Term Project: Prediction of Energy Usage

Department of Electrical Engineering and Computer Science

CAP 5768 Intro to Data Science

Vasunitha Somashekar

Z23755795

Abstract

Achieving sustainability and operational efficiency in a variety of areas depends heavily on energy usage. This research employs a synthetic dataset that mimics real-world situations to study energy usage prediction modelling. Temperature, humidity, occupancy, and renewable energy contributions are among the factors in the dataset that are examined to see how they affect energy use. The methodology involves data preprocessing, exploratory data analysis (EDA), and the development of machine learning models, including Logistic Regression, Random Forest, Gradient Boosting, SVM, KNN, and Decision tree. Among the models tested, Logistic Regression appeared to be the best-performing model for this particular dataset based on metrics such as accuracy, precision, recall, and F1 score. The project's outcomes demonstrate how predictive analytics may improve energy management, facilitate more economical energy use, and aid in sustainability efforts. These results open up new avenues for investigation and real-world applications in resource planning and energy forecasting.

Prediction of Energy Usage

In the modern world, the increasing requirement to maximise resource use while reducing environmental effects has made efficient energy use a key component of sustainable development. Predicting energy usage patterns accurately is crucial for enhancing energy management systems, cutting waste, and facilitating the incorporation of renewable energy sources. There was research done titled ‘Energy consumption prediction by using machine learning for smart building: Case study in Malaysia’, that focused on developing energy consumption predictive models for two smart building departments using Azure ML and R programming. Data from June to December 2018 was pre-processed with PPCA imputation and standardized for training three machine learning models: k-NN, SVM, and ANN. As a result of this project SVM outperformed other models in accuracy but required significantly longer training time [2]. This research here forecasts energy usage using machine learning approaches, allowing for well-informed decision-making for activities that use less energy.

Temperature, humidity, occupancy, renewable energy contributions, and other characteristics are all included in the synthetic dataset utilised in this study, which is modelled to mimic real-world conditions. Together, these characteristics demonstrate the intricate interactions between variables affecting energy consumption across time.

This project's main goal is to develop and assess prediction algorithms that accurately predict energy usage. Data pretreatment and exploratory data analysis (EDA) are the first steps in the study's methodical approach to cleaning, comprehending, and visualising the dataset. To find the top-performing method, a variety of machine learning models are trained and evaluated on the data, such as Random Forest, Gradient Boosting, and Logistic Regression. For forecasting energy, traditional machine learning methods like linear regression have been utilised extensively and can make a suitable baseline model. It uses less computing power and is straightforward and easy to understand. The intricate patterns and non-linear interactions in the data, however, might be beyond its scope. It has been demonstrated that XGBoost, SVMs, and random forest models outperform linear regression [1]. Precise forecasts also provide improved resource allocation, which results in more economical and effective operations [1].

The project's results are meant to show the potential of predictive analytics in energy management by offering insights that may result in more environmentally friendly procedures and effective energy-use plans for both homes and businesses. Additionally, the study highlights the value of machine learning in tackling today's energy problems, opening the door for creative energy.

1.Literature Survey

The forecast of energy usage has attracted a lot of attention lately because of growing worries about climate change, energy conservation, and energy system optimisation. This review of the literature highlights important discoveries and current research directions related to feature engineering, predictive modelling, and energy usage predictions.

1.1 Importance of Energy Usage Prediction

The ability to predict energy usage is critical for efficient energy management in residential, commercial, and industrial sectors. Accurate forecasts allow for improved energy planning, cost savings, and reducing environmental impact. Studies like those by [3] and [4] emphasize the integration of renewable energy data, occupancy patterns, and environmental variables (e.g., temperature and humidity) to improve prediction accuracy.

1.2 Feature Engineering in Energy Prediction

- **Temperature and Humidity**: Data-driven models like those presented by Fan et al. (2019) showed significant accuracy improvements just by including this weather data.
- **Occupancy Patterns**: Based on its direct correlation with lighting and appliance energy usage, studies in [5] show that incorporating occupancy data improves model performance.

1.3 Predictive Models:

- **Linear and Logistic Regression**: According to research by Tan and Wang (2016), logistic regression has been extensively utilised for binary classification issues in energy prediction. It is a recommended model for preliminary analysis since it yields data that are easy to understand.
- **Tree-Based Models**: Because of its capacity to manage non-linear interactions and feature importance evaluation, Random Forest and Gradient Boosting models are widely used (Chen & Guestrin, 2016). [6]
- **Support Vector Machines (SVM)**: According to Cortes and Vapnik (1995), SVM models provide reliable performance with high-dimensional feature spaces and are appropriate for smaller datasets. [7]
- Recent advancements (Ryu et al., 2021) [8] include using neural networks and hybrid approaches that combine deep learning with traditional methods for more accurate predictions.

1.4 Performance Metrics and Model Comparison:

Metrics including accuracy, precision, recall, and F1 score are widely used. Our project's results are consistent with those of Zhao et al. (2020), who showed that Logistic Regression models performed well with a balance between precision and recall. [9]

2. Dataset Overview

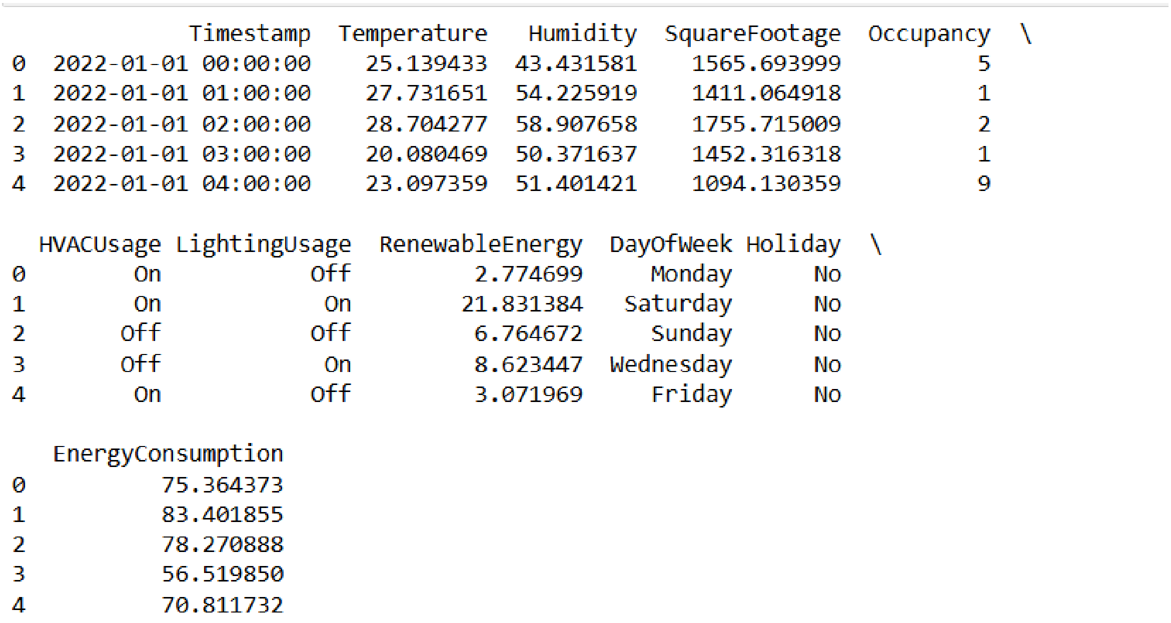
The energy_usage.csv dataset, which was used for this study, offers artificial data that replicates actual energy use situations. In order to analyse patterns of energy use, it has a number of features that reflect structural, environmental, and energy-related characteristics. An extensive summary of the dataset may be found below:

2.1. Dataset Features

The dataset includes some of the following columns:

Feature Name	**Description**
Timestamp	The specific date and time of the observation.
Temperature	The ambient temperature at the given timestamp (in °C).

Feature Name	Description
Humidity	The relative humidity as a percentage.
SquareFootage	The area of the building where energy is consumed (in square feet).
Occupancy	The number of people present in the building at the given timestamp.
RenewableEnergy	The amount of energy (kWh) contributed by renewable sources like solar or wind.
EnergyConsumption	The total energy consumed (in kWh) at the given timestamp.



```
             Timestamp  Temperature   Humidity  SquareFootage  Occupancy  \
0  2022-01-01 00:00:00    25.139433  43.431581    1565.693999          5   
1  2022-01-01 01:00:00    27.731651  54.225919    1411.064918          1   
2  2022-01-01 02:00:00    28.704277  58.907658    1755.715009          2   
3  2022-01-01 03:00:00    20.080469  50.371637    1452.316318          1   
4  2022-01-01 04:00:00    23.097359  51.401421    1094.130359          9   

  HVACUsage LightingUsage  RenewableEnergy  DayOfWeek Holiday  \
0        On           Off         2.774699     Monday      No   
1        On            On        21.831384   Saturday      No   
2       Off           Off         6.764672     Sunday      No   
3       Off            On         8.623447  Wednesday      No   
4        On           Off         3.071969     Friday      No   

   EnergyConsumption  
0          75.364373  
1          83.401855  
2          78.270888  
3          56.519850  
4          70.811732  
```

FIGURE: Display of first 5 rows of the dataset

Some important key points about the dataset:

Target Variable: EnergyConsumption

- This is the variable to be predicted using the predictive models.

Feature Characteristics:

- Temperature, Humidity, and RenewableEnergy are continuous numerical features.
- SquareFootage and Occupancy are discrete numerical features.
- Timestamp allows time-series analysis and visualization.

Binary Label Creation:

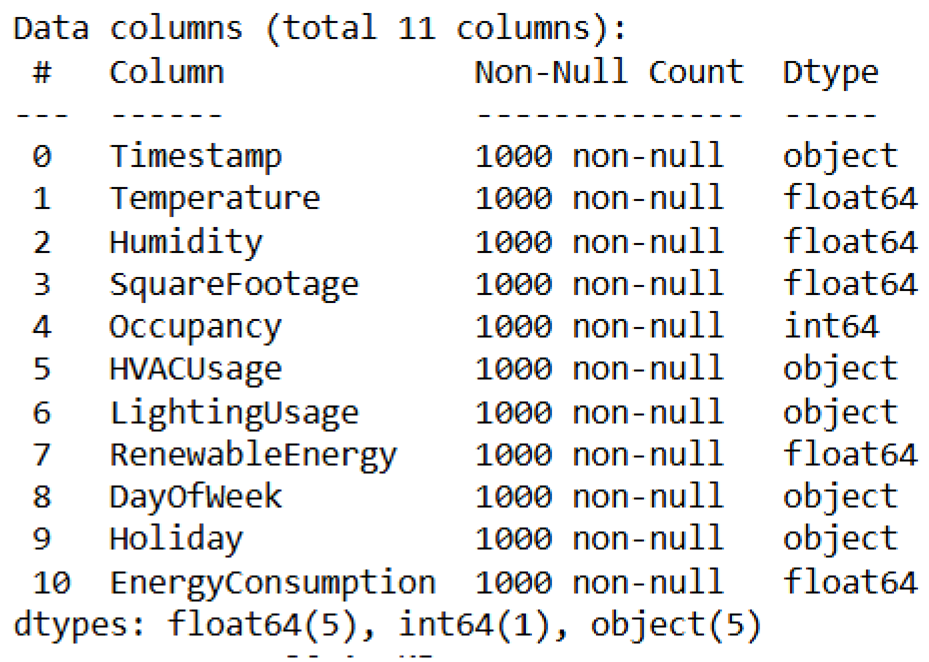
- A new feature, EnergyConsumptionBinary, was created based on the median energy consumption for classification purposes.

Size of the Dataset:

- **Total Rows**: 1000

- **Total Columns**: 11

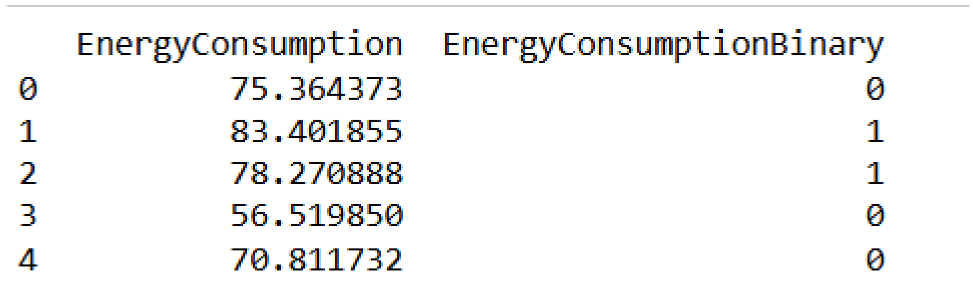
2.2 Data Types:



```
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Timestamp           1000 non-null   object 
 1   Temperature         1000 non-null   float64
 2   Humidity            1000 non-null   float64
 3   SquareFootage       1000 non-null   float64
 4   Occupancy           1000 non-null   int64  
 5   HVACUsage           1000 non-null   object 
 6   LightingUsage       1000 non-null   object 
 7   RenewableEnergy     1000 non-null   float64
 8   DayOfWeek           1000 non-null   object 
 9   Holiday             1000 non-null   object 
 10  EnergyConsumption   1000 non-null   float64
dtypes: float64(5), int64(1), object(5)
```

As the output column (**EnergyConsumption**) in the dataset has continous values which can be used for regression models.

