Evaluation of Linear and Logistic Regression Techniques Using Air Quality Data

Name: Avinash Angilikam
Student ID: 23037971
GitHub Link: https://github.com/AVINASH-ANGILIKAM/Evaluation-of-Linear-and-Logistic-Regression-Techniques-Using-Air-Quality-Data

1. Introduction

This report explores two regression techniques, Linear Regression and Logistic Regression, applied to the Air Quality dataset. Linear Regression is used to predict temperature values based on pollutant levels, while Logistic Regression classifies CO levels into High and Low categories. The report discusses preprocessing, modeling, and analysis, accompanied by relevant visualizations.

2. Data Preprocessing

Preparing the dataset for analysis involved several essential steps:

1. **Handling Missing Data**: Missing values in the dataset were replaced using column averages to ensure completeness.
2. **Standardization**: Numeric features were scaled using StandardScaler to normalize the data.
3. **Feature Selection**:
   - **Linear Regression**: Predictor variables included pollutant levels (CO(GT), PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC)), and the target variable was temperature (T).
   - **Logistic Regression**: Predictors included pollutant levels (PT08.S1(CO), NMHC(GT), C6H6(GT), PT08.S2(NMHC), PT08.S3(NOx)), while the binary target variable classified CO levels as High (>2.0) or Low.

3. Methodology

Linear Regression

Objective: To predict temperature (T) based on pollutant levels.

- **Steps**:
  - The dataset was divided into training (80%) and testing (20%) subsets.
  - A Linear Regression model was trained using pollutant levels as predictors and temperature as the target.
- **Performance Metrics**:
  - Mean Squared Error (MSE): 0.762
  - R^2 Score: 0.254
- **Visualization**: The scatter plot below shows predicted vs. actual temperature values. The weak correlation suggests the model explains only 25.4% of the variance in temperature, indicating room for improvement.

Explanation:
The scatter plot demonstrates the performance of Linear Regression. While some predictions align with the actual values, the spread of data points highlights a limited fit, indicating that the model struggles with capturing non-linear relationships in the data.

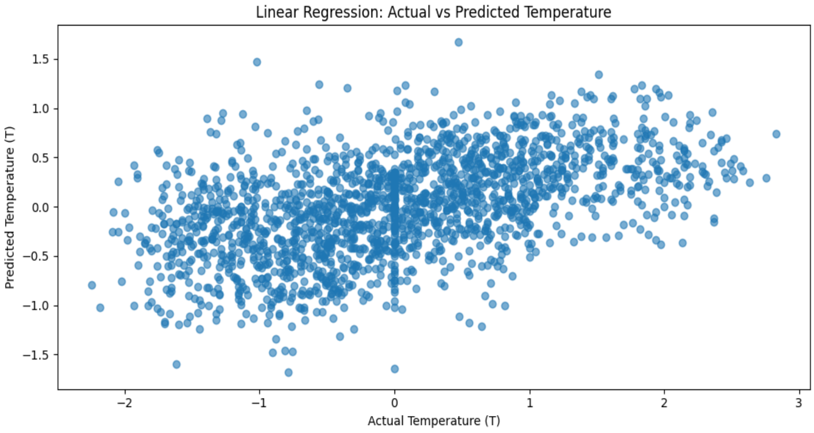

Figure 1: Linear Regression Scatter Plot

Logistic Regression
Objective: To classify CO levels as High or Low based on pollutant levels.

- **Steps**:
  - A binary target variable was created using a threshold for CO levels (>2.0).
  - The dataset was split into training (80%) and testing (20%) subsets.
  - A Logistic Regression model was trained using pollutant levels as inputs.
- **Performance Metrics**:
  - Accuracy: 79.49%
  - Precision: 84.56%
  - Recall: 75.03%
  - F1 Score: 79.51%

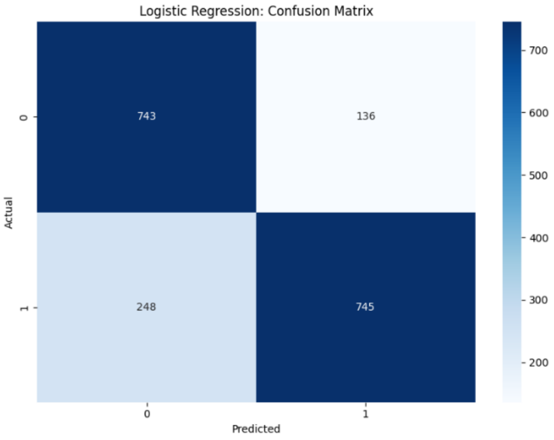

Figure 2: Logistic Regression Confusion Matrix

- **Visualization**: The confusion matrix below illustrates the model's classification performance, showing the distribution of true positives, true negatives, false positives, and false negatives.

Explanation:
The confusion matrix highlights that the model effectively classifies CO levels. A total of 743 true negatives (correctly identified Low CO levels) and 745 true positives (correctly identified High CO levels) reflect good precision and overall accuracy. However, the 248 false negatives indicate the need for improved recall.

4. Results and Discussion

Linear Regression Results:

- **Strengths**: Useful for predicting continuous variables with linear relationships.
- **Limitations**: The low R² score suggests the model fails to capture complex patterns in the data. Additional predictors or feature transformations are necessary to improve performance.

Logistic Regression Results:

- **Strengths**: Effective for binary classification tasks, with high precision and reasonable accuracy.
- **Limitations**: The relatively lower recall indicates a tendency to misclassify High CO levels as Low. Further feature engineering or hyperparameter tuning can address this issue.

5. Recommendations

- **Linear Regression**: Improve the model by adding non-linear transformations of features or exploring additional predictors.
- **Logistic Regression**: Focus on improving recall by using techniques such as regularization or incorporating domain-specific features.

6. Conclusion

This analysis showcases the application of Linear and Logistic Regression. Linear Regression is useful for continuous predictions but struggles with complex data. Logistic Regression, on the other hand, demonstrates robust classification performance with potential for further optimization. Both methods are valuable tools when applied with appropriate preprocessing and modeling strategies.

References

1. Tan, P-N., Steinbach, M., Karpatne, A. & Kumar, V., 2018. *Introduction to Data Mining*. 2nd ed. Pearson.
2. Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). SAGE Publications.
3. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.
4. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
5. Python Software Foundation, 2025. *Python 3.11 Documentation*. Available at: https://docs.python.org/3/ [Accessed 5 Jan. 2025].
6. Pedregosa, F. et al., 2011. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*, 12, pp. 2825–2830.
7. Hunter, J.D., 2007. *Matplotlib: A 2D Graphics Environment*. *Computing in Science & Engineering*, 9(3), pp. 90–95.
8. Waskom, M., 2021. *Seaborn: Statistical Data Visualization*. Available at: https://seaborn.pydata.org [Accessed 5 Jan. 2025].
9. Air Quality UCI Dataset, 2004. *Air Quality Data Set*. Available at: https://archive.ics.uci.edu/ml/datasets/Air+Quality [Accessed 5 Jan. 2025].