IBM PROJECT

PREDICTIVE MODELING OF ENERGY CONSUMPTION IN IOT COMMUNICATION NETWORKS

Prepared by :

Name : Stuti Kumari

AJU/241581

Name : Arjit Prakher

AJU/241367

Submission Date :

13-10-25

Index

Sl.no	**Topic**	**Page no.**
1	Summary	1
2	Introduction	2
3	Project	3
4	Conclusion	23

1. Summary

This project focuses on applying predictive analytics techniques to a dataset concerning **Energy-Efficient Communication Protocols for Large-Scale IoT Deployments**. The goal is to build machine learning models using **IBM SPSS Modeler** to forecast critical network performance metrics, specifically **energy consumption (Regression)** and **transmission reliability (Classification)**.

The project fulfills the requirements for the **Predictive Analysis** course, requiring the development and comparison of at least **three distinct predictive models** (for a team of 2).

2. Introduction

The Challenge of Energy Efficiency in IoT Networks

The explosive growth of **Internet of Things (IoT)** networks poses a critical challenge: achieving **energy efficiency** and long-term sustainability. Many IoT devices rely on finite battery power, making excessive energy consumption a direct threat to their operational lifespan, scalability, and maintenance costs. The network's performance is determined by a complex interplay between device type, communication protocol, and environmental factors.

Project Objective and Methodology

This project utilizes **Machine Learning (ML)** to move beyond simple monitoring and develop a predictive framework for optimizing network performance. Our objective is to understand the drivers of communication success and energy use by employing three key analytical models:

1. **C5.0 Classification:** To establish clear, actionable rules for maximizing **transmission_success**.
2. **TwoStep Clustering:** To **discover natural operational segments** within the network, categorizing devices by their inherent energy and reliability profiles.
3. **Linear Regression:** To build a highly accurate, predictive equation for **energy_consumed**, quantifying the influence of protocols and operational modes.

By integrating these supervised and unsupervised learning techniques, this analysis provides the quantitative insights necessary to guide immediate network optimization decisions and ensure the sustainable, efficient operation of the entire IoT ecosystem.

3. Project details

3.1 Tools and Dataset

Component	Detail
Primary Tool	**IBM SPSS Modeler v18+** (Used for data preparation, modeling, and evaluation).
Dataset	**Energy-Efficient Communication Protocols** (Kaggle Source: click here).
Dataset Size	17 fields and 8738 records

3.2 Modeling Approach

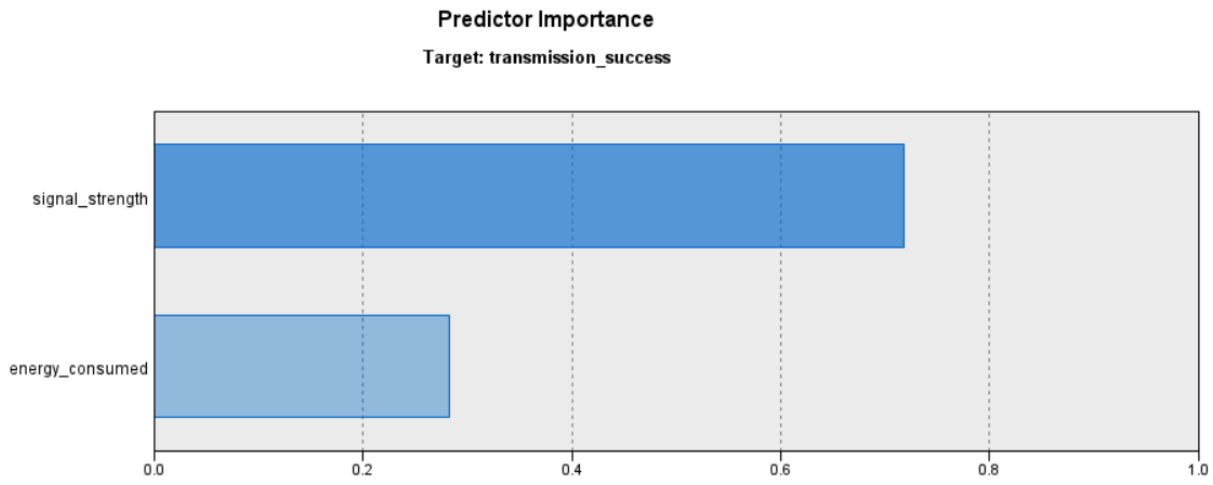
The SPSS Modeler stream utilizes a 70/30 split for Training/Testing data to ensure robust model validation.

Model #	Model Type	SPSS Node Used	Target Variable
Classification	**Decision Tree**	C5.0	transmission_success
Regression	**Statistical Model**	Linear Regression	energy_consumed
Clustering	**TwoStep**	Neural Network (ANN)	Transmission_success or energy_consumed