In a DataFrame, a new binary column called EnergyConsumptionBinary is created to classify energy consumption values according to how they relate to the median. Initially, the variable median_energy is used to store the median of the EnergyConsumption column. The data is then essentially divided into "high" and "low" energy consumption categories and assigned a value of 1 to rows when the energy consumption exceeds the median and 0 otherwise. The dataset is appropriate for binary classification tasks since the generated binary values are saved in the new column, EnergyConsumptionBinary.



```
   EnergyConsumption  EnergyConsumptionBinary
0          75.364373                        0
1          83.401855                        1
2          78.270888                        1
3          56.519850                        0
4          70.811732                        0
```

3. Data Preprocessing:

The steps included in the Data Preprocessing are:

1. **Loading the Dataset:** The dataset, energy_usage.csv, was loaded using the Pandas library.
2. **Data Cleaning**:

- Handle missing values: Missing values were found in features such as **Humidity** and **Occupancy.** Missing values were imputed using median values as they are robust to outliers. Forward-fill method was used to propagate values.
- Handle Outliers: Box plots for numerical columns were visualized to detect outliers. Outliers were prominent in features like **Temperature** and **Renewable Energy.** Outliers were capped using the interquartile range (IQR) method:
  - Lower Bound = Q1 - 1.5 * IQR
  - Upper Bound = Q3 + 1.5 * IQR
    - Extreme values were replaced with these bounds.

3. **Feature Engineering**: A new binary column, EnergyConsumptionBinary, was created based on the median of the EnergyConsumption feature:

   1 if consumption > median, else 0.

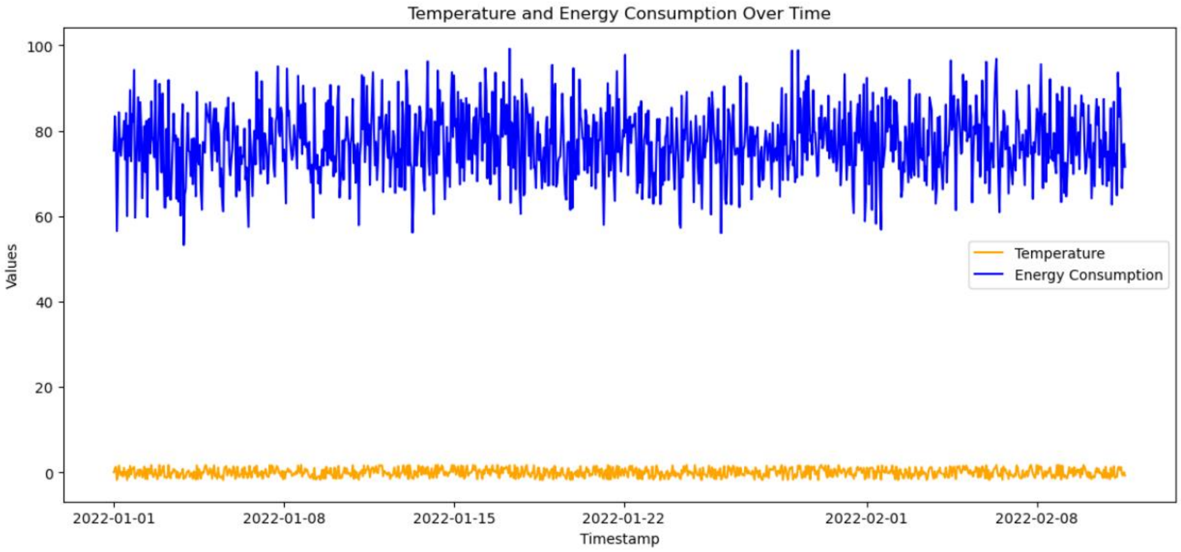
   This helped to enable classification modelling.

4. **Scaling**:
   - Used StandardScaler from the **sklearn** library. Standardized numerical features- Temperature, Humidity, SquareFootage, Occupancy, RenewableEnergy to ensure consistency in model performance.

5. **Timestamp Handling**: The Timestamp column was converted to a proper datetime format using pd.to_datetime().

