

**Big Data Analytics**

A MINI-PROJECT REPORT

ON

**“Energy Consumption Prediction System”**

BY

# Amey Pawar B – 721

# Anand Prajapati B – 726

# Roshan Prajapati B – 727

# Suraj Prajapati B – 728

Under the guidance of

**Dr. Shikha Gupta**



Juhu-Versova Link Road Versova, Andheri(W), Mumbai-53

Department of Computer Engineering

University of Mumbai

“**OCT-2025**”

I

# Declaration

We wish to state that the work embodied in this project titled **“Energy Consumption Prediction System''** forms our contribution to the work carried out under the guidance of **Dr. Shikha Gupta** at the Rajiv Gandhi Institute of Technology.

This written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Amey Pawar (B - 721)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Anand Prajapati (B - 726)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Roshan Prajapati (B - 727)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Suraj Prajapati (B - 728)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

Managing energy consumption efficiently is one of the biggest challenges faced by power companies, governments, and communities today. Electricity demand changes constantly throughout the day, week, and year, depending on factors like working hours, weekends, holidays, and seasonal patterns. If energy providers cannot predict these changes accurately, it can lead to serious problems such as power shortages, overloading of the grid, or wastage of resources.

The core problem is that energy consumption is not constant—it varies in complex patterns that are difficult to estimate just by looking at historical data. Sudden spikes in demand or unexpected drops can disrupt energy supply and affect both consumers and suppliers. Moreover, planning for future energy needs requires a clear understanding of trends and patterns in consumption, including which times of day or days of the week see the highest or lowest usage.

This project addresses the challenge of understanding and forecasting these energy consumption patterns using historical data. By analysing past trends, it aims to identify regular cycles, peak usage periods, and unusual behaviour in energy consumption. Accurate predictions can help energy planners make informed decisions about how much energy to generate, store, or distribute at different times, reducing costs and avoiding inefficiencies.

The ultimate goal is to provide a way to anticipate energy needs, improve resource allocation, and ensure a reliable and consistent power supply. By understanding the problem clearly, this project focuses on helping energy providers and decision-makers plan smarter and avoid disruptions, ultimately leading to more efficient and sustainable energy management.

Contents

**Contents ..........................................................................................................................**

**List of Figures ..........................................................................................................**

**CHAPTER 1 ....................................................................................................................** Introduction .................................................................................................................

1.1 Introduction .....................................................................................................

**CHAPTER 2 ....................................................................................................................** System Architecture ....................................................................................................

2.1 Proposed Statement .......................................................................................... **CHAPTER 3 .................................................................................................................... Results ............................................................................................................................. CHAPTER 4 ....................................................................................................................**

Conclusion...................................................................................................................

References .................................................................................................................. **Acknowledgement ...........................................................................................................**

# List of Figures

|  |  |
| --- | --- |
| **Figure** | **Page no.** |
| Energy consumption prediction workflow | 6 |
| Energy Consumption Pattern | 7 |
| Average Energy Consumption by Day of Week | 7 |
| Average Energy Consumption by Hour | 8 |
|  |  |

V

# CHAPTER 1 Introduction

**1.1 Introduction**

Energy plays a crucial role in our daily lives and in the functioning of modern society. Power companies need to ensure that electricity is available whenever it is needed, but predicting how much energy will be consumed at different times is a complex task. Energy consumption does not stay the same—it changes throughout the day, across different days of the week, and over the months and seasons of the year. Factors like working hours, weekends, holidays, and seasonal changes can all affect how much electricity people use.

Accurate energy forecasting is important because it helps energy providers plan how much electricity to generate, distribute, or store. Without good predictions, there can be serious problems such as power shortages, overloaded grids, wasted energy, and higher operational costs. Understanding consumption patterns allows planners to make informed decisions, optimize energy distribution, and improve overall efficiency.

This project focuses on **forecasting energy consumption** based on historical data from American Electric Power (AEP), which includes hourly electricity usage over a long period. By analyzing this data, the system can detect patterns such as daily peaks, weekly trends, and unusual consumption events. It also provides insights into how energy usage changes over time, helping decision-makers understand trends and plan better.

The goal of this project is to **help energy providers and planners predict future energy needs**, reduce waste, and ensure that power is delivered reliably. By identifying consumption trends and understanding how usage varies over time, this system can support smarter energy management and contribute to a more sustainable and efficient energy system.

# CHAPTER 2

# System Architecture

**2.1 Proposed Statement**

The proposed project aims to develop an **intelligent energy consumption forecasting system** that helps energy providers, planners, and policymakers predict electricity usage accurately. By analyzing historical energy consumption data, the system can identify patterns, trends, and peak usage periods. This allows for better planning, efficient energy distribution, and reduction of waste or shortages.

The system will be designed to handle **large time-series datasets**, transform them into meaningful features, and deliver predictions through multiple access points. Users will be able to obtain predictions for a single date and time, a range of dates, or analyze consumption patterns over time. The architecture ensures **scalability, reliability, and maintainability**, making it suitable for both research and real-world operational use.

Key goals of the project include:

1. Providing **accurate energy consumption predictions** using historical data.
2. Identifying **patterns and trends** in electricity usage across hours, days, weeks, and seasons.
3. Offering **multiple access options**, including dashboards and APIs, for both interactive and programmatic usage.
4. Ensuring the system is **modular, scalable, and easy to maintain** for future enhancements.