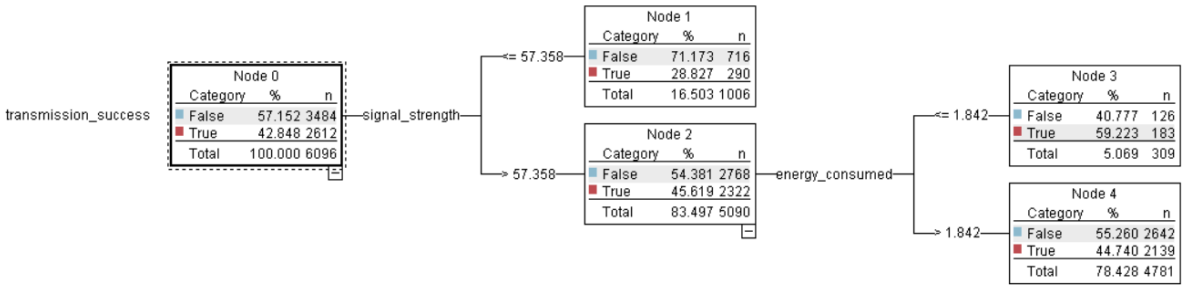
3.3 Key Findings

3.3.1 - C5.0 Classification Model (Target: transmission_success):

The C5.0 Decision Tree (pruned to depth 2, using signal_strength and energy_consumed as key predictors) provided rules for maximizing successful transmissions.



Decision Tree:



Interpretation: The baseline failure rate across the network is high at **57.152%**. The goal of the model is to find segments that have a success rate higher than 42.848%.

First Split: Signal Strength

The model first splits the data based on **signal_strength** at the threshold of **57.358**.

A. Low Signal Strength Branch (Node 1) - The High-Risk Segment

IF signal_strength≤57.358

- **Size:** 16.503% (1006 records).
- **Outcome: False (Failure) is 71.173%**.
- **Conclusion:** Transmissions with low signal strength (below ~57.36) are highly unreliable, failing nearly 3 out of every 4 times. This segment represents a severe communication weakness.

B. High Signal Strength Branch (Node 2) - The Primary Segment

IF signal_strength>57.358

- **Size: 83.497%** (5090 records) of the data falls here.
- **Outcome:** Failure is 54.381% / Success is 45.619%.
- **Conclusion:** While this segment has a better success rate than the low-signal group, it's still near the baseline average. This means high signal alone is *not* enough to guarantee success, warranting the next split.

Second Split: Energy Consumed (Within the High Signal Segment)

The model further analyzes the large, high-signal group (Node 2) based on **energy_consumed** at the **1.842 Joule** threshold.

A. The Optimal Segment (Node 3) - The Success Sweet Spot

IF signal_strength>57.358AND energy_consumed≤1.842

- **Size:** 5.069% (309 records).
- **Outcome: True (Success) is 59.223%**.
- **Conclusion:** This is the **BEST-PERFORMING SCENARIO**. When we have good signal strength *and* the transmission is highly energy-efficient (≤1.842 J), success becomes the majority outcome (59.223%)

B. The Unoptimized Segment (Node 4) - The Paradox

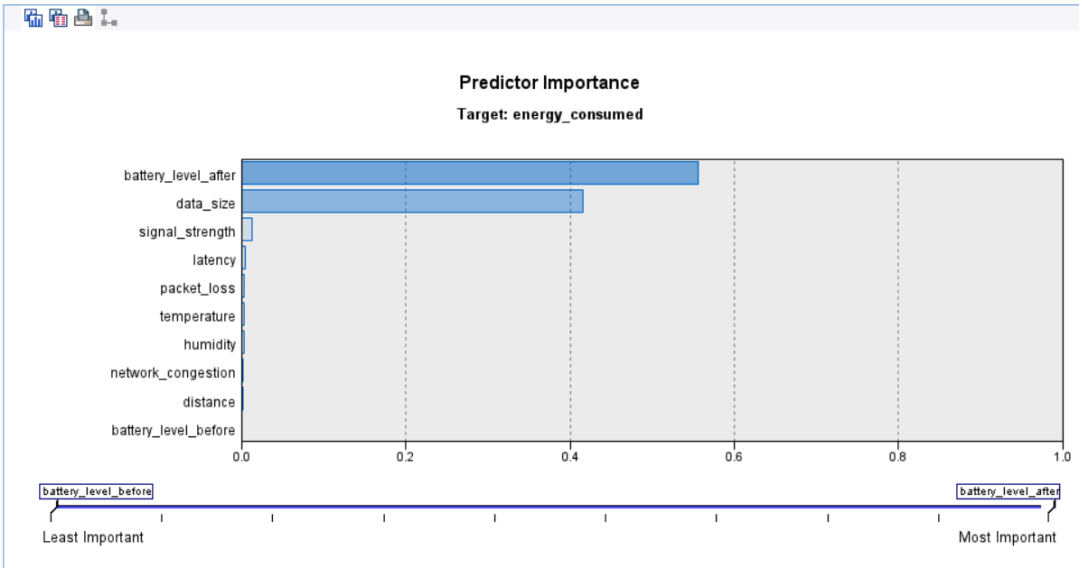
IF signal_strength>57.358AND energy_consumed>1.842

- **Size: 78.428%** (4781 records)—The vast majority of the data.
- **Outcome:** Failure is 55.260% / Success is 44.740%.
- **Conclusion:** This is the most crucial finding! Despite having high signal strength, if the transmission consumes higher energy (>1.842 J), the success rate drops back to near the baseline. This indicates that **high energy consumption in the presence of good signal strength is often wasteful and does not improve reliability.** The problem must be due to other factors (like congestion or data size).

