Danny Nguyen - CISD 43 Spring 2025 Final Project

Introduction

This project will apply a supervised machine learning analysis on a dataset about Smart Home Device Efficiency. This dataset was found on Kaggle and is described as follows: Predict Smart Home Device Efficiency Dataset

Description:

This dataset captures smart home device usage metrics, offering insights into user behavior, device efficiency, and preferences. It includes data on device types, usage patterns, energy consumption, malfunction incidents, and user satisfaction metrics.

Features:

- **UserID:** Unique identifier for each user.
- **DeviceType:** Type of smart home device (e.g., Lights, Thermostat).
- **UsageHoursPerDay:** Average hours per day the device is used.
- **EnergyConsumption:** Daily energy consumption of the device (kWh).
- **UserPreferences:** User preference for device usage (0 - Low, 1 - High).
- **MalfunctionIncidents:** Number of malfunction incidents reported.
- **DeviceAgeMonths:** Age of the device in months.
- **SmartHomeEfficiency (Target Variable):** Efficiency status of the smart home device (0 - Inefficient, 1 - Efficient).

Objective:

The objective of this project will be to develop an algorithm that predicts whether a smart home device is considered Efficient or Inefficient. The data will be split 80/20 into a training set and a testing set.

The analysis will utilize two supervised machine learning types suitable for this type of binary outcome: K-Nearest Neighbor and Logistic Regression.

Conclusion:

This project used both the K-NN and Logistic Regression supervised machine learning models to predict the SmartHomeEfficiency Target Variable, or whether a device is efficient or inefficient.

Both K-NN and Logistic Regression produced similar results, though K-NN performed slightly better. Both models produced accuracy rates of approximately 90% in Python, with similarly high Precision, Recall, and F1-scores. The results produced in Rapidminer were also similar.

I am not surprised at the results because this project used a simple synthetic dataset that was produced for educational purposes on Kaggle. If a more complex, real-world dataset were used, there would possibly be a need to explore other more advanced methodologies.