Energy Consumption Analysis and Prediction

1. Data Summary and Basic Statistics

The dataset 'energydata_complete.csv' contains 19,735 observations and 29 columns.
It records household energy consumption data (Appliances in Wh) along with temperature, humidity, and weather conditions. There are no missing values, and most columns are numeric except 'date', which was converted to datetime format.

There are temperature and relative humdity is provided for eight different parts of the house.

Also T6 and RH_6 are temperature and RH for outside house and T_out and RH_out for weather station.

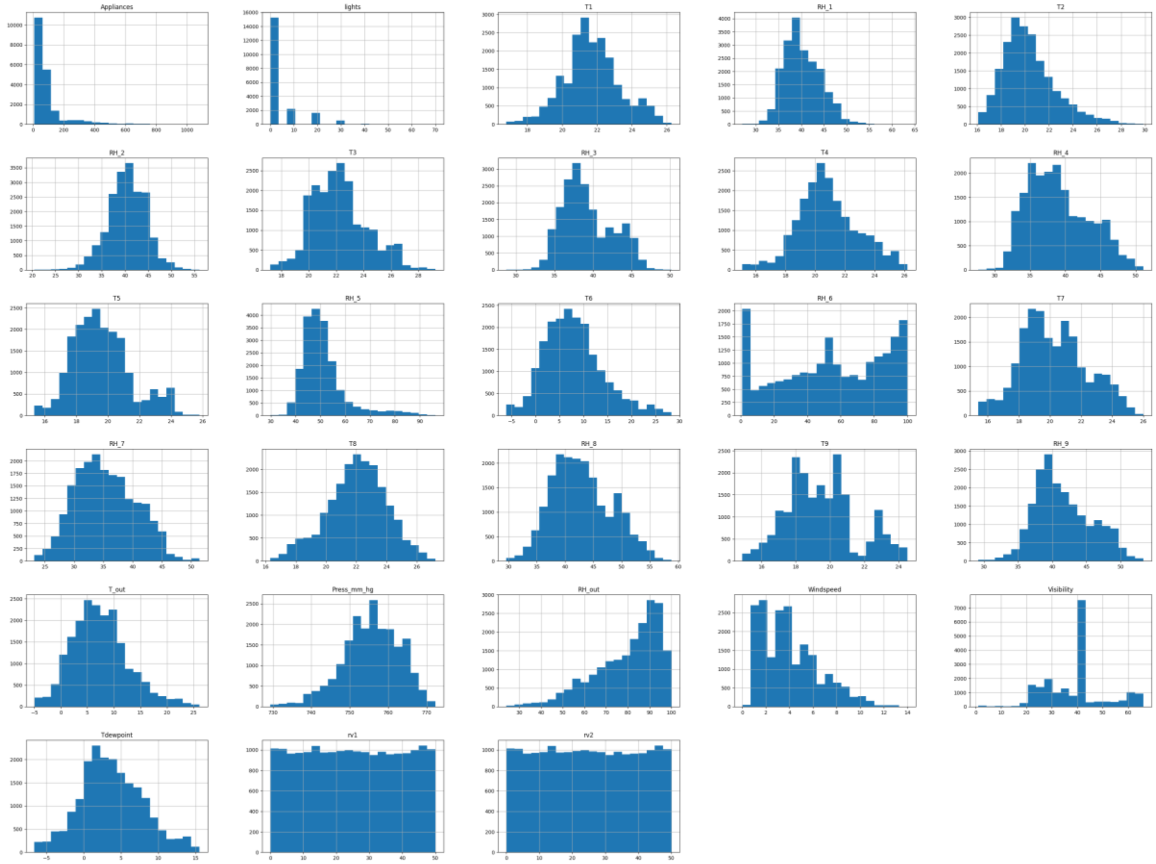
Date,Visbility, lights, Tdewpoint, Windspeed, rv1, and rv2 are other attributes

Basic Statistical Summary:
- Mean appliance energy use: 97.7 Wh
- Median: 60 Wh
- Standard deviation: 102.5 Wh
- Average indoor temperature: ~16.9°C
- Average humidity: ~46.4%