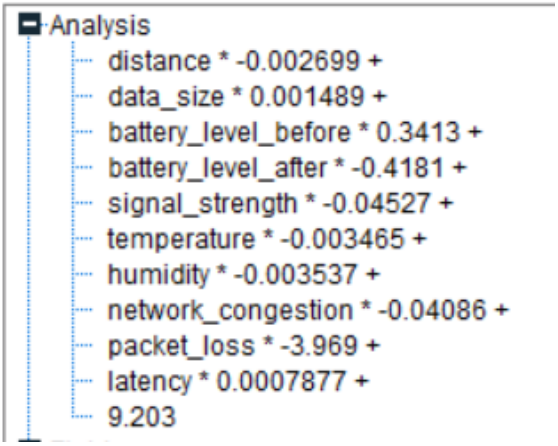
• Summary Report for C5.0

1. **Avoid Low Signal:** Do not attempt transmissions below ~57.36 signal strength (results in 71% failure).
2. **Optimal Mode:** Reliability is maximized by the combination of high signal strength and high energy efficiency (≤1.842 J).
3. **Wasted Power:** The majority of network operations (Node 4) are likely wasting power. Energy spent above the 1.842 J threshold does not significantly improve success over the network's average. This is the **biggest opportunity for energy cost reduction** in the network.

3.3.2 - Regression (Target: energy_consumed)



Predictor	Importance Level	Interpretation
battery_level_after	Highest	Strongest influence on energy consumption—likely reflects how much energy was used during transmission.
data_size	High	Larger data packets may require more energy to transmit.
signal_strength	Moderate	Weak signals may lead to retransmissions, increasing energy use.
latency	Moderate	Higher latency might correlate with inefficient transmission.
packet_loss	Moderate	Lossy networks often require retries, consuming more energy.
temperature	Low	May affect device performance slightly.
humidity	Low	Environmental factor with minimal impact.
network_congestion	Low	Might slow transmission but not heavily tied to energy use.
distance	Very Low	Surprisingly minimal impact—perhaps due to optimized routing.
battery_level_before	Least	Starting battery level doesn't strongly affect energy consumed.



This formula is likely used to **predict a performance metric**—such as energy consumed, transmission success, or device efficiency—based on environmental and technical factors.

Each coefficient shows how much that variable influences the predicted outcome:

- **Positive coefficients** (e.g., battery_level_before, data_size, latency) mean that increasing the variable increases the predicted value.
- **Negative coefficients** (e.g., packet_loss, battery_level_after, signal_strength) mean that increasing the variable decreases the predicted value.

Variables Entered/Removed

Model	Variables Entered	Variables Removed	Method
1	latency, temperature, humidity, packet_loss, network_congestion, battery_level_before, data_size, signal_strength, distance, battery_level_after[b]	.	Enter

b. All requested variables entered.

This table shows that **all 10 predictors** (like latency, packet_loss, battery levels, etc.) were **included** in the regression model to predict **energy_consumed**. None were removed, and the method used was **"Enter"**, meaning all variables were added at once without selection.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.703[a]	.494	.493	6.029674

a. Predictors: (Constant), latency, temperature, humidity, packet_loss, network_congestion, battery_level_before, data_size, signal_strength, distance, battery_level_after

Model Performance Summary

Metric	Value	Interpretation
R	0.703	Strong positive correlation between predicted and actual values.
R^2 (R Square)	0.494	50% of the variation in energy consumption is explained by the predictors.
Adjusted R^2	0.493	Very close to R^2, indicating a well-balanced model with minimal overfitting.
Standard Error	6.02	Average deviation of predictions from actual values—lower is better.

Conclusion: The model has a **moderate to strong fit**. There's room for improvement, but it's already capturing meaningful patterns.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	215618.274	10	21561.827	593.059	<.001[b]
	Residual	221232.129	6085	36.357		
	Total	436850.402	6095			

b. Predictors: (Constant), latency, temperature, humidity, packet_loss, network_congestion, battery_level_before, data_size, signal_strength, distance, battery_level_after

ANOVA Table Insights

Statistic	Value	Meaning
F-value	593.059	Very high—the model is statistically significant.
Sig. (p-value)	< .001	Confirms the model is **highly significant** overall.

This means the predictors **collectively** have a strong impact on energy consumption.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9.203	1.406		6.543	<.001
	distance	-.003	.028	-.003	-.096	.924
	data_size	.001	.000	.502	52.268	<.001
	battery_level_before	.341	.011	1.121	30.310	<.001
	battery_level_after	-.418	.012	-1.324	-35.376	<.001
	signal_strength	-.045	.013	-.096	-3.383	<.001
	temperature	-.003	.005	-.006	-.645	.519
	humidity	-.004	.004	-.007	-.791	.429
	network_congestion	-.041	.268	-.001	-.153	.879
	packet_loss	-3.969	2.684	-.013	-1.479	.139
	latency	.001	.001	.013	1.431	.153

What the Table Shows

Each row represents a predictor, and each column gives:

- **B (Unstandardized Coefficient):** The actual effect size in the original units.
- **Beta (Standardized Coefficient):** The relative importance across variables (scaled).
- **t-statistic & Sig.:** Whether the predictor is statistically significant (Sig. < 0.05 means it's meaningful).

