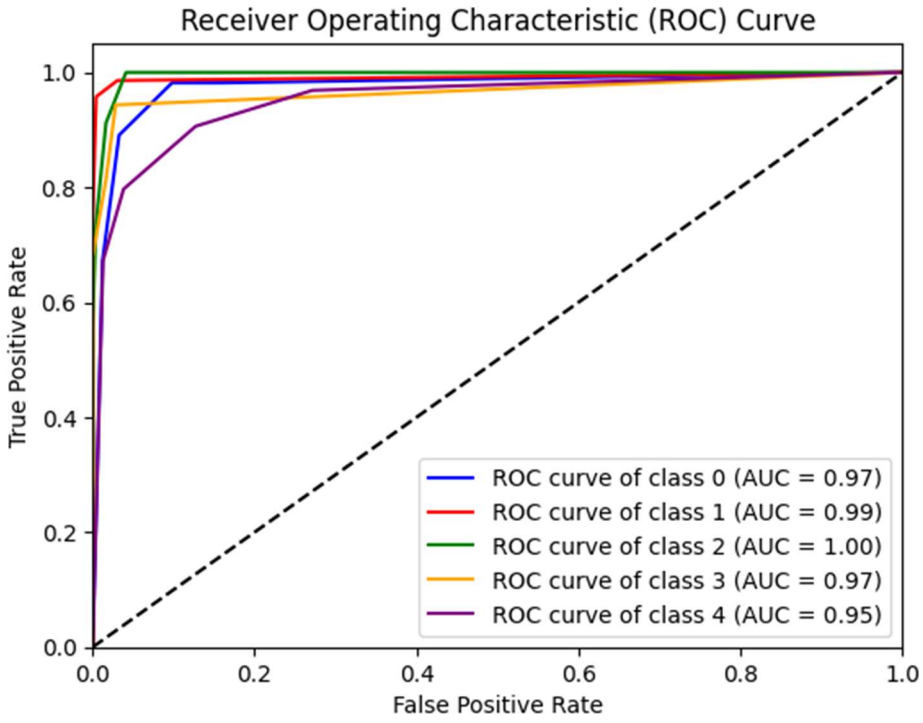


3. ROC:

- Logistic



- KNN



- Accuracy: Logistic Regression achieves 97.3% accuracy, significantly outperforming KNN, which achieves 86.3% accuracy.

- Precision, Recall, F1-Score:
  - Logistic Regression shows higher scores across all metrics for each class, indicating better overall performance.
  - KNN struggles slightly in Class 3 and Class 4 compared to Logistic Regression.

- Precision, Recall, F1-Score:
  - Logistic Regression shows higher scores across all metrics for each class, indicating better overall performance.
  - KNN struggles slightly in Class 3 and Class 4 compared to Logistic Regression.

- Conclusion:

  Logistic Regression clearly outperforms KNN in terms of accuracy, precision, recall, and F1-score. It generalizes better on the Oxford 102 Flowers dataset, making it the preferred model for this classification task.

Tools & Techniques Used

- os: Interact with the operating system (file paths, directories).
- numpy: Numerical operations on arrays and matrices.
- matplotlib.pyplot: Data visualization library for plots and graphs.
- seaborn: Simplifies statistical data visualization.
- classification_report: Generates precision, recall, and F1-score for classification.
- confusion_matrix: Evaluates model performance with a confusion matrix.
- roc_curve / auc: Plots ROC curve and calculates the area under it.
- precision_recall_curve / average_precision_score: Precision-recall tradeoff analysis.
- to_categorical: Converts labels to one-hot encoded format.
- KNeighborsClassifier: KNN algorithm for classification.
- train_test_split: Splits data into training and testing sets.
- learning_curve: Analyzes model learning performance.
- PIL.Image: Loads and processes images.
- loadmat: Loads MATLAB .mat files.
- resample: Balances datasets through sampling.
- StandardScaler: Standardizes features for consistent scaling.
- label_binarize: Converts labels into binary format.

Thank you