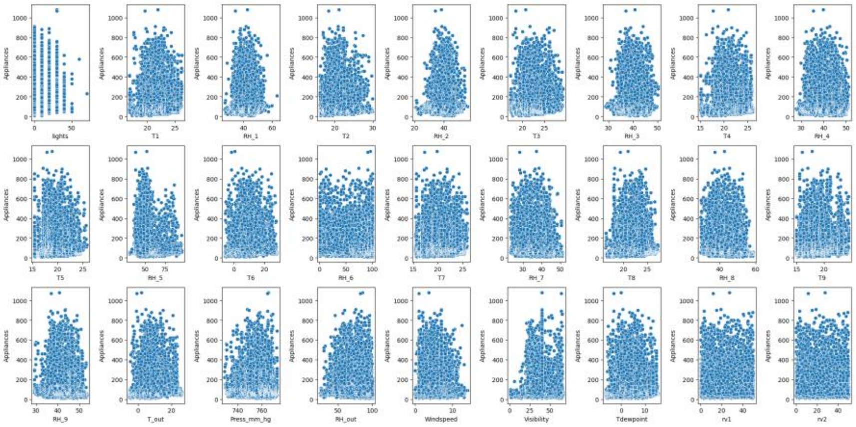
2. Exploratory Data Analysis (EDA)

Histogram

Most features are near-normal or slightly right-skewed.
'Appliances' and 'lights' have strong right tails, while temperature sensors show bell-shaped distributions.



Scatter Plot

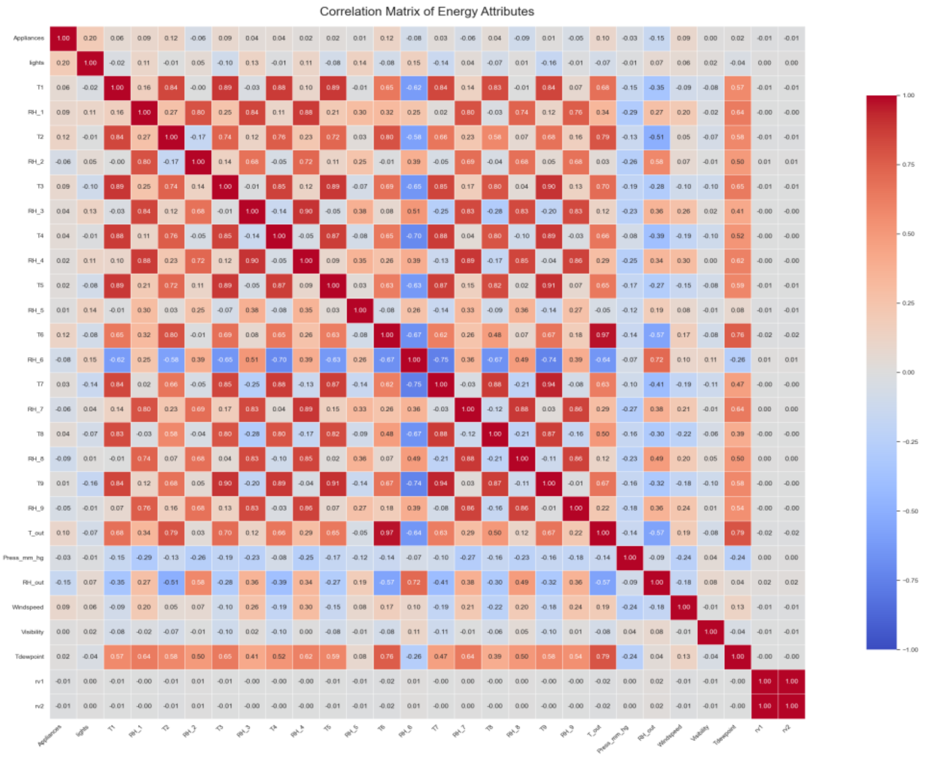


Indoor temperature features (like T2, T6, T5, T9) show positive correlations with 'Appliances'.
As room temperature increases, appliance usage tends to rise moderately.

KDE Plot Insights

Temperature features display unimodal, near-normal distributions.
'Appliances' and 'lights' are right-skewed, showing most readings are low with few peaks.

Correlation Heatmap



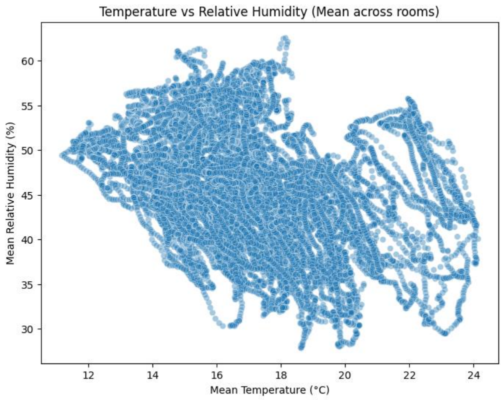
Strong multicollinearity is observed among temperature (T1–T9) and humidity (RH_1–RH_9) sensors.
'Appliances' correlates moderately with indoor temperatures and weakly with outdoor variables.
Appliances Correlation

Temperature-related features dominate the top correlations with 'Appliances'.