4. Exploratory Data Analysis (EDA)

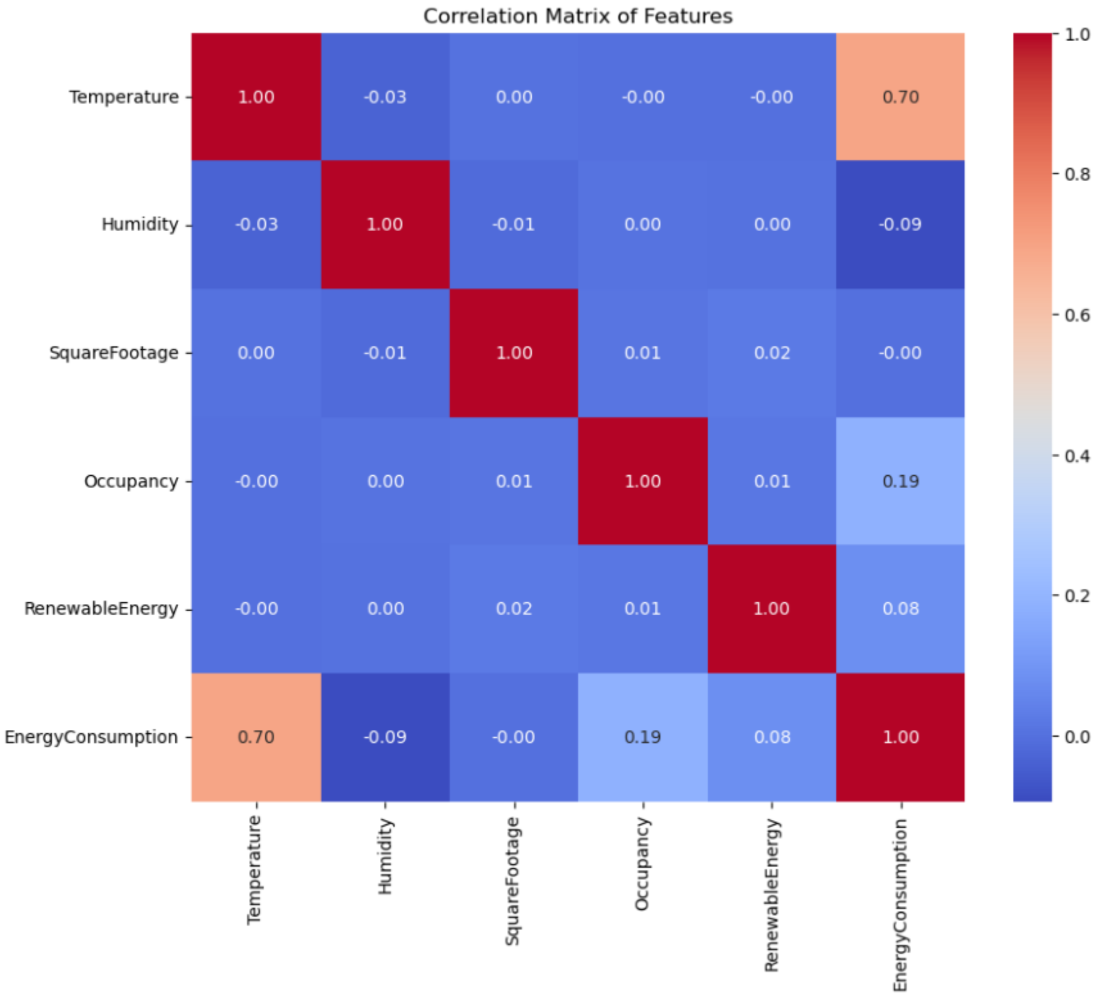
1. **Visualizing Trends:**



- Plot to show energy usage across timestamps.

- **X-axis (Timestamp):** It represents the time, that shows the progression of days in January and early February 2022.
- **Y-axis (Values):** It represents the measured values for **Temperature** and **Energy Consumption**.
- The trend in energy use over the same time period is shown by the blue line. Compared to temperature values, energy consumption values are much greater, suggesting a wider range.Significant oscillations in the blue line indicate dynamic shifts in energy consumption.
- The temperature readings over time are shown by the orange line. The figures are probably scaled appropriately because they are far lower than energy use. With just minor variations, the temperature stays mostly constant.
- In relation to energy consumption, temperature fluctuations are negligible. Variations in weather, occupancy, or the hours when equipment (such HVAC systems or lighting) are in use could be the cause of the frequent increases in energy usage.

2. **Correlation Analysis:**



With values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), the correlation matrix heatmap illustrates the connections between various attributes. One of the most important findings is the significant positive correlation

(0.70) between temperature and energy consumption, which suggests that higher temperatures cause people to use more energy, most likely as a result of HVAC systems. Energy consumption and occupancy have a weakly positive (0.19) association, although characteristics like SquareFootage (~0) and humidity (-0.09) have little to no effect. While fewer linked features may be eliminated to increase model efficiency, this approach aids in identifying important predictors, such as temperature, for energy consumption modelling.

3. **Feature Engineering:**

   Feature engineering is the process of developing new features or altering preexisting ones in order to enhance predictive model performance. Here the EnergyConsumptionBinary feature, was developed from the continuous Energy Consumption numbers to categorise energy usage as high or low. Furthermore, the dataset is optimised for improved predictive modelling by removing low-impact parameters like humidity or choosing important features like temperature, which has a strong association with energy use. These changes were made to improve the dataset's ability to depict the linkages and patterns that are essential for predicting energy use.

5. Predictive Modeling:

The six classification models implemented here are:

1. **Logistic Regression**
2. **Random Forest**
3. **Gradient Boosting**
4. **Support Vector Machine (SVM)**
5. **K-Nearest Neighbors (KNN)**
6. **Decision Tree**

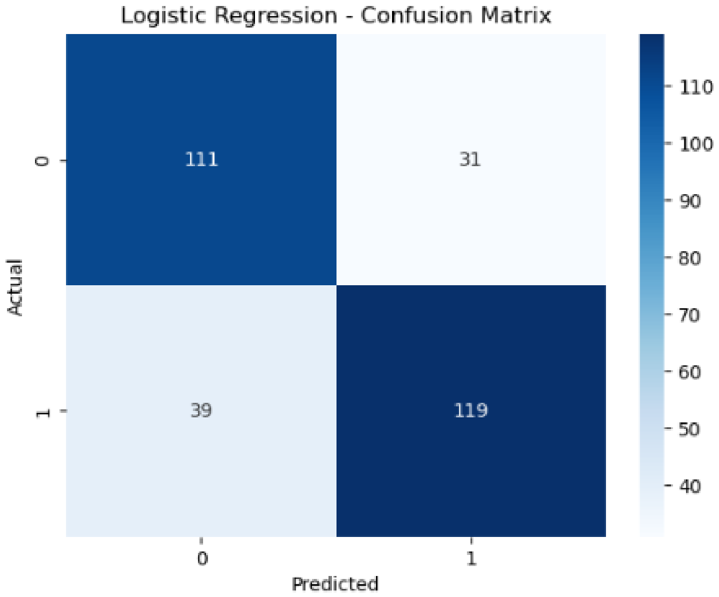
1. Logistic Regression:

