Project Report

Dataset Descriptions

Numerical Dataset: Sales and Satisfaction Data

Source: Sales and Satisfaction Dataset on Kaggle (https://www.kaggle.com/datasets/matinmahmoudi/sales-and-satisfaction)

This synthetic dataset contains information on sales and customer satisfaction before and after an intervention. It also includes purchase data for control and treatment groups. The dataset includes missing values (NaNs) in some of its entries.

Features:

1. **Group**
    - **Description**: Indicates whether the data point belongs to the Control or Treatment group.
    - **Categories**: Control, Treatment
2. **Customer_Segment**
    - **Description**: Categorizes customers based on their value.
    - **Categories**: High Value, Medium Value, Low Value
3. **Sales_Before**
    - **Description**: Sales figures before the intervention.
    - **Data Type**: Numerical
4. **Sales_After**
    - **Description**: Sales figures after the intervention.
    - **Data Type**: Numerical
5. **Customer_Satisfaction_Before**
    - **Description**: Customer satisfaction scores before the intervention.
    - **Data Type**: Numerical
6. **Customer_Satisfaction_After**
    - **Description**: Customer satisfaction scores after the intervention.
    - **Data Type**: Numerical
7. **Purchase_Made**

- **Description**: Indicates whether a purchase was made after the intervention.
- **Categories**: Yes, No

Dataset Details:

- **Number of Records**: 10,000 entries
- **Missing Values**: Yes, certain features contain NaN values.
- **Number of Classes**: N/A (Regression task)

Image Dataset: Traffic Sign Classification

Source: Traffic Sign Dataset on Kaggle (https://www.kaggle.com/datasets/ahemateja19bec1025/traffic-sign-dataset-classification)

This dataset consists of images of traffic signs, specifically speed limit signs. For this project, we focused on 5 classes representing different speed limits.

Classes:

1. **Speed Limit 40 km/h**
2. **Speed Limit 50 km/h**
3. **Speed Limit 60 km/h**
4. **Speed Limit 70 km/h**
5. **Speed Limit 80 km/h**

Dataset Details:

- **Number of Images**: Varies per class; total images across all classes used.
- **Image Size**: Standardized dimensions used during preprocessing.
- **Missing Values**: No missing images or labels.
- **Number of Classes**: 5 (Classification task)

Algorithms Applied

Numerical Dataset (Regression Task)

1. **Linear Regression**:
    - A fundamental regression technique that models the relationship between a dependent variable and one or more independent variables by fitting a linear equation.
    - **Usage**: To predict sales figures after the intervention based on the available features.
2. **K-Nearest Neighbors (KNN) Regression**:

- A non-parametric method that predicts the value of a variable based on the average of the nearest K neighbors.
- **Usage**: To provide an alternative prediction model for sales figures, allowing comparison with Linear Regression.

Image Dataset (Classification Task)

1. **Logistic Regression**:

   - A statistical model used for binary classification, extended here for multi-class classification using a one-vs-rest approach.
   - **Usage**: To classify images of traffic signs into one of the five speed limit categories.

2. **K-Nearest Neighbors (KNN) Classification**:

   - A simple, instance-based learning algorithm that classifies a data point based on majority voting among its k nearest neighbors.
   - **Usage**: To classify traffic sign images, offering a comparison with the Logistic Regression model.

Results Comparison

Numerical Dataset Regression Results

Evaluation Metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors in a set of predictions, without considering their direction.
- **Mean Squared Error (MSE)**: Measures the average squared difference between the estimated values and actual value.
- **R-Squared Score (R²)**: Represents the proportion of variance for a dependent variable that's explained by an independent variable.

Training Data Evaluation

Metric	Linear Regression	KNN Regression
Mean Absolute Error	9.13	8.04
Mean Squared Error	134.46	102.91
R-Squared Score	0.51	0.62

Test Data Evaluation

Metric	Linear Regression	KNN Regression
Mean Absolute Error	9.47	10.16
Mean Squared Error	145.40	164.94
R-Squared Score	0.48	0.41

Interpretation:

- **Training Data**:
  - KNN Regression outperforms Linear Regression, showing lower MAE and MSE, and a higher R² score.
- **Test Data**:
  - Linear Regression performs slightly better than KNN Regression on MAE and MSE.
  - R² scores indicate that both models explain a similar proportion of variance, with Linear Regression performing marginally better.

Image Dataset Classification Results

Evaluation Metrics:

- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives.
- **Recall**: The ratio of correctly predicted positive observations to all actual positives.
- **F1-Score**: The weighted average of Precision and Recall.
- **Support**: The number of actual occurrences of each class in the specified dataset.

Training Data Evaluation

Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
0	1.00	1.00	1.00	255
1	1.00	1.00	1.00	122
2	1.00	1.00	1.00	190
3	1.00	1.00	1.00	89
4	1.00	1.00	1.00	166
Accuracy			**1.00**	822

KNN Classification Model

Same performance as Logistic Regression on training data.

Test Data Evaluation

Logistic Regression Model

Class	Precision	Recall	F1-Score	Support
0	1.00	0.94	0.97	63
1	0.94	1.00	0.97	34
2	0.96	0.93	0.94	54
3	1.00	1.00	1.00	19
4	0.90	1.00	0.95	36

Class	Precision	Recall	F1-Score	Support
Accuracy			**0.96**	206

KNN Classification Model

Class	Precision	Recall	F1-Score	Support
0	0.91	0.94	0.92	63
1	1.00	0.76	0.87	34
2	0.96	0.93	0.94	54
3	1.00	1.00	1.00	19
4	0.82	1.00	0.90	36
Accuracy			**0.92**	206

Interpretation:

- **Training Data**:
  - Both models achieved perfect scores, indicating potential overfitting.
- **Test Data**:
  - Logistic Regression outperforms KNN, achieving higher accuracy and F1-scores across most classes.
  - KNN shows lower recall for class 1, indicating some misclassifications.

ROC Curve Analysis

Note: The ROC (Receiver Operating Characteristic) curves provide a graphical representation of the model's diagnostic ability.

- **Logistic Regression**:
  - Expected to have a higher AUC (Area Under Curve), indicating better performance in distinguishing between classes.
- **KNN Classification**:
  - May show lower AUC compared to Logistic Regression, especially if overfitting occurred during training.

Since ROC curves are visual, please refer to the attached graphs for a detailed comparison.

Conclusion

- **Numerical Dataset**:
  - Both Linear Regression and KNN Regression models have comparable performance.
  - Linear Regression shows better generalization on test data.
- **Image Dataset**:

    - Logistic Regression provides better classification performance on test data than KNN.
    - Overfitting is a concern, as both models perform perfectly on training data.

Recommendations: (*based on our implementation*)

- For the numerical dataset, consider using Linear Regression for better generalization.
- For the image dataset, Logistic Regression is preferred over KNN for its higher accuracy and robustness.
- Implement regularization techniques to address overfitting in future models.