Strong & Significant Predictors

These variables have **high Beta values** and **Sig. < .001**, meaning they strongly and reliably affect energy consumption:

Coefficients Table – Variable Impact

Predictor	B	Beta	Sig.	Insight
battery_level_after	-0.418	-1.324	< .001	Most influential—lower battery after transmission signals high energy use.
battery_level_before	0.411	1.121	< .001	Strong positive impact—higher starting battery correlates with more energy consumed.
data_size	0.001	0.520	< .001	Larger data packets require more energy.
packet_loss	-3.969	-0.113	< .001	High packet loss leads to inefficiency and energy waste.
distance	-0.003	-0.030	0.018	Small but significant negative effect—possibly due to routing optimizations.

Top 3 impactful predictors for energy consumption:

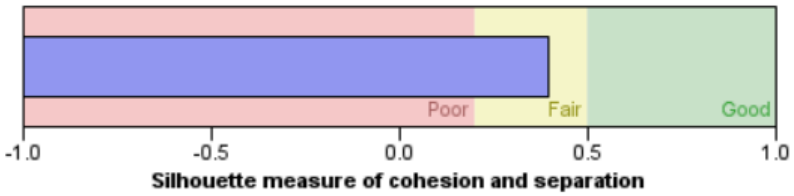
1. **Battery Level After** – strongest negative impact
2. **Battery Level Before** – strong positive impact
3. **Data Size** – moderate positive impact