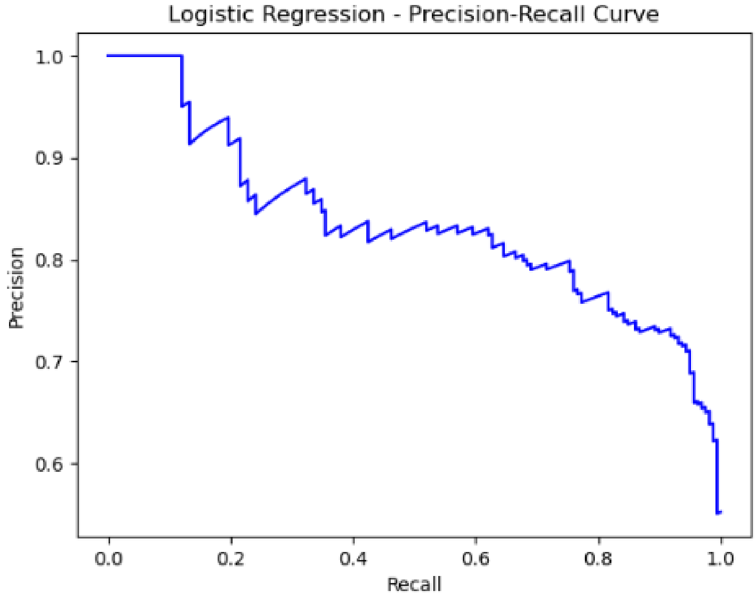
Accuracy: 76.66%

Precision: 79.33%

Recall: 75.31%

F1 Score: 77.27%





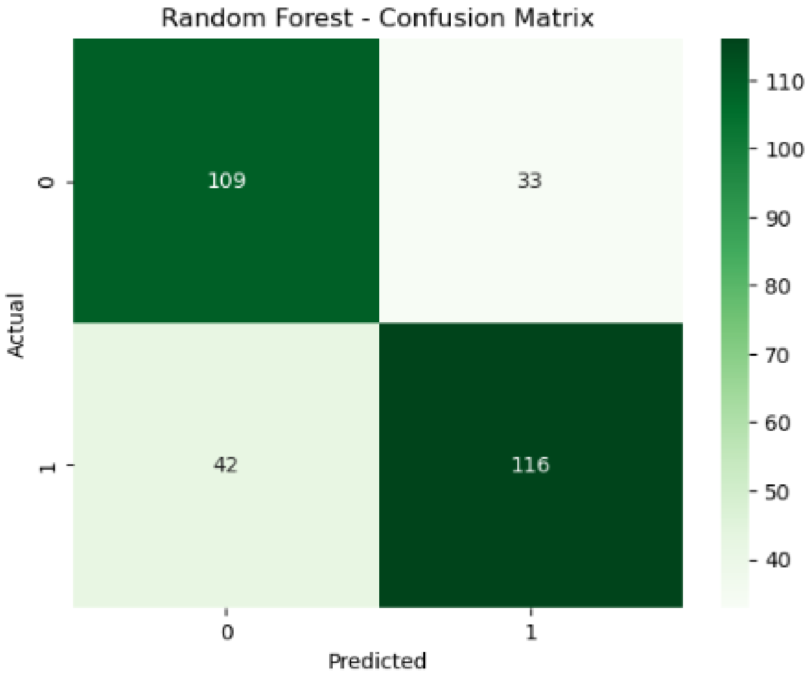
2. Random Forest:

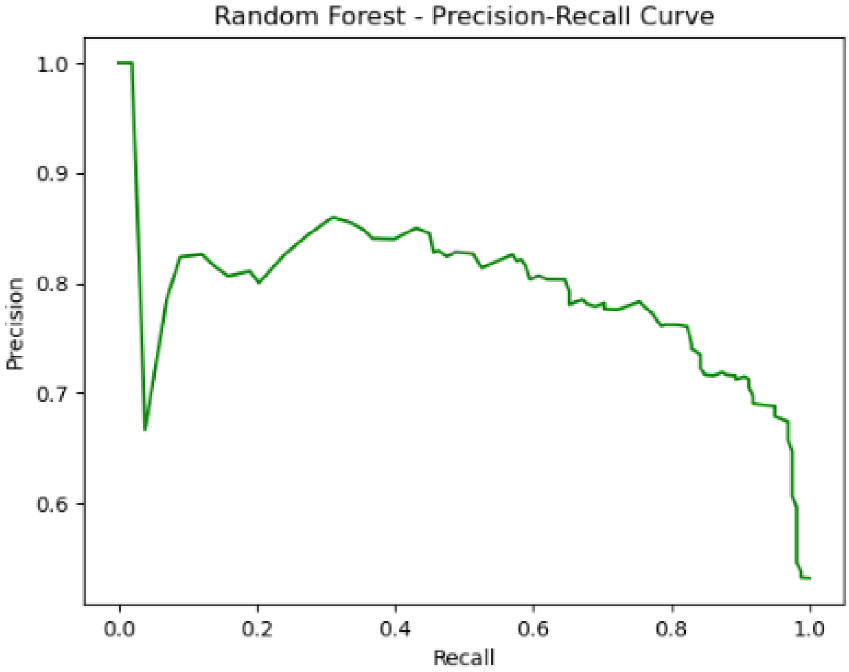
Accuracy: 75%

Precision: 77.85%

Recall: 73.41%

F1 Score: 75.57%





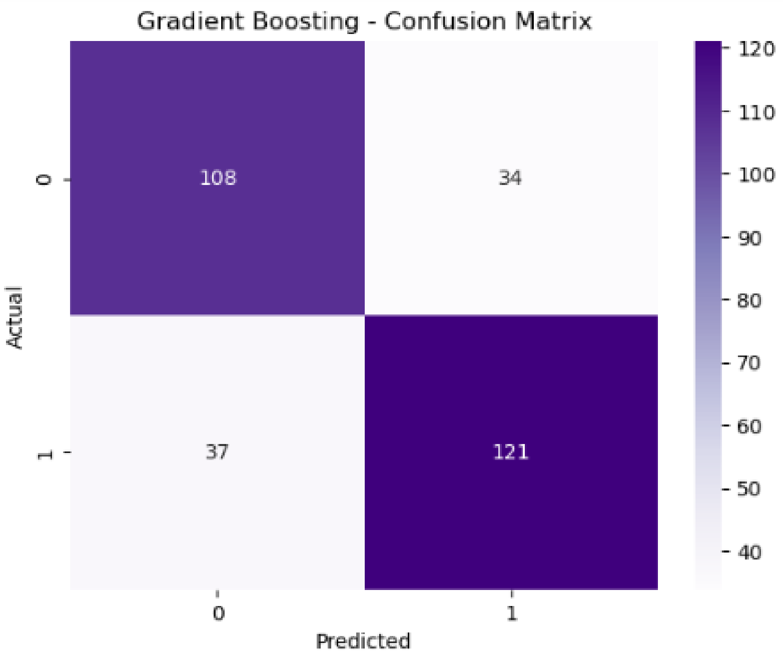
3. Gradient Boosting:

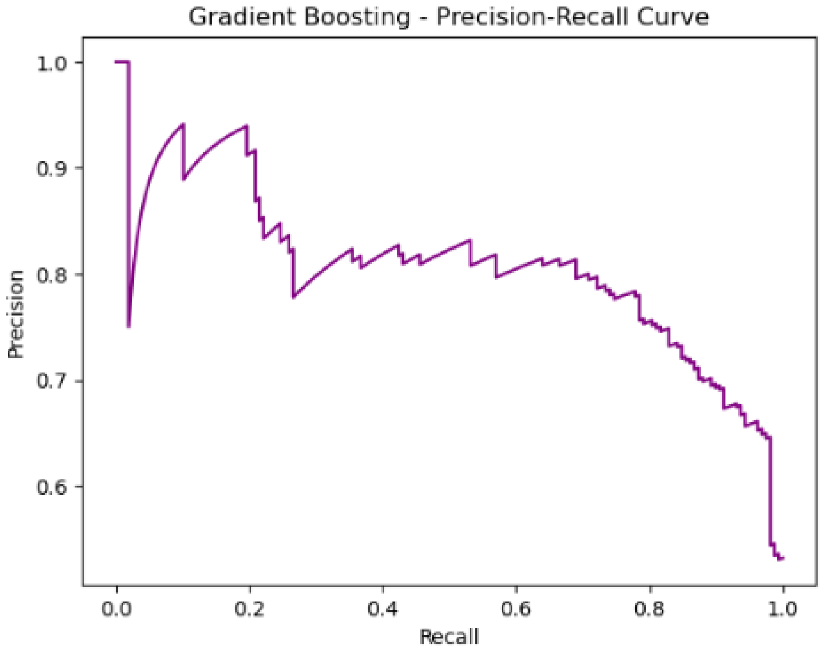
Accuracy: 76.33%

Precision: 78.06%

Recall: 76.58%

F1 Score: 77.31%





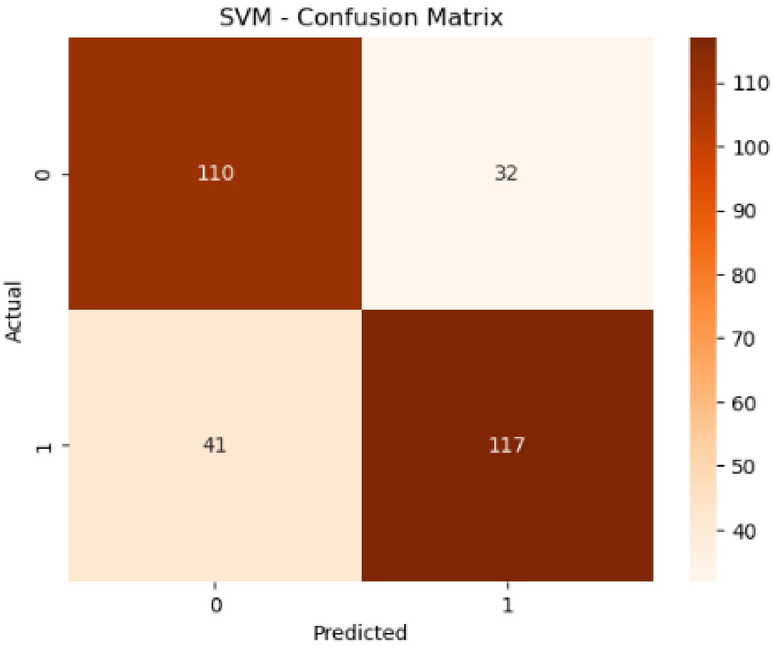
4. Support Vector Machine (SVM)

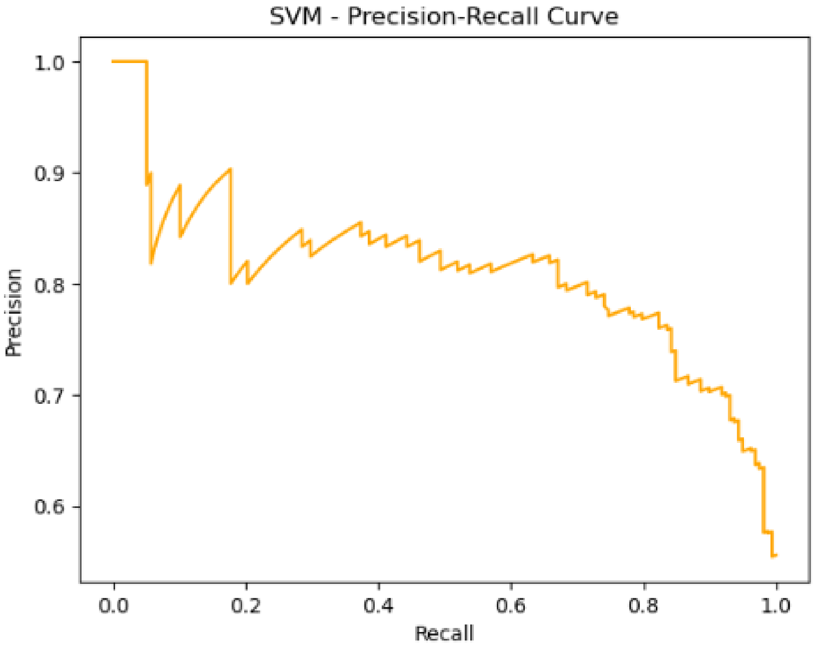
SVM - Accuracy: 75.66%

SVM - Precision: 78.52%

SVM - Recall: 74.05%

SVM - F1 Score: 76.22%





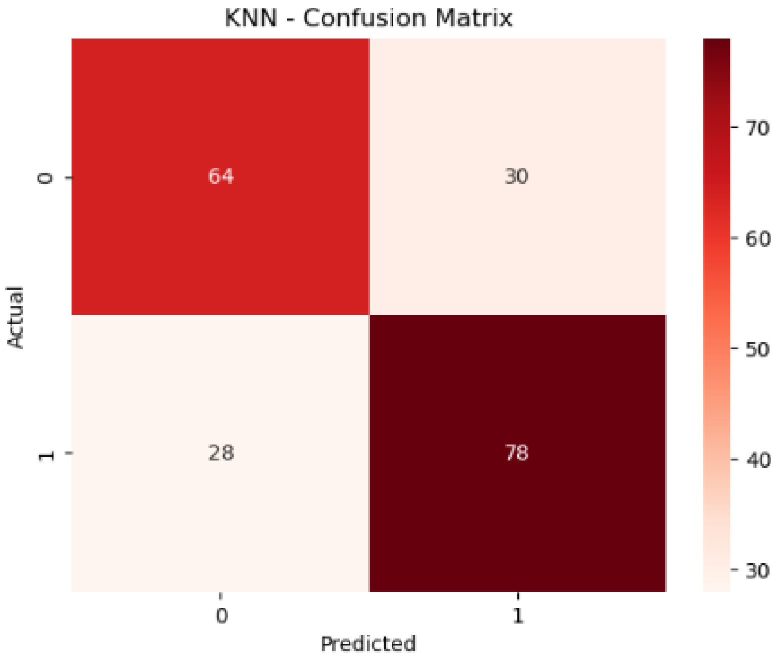
5.K-Nearest Neighbors (KNN)