Boxplots

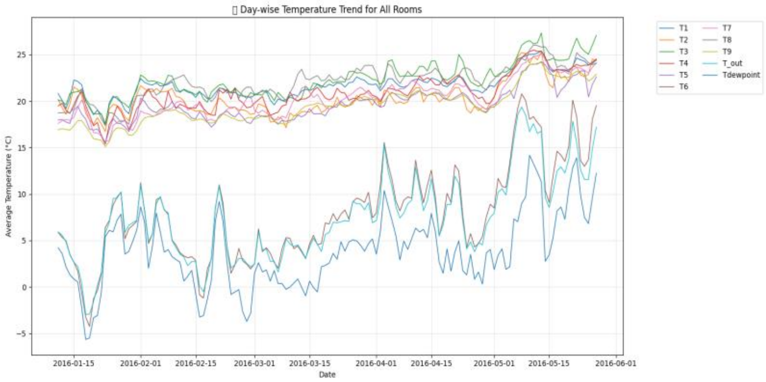
Outliers are present mainly in almost all columns mostly in 'Appliances' and 'lights'.
Temperature features show small IQRs, indicating stable sensor readings.

Temperature and Humidity Relation



An inverse relationship exists between temperature and humidity —
as temperature increases, relative humidity decreases.

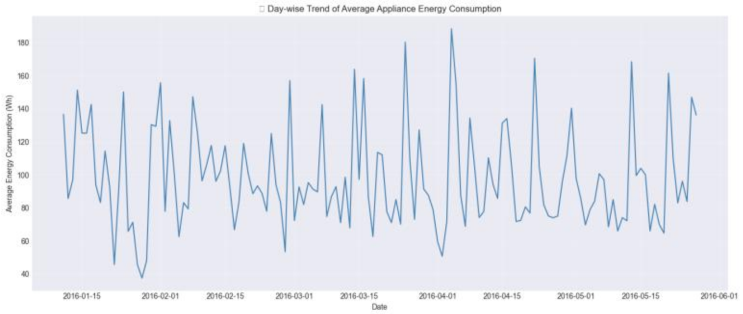
Time-Series Analysis: Temperature



Indoor room temperatures remain relatively stable throughout the day,
while outdoor temperatures fluctuate significantly — showing controlled indoor environments.

Time-Series Analysis: Appliances

Appliance energy usage fluctuates daily but shows clear weekly cycles.
A 7-day moving average highlights recurring peaks during weekends or extreme temperature days.



3. Preprocessing

1. Converted 'date' to datetime and extracted time features (year, month, day, hour, day_of_week, is_weekend).
2. Dropped non-informative columns: date, date_only, lights, rv1, rv2, Visibility.
3. Capped outliers using IQR to minimize extreme value impact.

Appliances: 2138 outliers capped. ◈ lights: 4483 outliers capped. ◈ T1: 515 outliers capped. ◈ RH_1: 146 outliers capped. ◈ T2: 546 outliers capped. ◈ RH_2: 235 outliers capped. ◈ T3: 217 outliers capped. ◈ RH_3: 15 outliers capped. ◈ T4: 186 outliers capped. ◈ T5: 179 outliers capped. ◈ RH_5: 1330 outliers capped. ◈ T6: 515 outliers capped. ◈ T7: 2 outliers capped. ◈ RH_7: 42 outliers capped. ◈ T8: 71 outliers capped. ◈ RH_8: 17 outliers capped. ◈ RH_9: 21 outliers capped. ◈ T_out: 436 outliers capped. ◈ Press_mm_hg: 219 outliers capped. ◈ RH_out: 239 outliers capped. ◈ Windspeed: 214 outliers capped. ◈ Visibility: 2522 outliers capped. ◈ Tdewpoint: 10 outliers capped.

4. Removed redundant features with correlation above 0.9 to reduce multicollinearity.

Dropped Columns: RH_3, RH_4, RH_6, RH_7, RH_9, T1, T4, T5, T6, T7, T8, T9, T_out, Temp_mean, month, rv2

5. Applied StandardScaler for normalization (mean=0, std=1).
6. Performed PCA retaining 95% variance for dimensionality understanding.