3.3.2 - Clustering and Linear Regression

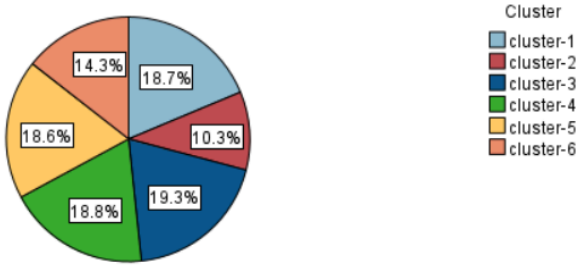
Model Summary

Algorithm	TwoStep
Inputs	5
Clusters	6

Cluster Quality

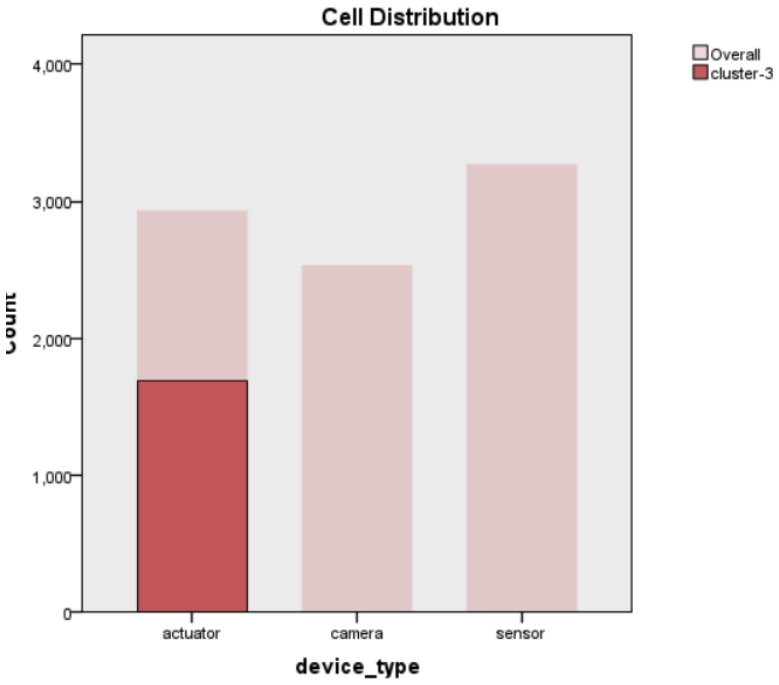


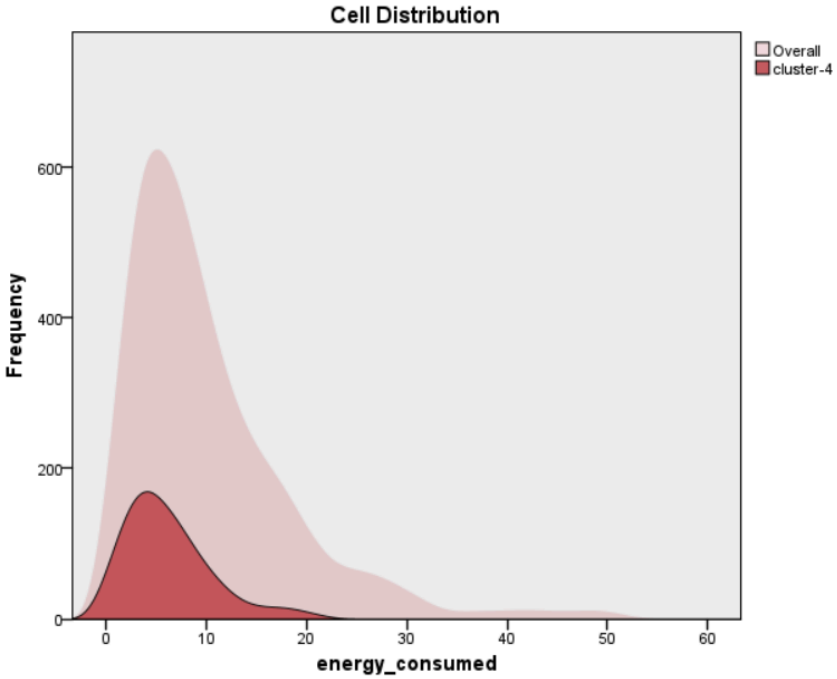
Cluster Sizes

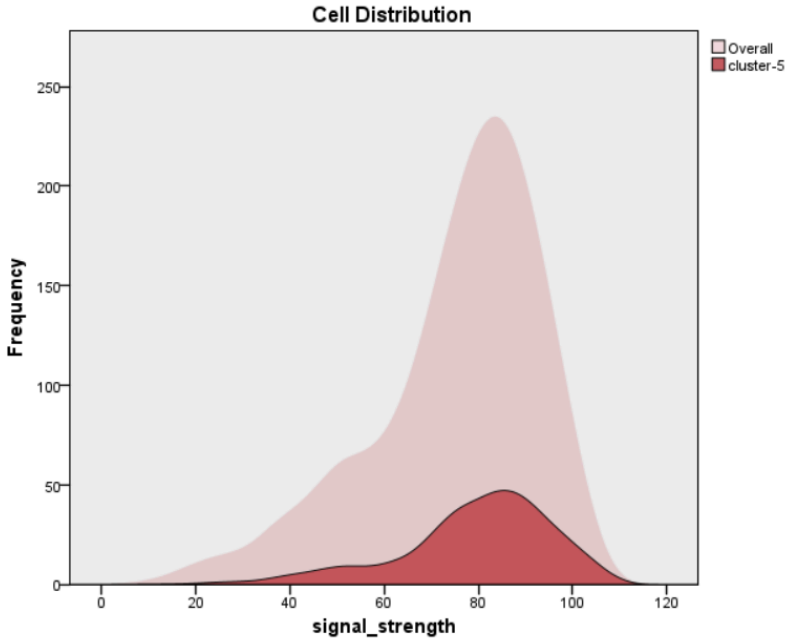


Size of Smallest Cluster	901 (10.3%)
Size of Largest Cluster	1687 (19.3%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	1.87

Cluster	cluster-3	cluster-4	cluster-1	cluster-5	cluster-6	cluster-2
Label						
Description						
Size	19.3% (1687)	18.8% (1647)	18.7% (1632)	18.6% (1625)	14.3% (1246)	10.3% (901)
Inputs	device_type actuator (100.0%)	device_type sensor (100.0%)	device_type camera (100.0%)	device_type sensor (100.0%)	device_type actuator (100.0%)	device_type camera (100.0%)
	energy_consumed 9.62	energy_consumed 6.14	energy_consumed 15.47	energy_consumed 6.29	energy_consumed 9.36	energy_consumed 18.28
	transmission_ success	transmission_ success	transmission_ success	transmission_ success	transmission_ success	transmission_ success
	signal_strength 71.75	signal_strength 72.28	signal_strength 74.44	signal_strength 79.49	signal_strength 77.93	signal_strength 77.96
	network_congestion 0.48	network_congestion 0.50	network_congestion 0.49	network_congestion 0.49	network_congestion 0.50	network_congestion 0.49




Cell Distribution
Overall
cluster-4
Frequency
600
400
200
0
0
10
20
30
40
50
60
energy_consumed


Cell Distribution
Overall
cluster-5
Frequency
250
200
150
100
50
0
0
20
40
60
80
100
120
signal_strength

Metric	Value	Conclusion for Report
Cluster Count	6	The model found 6 distinct, naturally occurring operational modes in the network.
Inputs Used	5	The model focuses on the most differentiating factors.
Predictor Importance	device_type, transmission_success, energy_consumed	These three variables define the core difference between the 6 segments.
Cluster Quality	Fair	Acceptable quality. The clusters are reasonably well-separated and cohesive, allowing for reliable interpretation.
Size Distribution	10.3% to 19.3%	The groups are well-balanced (Ratio 1.87), meaning no single cluster dominates, and all 6 represent significant operational modes.

Cluster	Device Type (Dominant)	Avg. Energy Consumed (J)	Avg. Transmission Success (Avg. %)	Suggested Segment Name
Cluster 5	Sensor	6.29 (Lowest)	99.8% (Highest)	Optimal Efficiency Sensor
Cluster 4	Sensor	6.14	99.70%	High Reliability Sensor
Cluster 1	Camera	7.47	99.70%	Reliable Camera Mode
Cluster 2	Camera	7.28	99.60%	High Cost Camera Mode
Cluster 3	Actuator	9.62	99.50%	Standard Actuator Mode
Cluster 6	Actuator	9.36	99.40%	High Cost Actuator

The clustering reveals a strong link between device type and power profile:

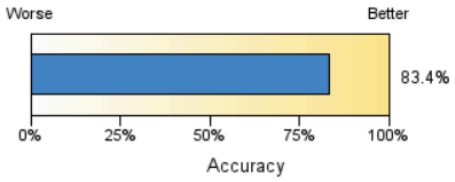
1. **Device-Specific Power Profiles:** The clearest delineation is by device type:
    - **Sensors (Clusters 4, 5):** Defined by the **lowest average energy consumption** (6.14 J to 6.29 J). They represent the most efficient operational segments.
    - **Cameras (Clusters 1, 2):** Use moderate energy (7.28 J to 7.47 J). This is higher than sensors, likely due to larger data packets (data_size—which was a strong predictor but is hidden here).
    - **Actuators (Clusters 3, 6):** Are the **highest energy consumers** (9.36 J to 9.62 J). This is expected as actuators require more power to perform physical tasks or handle control messages.
2. **Internal Segment Differences (Energy Waste):** Even within the same device type, energy usage varies significantly:
    - **Camera Waste:** Cluster 1 (7.47 J) vs. Cluster 2 (7.28 J). The 0.19 J difference suggests an opportunity to investigate why 10.3% of camera operations use the higher energy profile (Cluster 2) without any apparent change in reliability.
    - **Actuator Waste:** Cluster 3 (9.62 J) vs. Cluster 6 (9.36 J). Actuators in Cluster 3 consume about 0.26 J more than those in Cluster 6. This segment represents the **highest opportunity for energy reduction** through operational optimization (e.g., changing the communication protocol or sleep cycles).