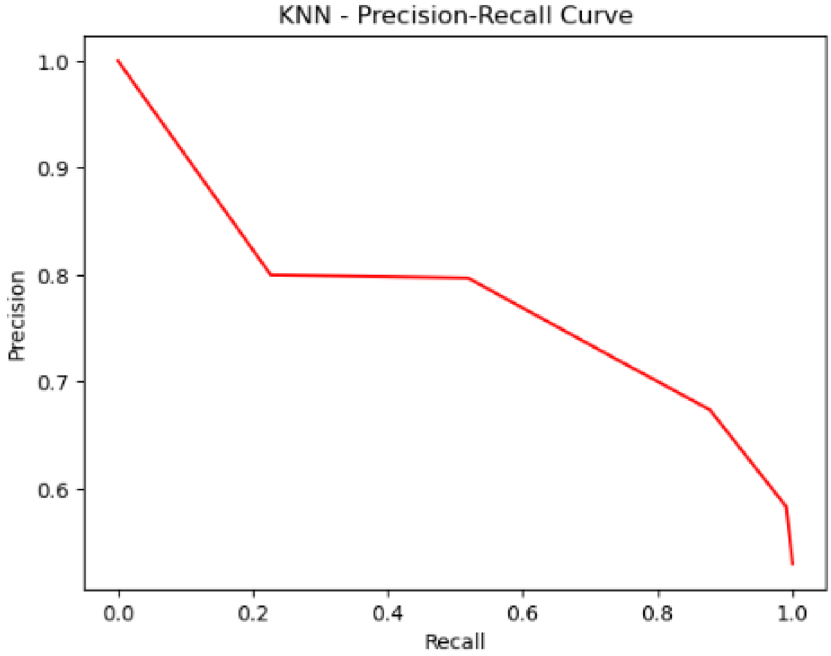
Accuracy: 71%

Precision: 72.22%

Recall: 73.58%

F1 Score: 72.89%





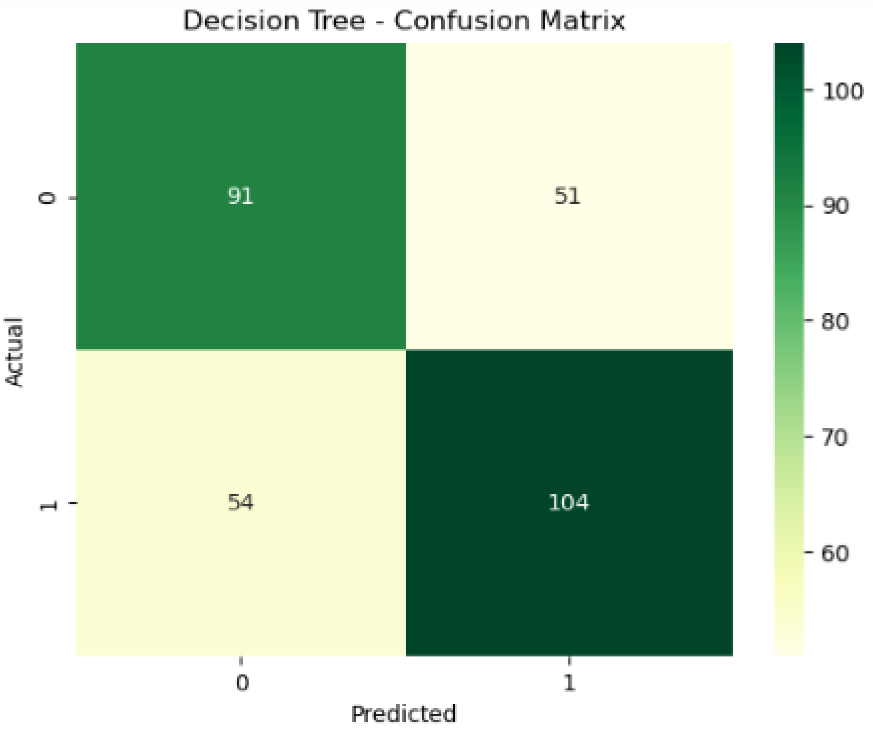
6.Decision Tree

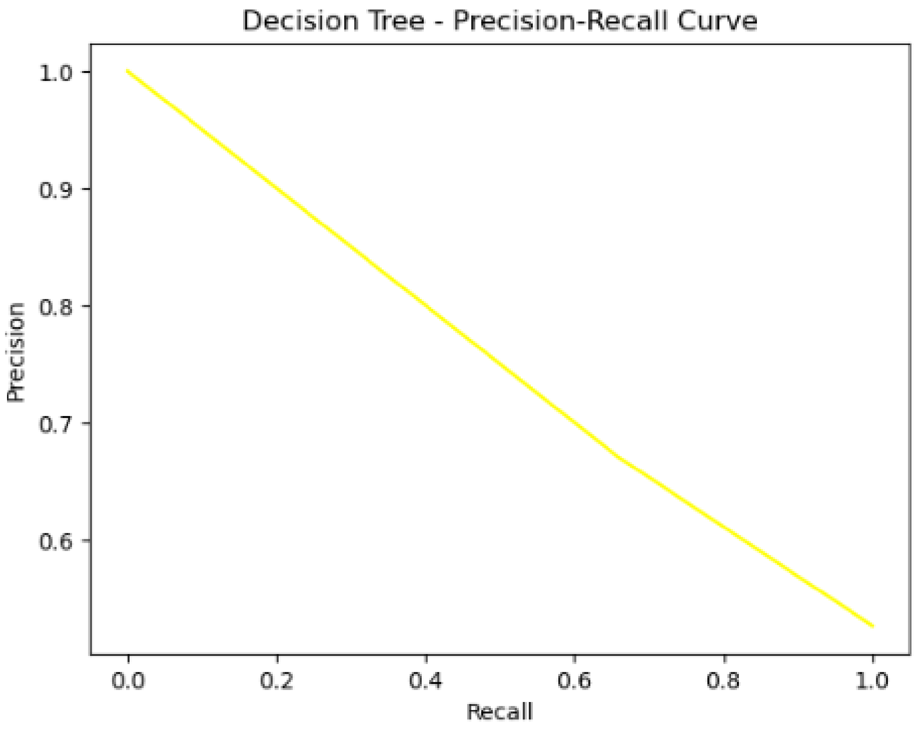
Accuracy: 65%

Precision: 67.09%

Recall: 65.82%

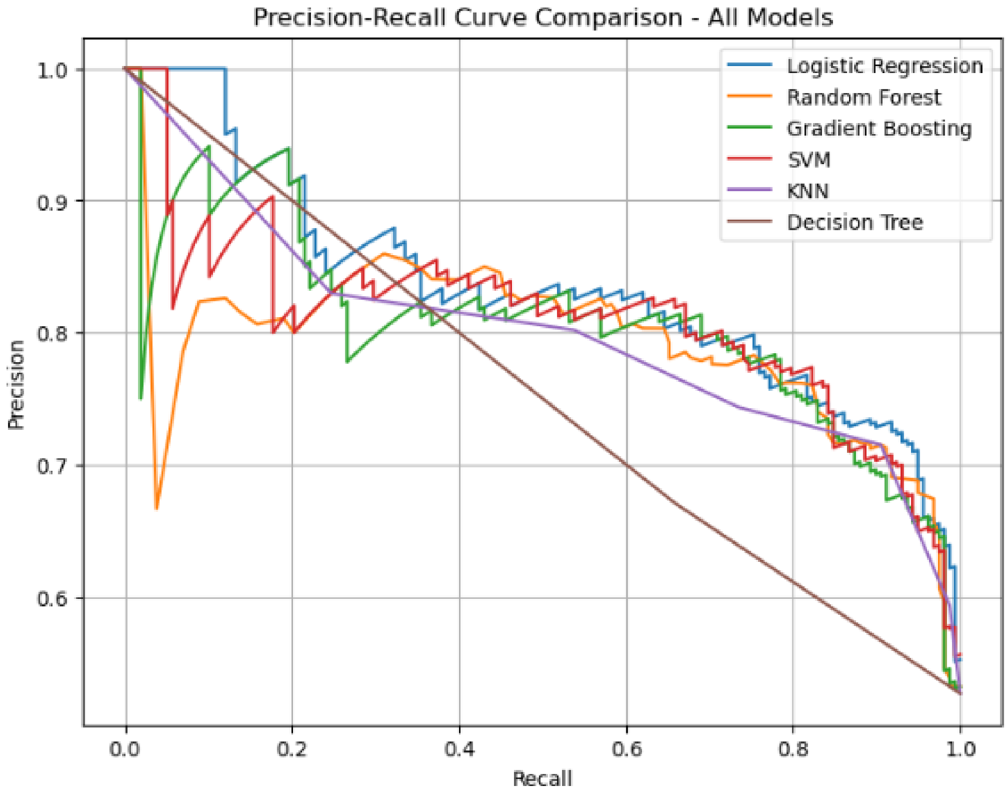
F1 Score: 66.45%

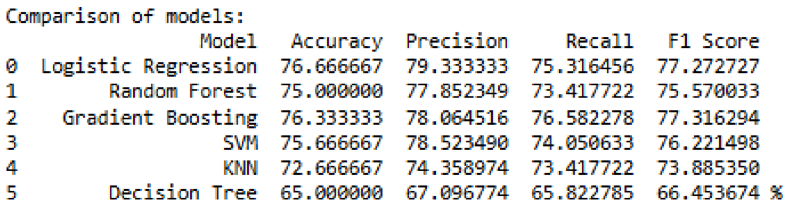




6. MODEL EVALUATION:

Six distinct machine learning models—Logistic Regression, Random Forest, Gradient Boosting, SVM, KNN, and Decision Tree—are evaluated for their precision-recall performance on a classification test. Each model's precision-recall curve is displayed on the graph; the curve nearest the upper-right corner denotes the model with the best overall performance. According to the curves, KNN performs the worst out of all the models displayed, whereas Logistic Regression seems to have the best precision and recall. The best suitable model for the particular needs of the classification problem may be chosen thanks to this kind of visualisation, which enables a thorough assessment of the trade-offs between precision and recall.





```
Comparison of models:
                 Model   Accuracy  Precision     Recall   F1 Score
0  Logistic Regression  76.666667  79.333333  75.316456  77.272727
1        Random Forest  75.000000  77.852349  73.417722  75.570033
2    Gradient Boosting  76.333333  78.064516  76.582278  77.316294
3                  SVM  75.666667  78.523490  74.050633  76.221498
4                  KNN  72.666667  74.358974  73.417722  73.885350
5        Decision Tree  65.000000  67.096774  65.822785  66.453674 %
```

Based on the comparison of models, Logistic Regression appears to be the best-performing model for this particular dataset. It has the highest accuracy (76.67%) and a strong balance between precision (79.33%) and recall (75.32%), leading to a high F1 score of 77.27%. While other models like Gradient Boosting and SVM also show competitive performance with slightly higher recall values, Logistic Regression outperforms them in terms of precision and overall F1 score. Decision Tree, on the other hand, underperforms with the lowest accuracy (65%) and F1 score (66.45%), indicating that it might not generalize as well for this dataset. Therefore, Logistic Regression would be the preferred choice for its overall robust performance across multiple evaluation metrics.