Remaining Attributes: PC1 – PC9

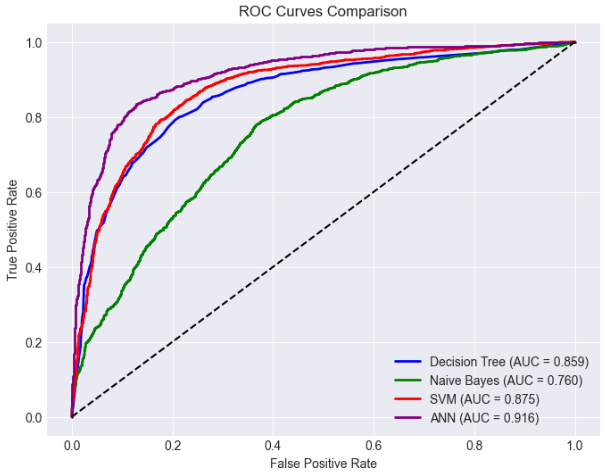
7. Created binary target variable: Appliances > median → 1, else 0.

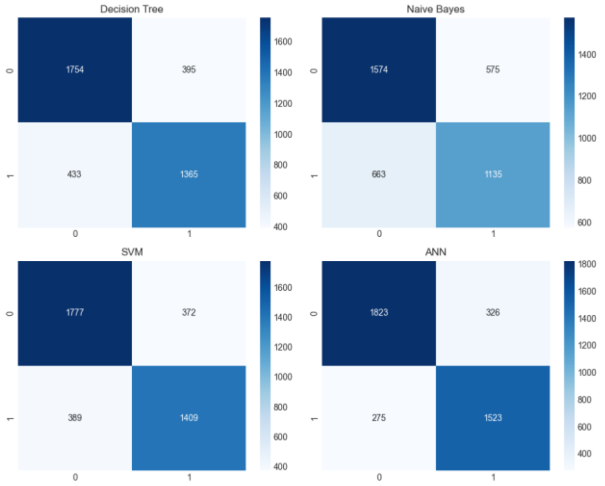
Median of Appliances = 60

4. Model Performance

All models were trained with an 80-20 train-test split on scaled data.
Each model's parameters and performance metrics are summarized below.

Model	Accuracy	Precision	Recall	F1-Score	ROC_AUC
Decision Tree	0.8753	0.8568	0.8721	0.8644	0.8751
GaussianNB	0.6752	0.6539	0.6096	0.6310	0.7445
SVM	0.8343	0.8217	0.8126	0.8171	0.9129
ANN	0.8885	0.8920	0.8593	0.8754	0.9522





5. Model Evaluation

The ANN Classifier achieved the best performance, with the highest Accuracy (0.8885) and ROC_AUC (0.9522), showing excellent discriminative power.
SVC performed similarly well (AUC ≈ 0.91), while Decision Tree was simpler but prone to overfitting.
GaussianNB lagged due to correlated and non-Gaussian data.

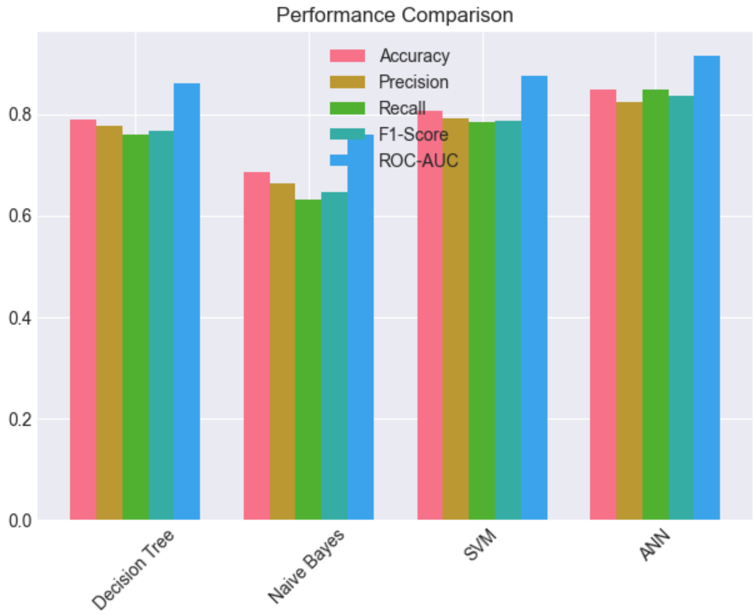
Best Model: ANN Classifier
- hidden_layer_sizes = (64, 32)
- activation = relu
- solver = adam
- max_iter = 300
- random_state = 42

 Insights:

• ANN provides best discrimination between classes (highest ROC area).

• Indoor temperature and humidity variations are key predictors.

• Future work: hyperparameter tuning, SHAP explanations, and inclusion of external weather forecasts.



Group Members	Contribution:
1. Ayush Patidar (Roll No. 251140004)	Data Collection and EDA
2. Ankit Kumar (Roll No. 251140002)	Report Writing
3. Subodh Singh (Roll No. 251140024)	Data Preprocessing
4. Vishal Dixit (Roll No. 251140026)	Model Preparation and evaluation