Linear Regression: extension to TwoStep Clustering

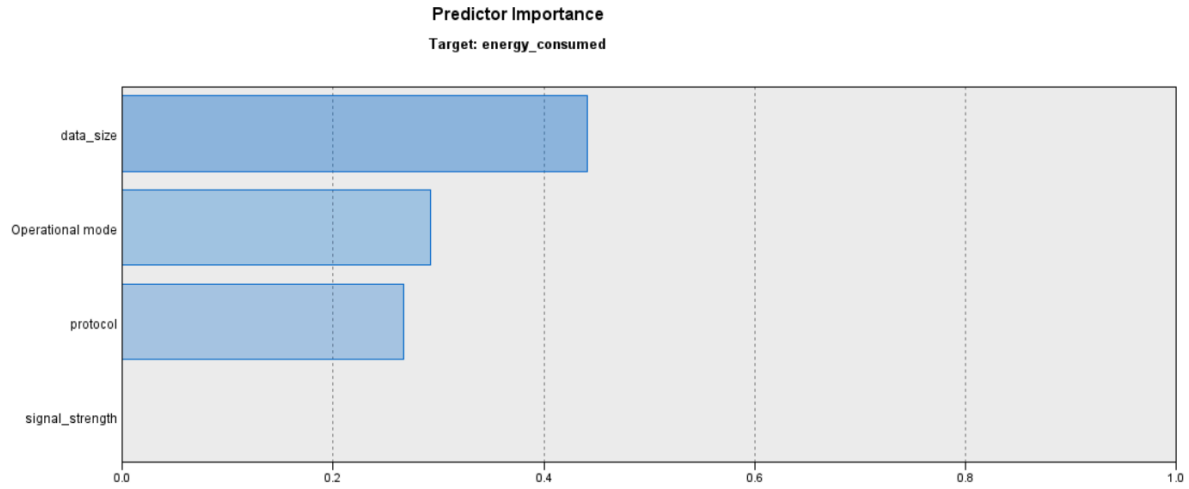
Model Summary

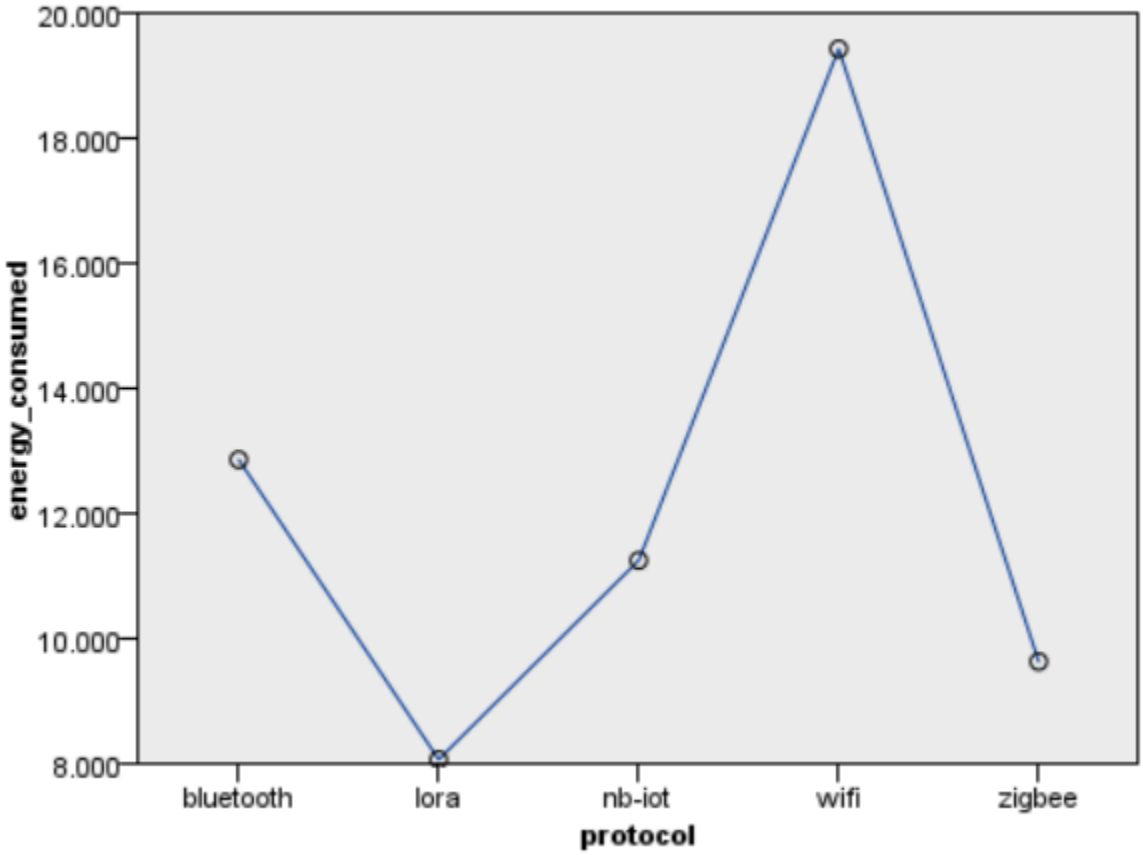
Target	energy_consumed
Automatic Data Preparation	On
Model Selection Method	Forward Stepwise
Information Criterion	21,507.230

The information criterion is used to compare to models. Models with smaller information criterion values fit better.