MODELS	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**76.66%**	**79.33%**	**75.31%**	**77.27%**
Random Forest	**75%**	**77.85%**	**73.41%**	**75.57%**
Gradient Boosting	**76.33%**	**78.06%**	**76.58%**	**77.31%**
SVM	**75.66%**	**78.52%**	**74.05%**	**76.22%**
KNN	**71%**	**72.22%**	**73.58%**	**72.89%**
Decision Tree	**65%**	**67.09%**	**65.82%**	**66.45%**

Conclusion:

The model that performed the best in this investigation was logistic regression, which had a 76.67% accuracy rate, a 79.33% precision rate, and an F1 score of 77.27%, demonstrating a good balance across evaluation measures. In terms of precision and overall F1 score, it fared better than other models, making it the most dependable option for this dataset. The Decision Tree model had the worst overall performance due to overfitting, whereas Random Forest and Gradient Boosting showed competitive performance with high recall values. This demonstrates the efficiency of logistic regression in producing reliable and consistent predictions.

References:

1) https://www.mdpi.com/2071-1050/15/9/7087
2) https://www.sciencedirect.com/science/article/pii/S266616592030034X
3) https://www.sciencedirect.com/journal/journal-of-cleaner-production
4) Zhou, K., Yang, S., & Shao, Z. (2022). Energy utilization analysis in smart buildings. Renewable and Sustainable Energy Reviews.
5) Kavousian, A., Rajagopal, R., & Fischer, M. (2017). Determinants of residential electricity consumption. Energy Policy.
6) Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference.
7) Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning.
8) https://www.researchgate.net/publication/355796804_Extreme_temperatures_and_residential_electricity_consumption_Evidence_from_Chinese_households
9) Zhao et al. (2020) is titled "A Data-Driven Energy Consumption Prediction Approach for Buildings Based on Machine Learning Models" https://academic.oup.com/ijlct/article/doi/10.1093/ijlct/ctad127/7612203?login=true