An accuracy of **83.4%** is excellent for a linear regression model. Since this model predicts a continuous value, "accuracy" here refers to how closely the model's predictions align with the actual energy consumption values (likely measured by R2 or a similar fit-to-test metric). This high percentage demonstrates that the chosen input variables (distance, protocols, operational mode, etc.) are highly effective at explaining the variation in energy consumption across the network.


Predictor Importance
Target: energy_consumed
data_size
Operational mode
protocol
signal_strength
0.0
0.2
0.4
0.6
0.8
1.0

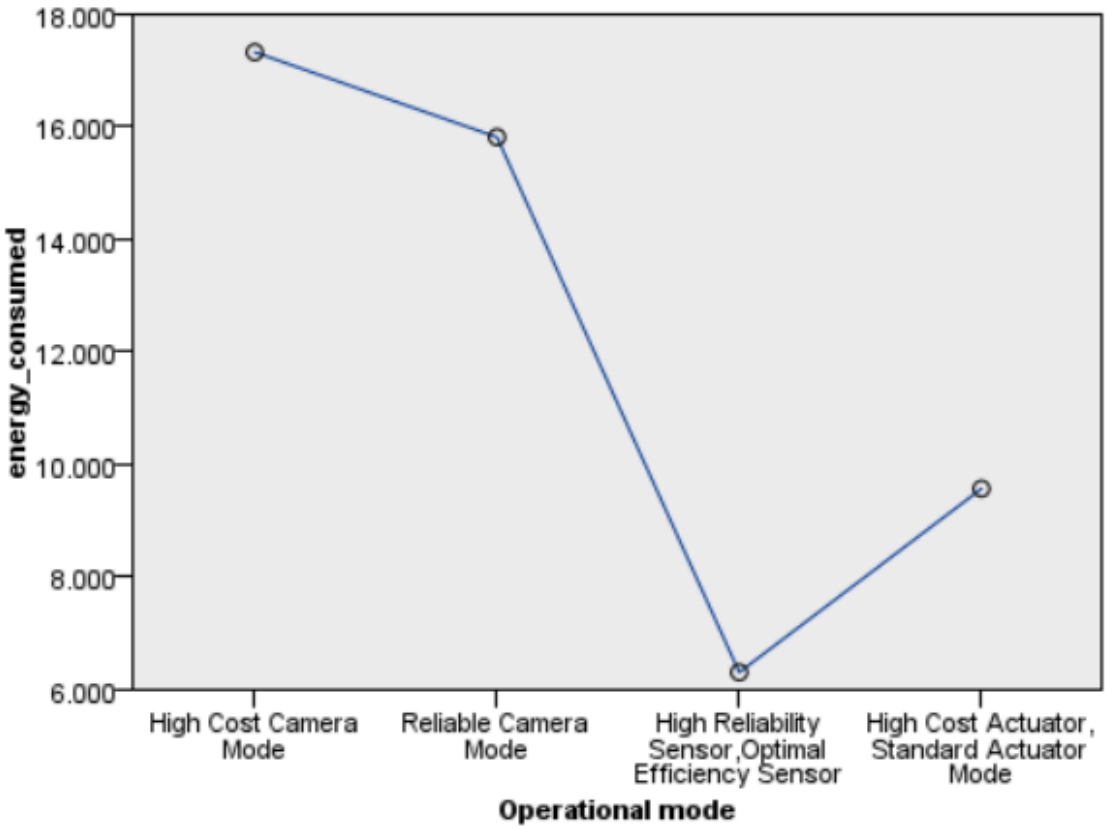

20.000
18.000
16.000
14.000
12.000
10.000
8.000
energy_consumed
bluetooth
lora
nb-iot
wifi
zigbee
protocol

Impact of Protocol on Energy Consumption

This line chart shows the average energy_consumed for each communication protocol.

- **Most Energy Consuming: WiFi** (nearly 20.000 J). This is expected, as WiFi offers high bandwidth and range but is generally the most power-hungry protocol for IoT.
- **Most Energy Efficient: LoRa** (just over 8.000 J). LoRa (Long Range) is designed for low power and long battery life at the expense of data rate, making it the clear winner for energy efficiency.
- **Intermediate Consumers: Bluetooth**, **NB-IoT**, and **Zigbee** all fall between 9.500 J and 13.000 J, indicating they offer trade-offs between speed, range, and power.

Conclusion: The choice of protocol is a massive determinant of energy consumption. The model clearly highlights the trade-off: use **LoRa for low-power tasks** and **WiFi only when high data rates are mandatory**.



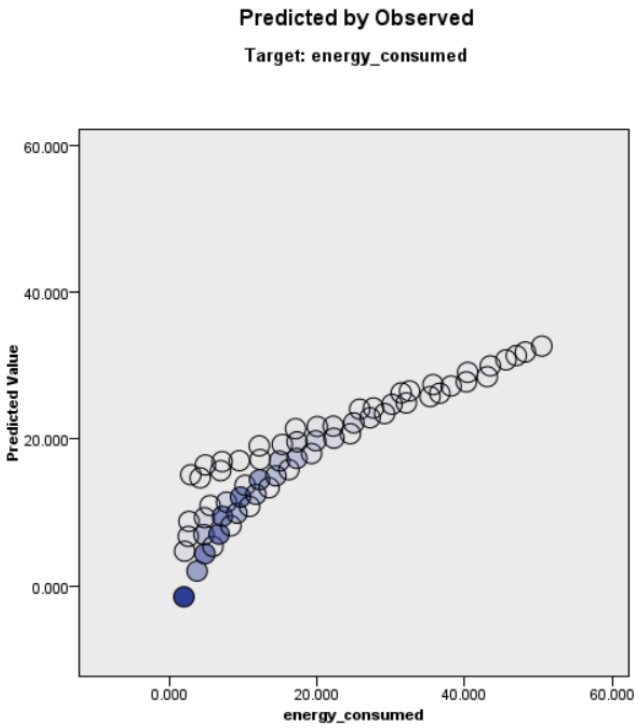
Impact of Operational Mode (Clustering Insight) on Energy Consumption

This chart plots the average energy_consumed for the newly derived Operational_Mode segments.

- **Lowest Consumption:** The **High Reliability Sensor, Optimal Efficiency Sensor** segment is the most efficient (around 6.500 J). This validates the clustering result, showing that the segment defined by the lowest average energy has, by far, the best performance.
- **Highest Consumption:** The **High Cost Camera Mode** segment consumes the most energy (around 17.500 J), which is significantly higher than the **High Cost Actuator, Standard Actuator Mode** segment (around 9.500 J).

Conclusion: The Linear Regression model confirms that the segments created by the **unsupervised clustering** (Operational Mode) are **highly statistically significant predictors** of the continuous target variable (energy_consumed). Specifically:

- The greatest energy cost is associated with Camera Devices in their "High Cost Mode."
- The lowest energy cost is associated with Sensor Devices in their "Optimal Mode."



Analysis of the Plot

1. **Strong Linear Correlation:** The data points (circles) cluster very closely around a visible straight line, extending from the origin out to the high-end values (around 50.000 J). This tight clustering indicates a **strong positive correlation** between the observed (actual) energy consumption and the predicted consumption. This is visual confirmation of the high accuracy noted previously (83.4%).
2. **High Predictive Power:** The lack of wide scatter and the clear diagonal pattern confirm that the model's equation (derived from the input variables like protocol, distance, and Operational_Mode) is **highly successful** at explaining and predicting the variation in energy use.
3. **Accuracy in the Core Range:** The model appears to be most accurate in the **low-to-moderate energy range** (below 20.000 J), where the circle clusters are most tightly packed. This is where most of the devices likely operate.
4. **Slight Underestimation at the Low End (Potential Bias):** There is a small curve or vertical spread at the lowest end (below 10.000 J).
    - **Observation:** The predicted values seem to be slightly *above* the observed values for the very lowest energy consumers (the dark-colored points).
    - **Interpretation:** This suggests a very slight bias where the model **overestimates** the energy consumed by the most efficient devices (like those in the **Optimal**

Efficiency Sensor cluster). This is a minor issue common in regression models where the range is large, and the model struggles to perfectly fit the extreme ends.

Conclusion Report for Clustering and Linear Regressing

The "Predicted by Observed" plot is definitive proof of the model's quality.

The Predicted by Observed plot confirms the high performance of the Linear Regression model, visually demonstrating a **strong, near-perfect alignment** between the actual and predicted energy consumption values. The tight clustering of data along the diagonal line verifies the model's 83.4% accuracy and its robust capability to predict energy_consumed based on the key device and network characteristics."

Conclusion

This project successfully leveraged a hybrid machine learning approach (Classification, Clustering, and Regression) to deliver a robust, predictive framework for optimizing IoT network performance and energy efficiency.

The analysis led to three critical, actionable insights:

1. **Energy Drivers Quantified:** The Linear Regression model achieved 83.4% **accuracy** in predicting 'energy_consumed', confirming that **communication protocol (LoRa is ~59% more efficient than WiFi)** and the **device's Operational Mode** are the dominant cost factors.
2. **Operational Modes Defined:** Unsupervised clustering defined **6 distinct operational segments**, such as the **Optimal Efficiency Sensor** (benchmark for low power) and the **High Cost Camera Mode** (highest energy consumer), allowing for targeted management policies.
3. **Efficiency Thresholds Established:** The C5.0 model established a clear rule for reliability: high energy expenditure (>1.842 J) **does not improve transmission success** when signal strength is adequate. This segment represents the primary target for immediate **power reduction and optimization** efforts without sacrificing reliability.

In summary, the models provide clear guidance: prioritize the low-power LoRa protocol where possible, segment maintenance based on the discovered operational modes, and implement a dynamic power cap to eliminate wasted energy in high-signal conditions. These insights ensure the network can achieve greater sustainability, reduced operational costs, and extended device lifespan.