**2.2 Steps Involved**

**Step 1: Data Collection and Preprocessing**

* **Collect Historical Data:** Use the AEP hourly electricity consumption dataset from 2004–2018 (~121,000 records).
* **Data Validation:** Check for missing values, outliers, and inconsistencies.
* **Data Cleaning:** Handle missing data and anomalies to ensure accuracy.
* **Time Indexing:** Convert date and time strings into a structured datetime format for easy analysis.

**Step 2: Feature Engineering**

* **Extract Temporal Features:**
  + Hour of the day (0–23)
  + Day of the week (0–6)
  + Month of the year (1–12)
  + Day of the year (1–365)
  + Quarter (1–4)
  + Week of the year (1–52)
* **Transform Features:** Apply cyclical encoding for hours, days, or months to capture periodic patterns.
* **Create Statistical and Lag Features:** Generate moving averages or lag values to capture trends.
* **Select Features:** Identify the most important features for predictions and optimize the dataset to remove irrelevant variables.

**Step 3: Model Training and Validation**

* **Split Data:** Divide data into training (80%) and validation (20%) sets using time-based splitting.
* **Train Model:** Train the machine learning model to learn consumption patterns from historical data.
* **Validate Model:** Assess performance using appropriate metrics to ensure accurate predictions.
* **Save Model:** Serialize the trained model and feature pipelines for later use.

**Step 4: Service Layer Implementation**

* **Load Model and Features:** Deserialize saved models and initialize prediction services.
* **Provide Prediction Services:**
  + Single datetime prediction
  + Range-based predictions for multiple dates/times
  + Feature importance analysis to understand drivers of consumption
* **Utility Services:**
  + Generate charts and visualizations
  + Perform data validation and error handling

**Step 5: Web Application Integration**

* **Interactive Dashboard (Streamlit):** Allow users to enter input and view predictions, patterns, and charts.
* **Web API (Flask):** Provide programmatic access to predictions, supporting integration with external applications.
* **Handle User Input:** Validate user inputs, extract parameters, and call service layer functions.
* **Display Results:** Present predictions, statistical summaries, and visual insights in an easy-to-understand format.

**Step 6: User Interaction Workflows**

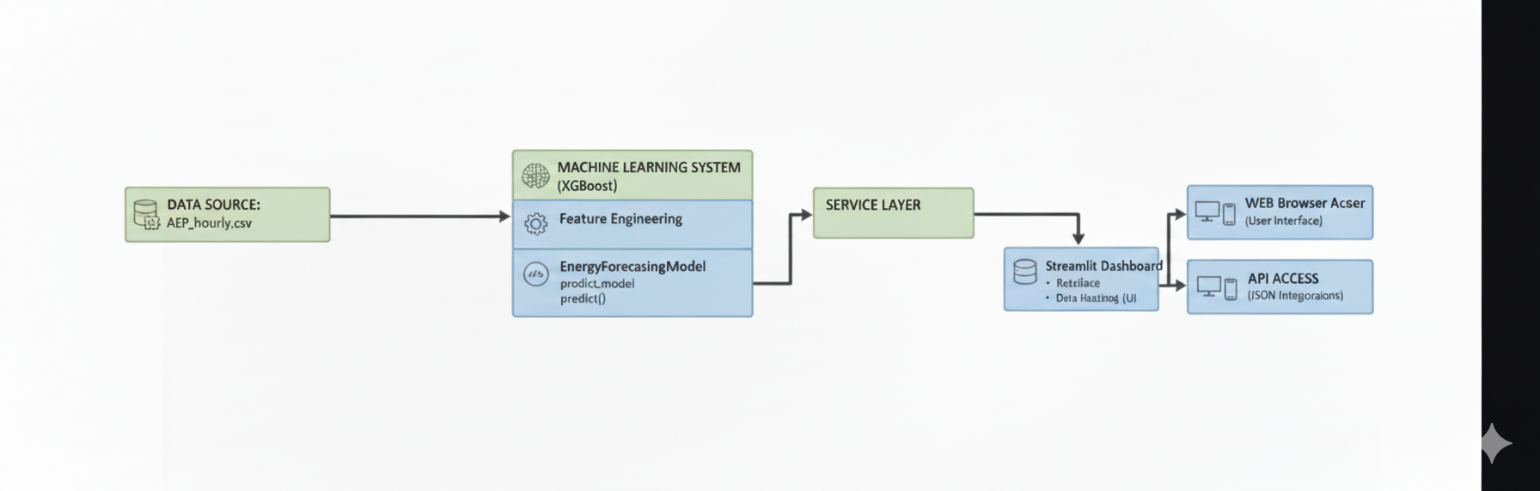
* **Single Prediction Workflow:**
  1. User selects a date and time
  2. Input is validated
  3. Features are generated
  4. Prediction is calculated
  5. Results are displayed with charts
* **Range Prediction Workflow:**
  1. User selects start and end dates with frequency
  2. Batch feature generation occurs
  3. Predictions are computed for all dates in range
  4. Statistical summary and charts are generated
* **Pattern Analysis Workflow:**
  1. User selects analysis parameters
  2. Data is aggregated and analyzed
  3. Consumption patterns and trends are identified
  4. Insights are visualized for decision-making

**Step 7: System Integration**

* **Data Flow Integration:** Ensure smooth flow of data from CSV → Features → Model → Predictions.
* **Service Integration:** Connect model services with web applications for consistent results.
* **API Integration:** Allow external systems to request predictions via REST endpoints.
* **User Interface Integration:** Provide consistent access through both interactive dashboards and programmatic APIs.

**Step 8: Deployment and Quality Assurance**

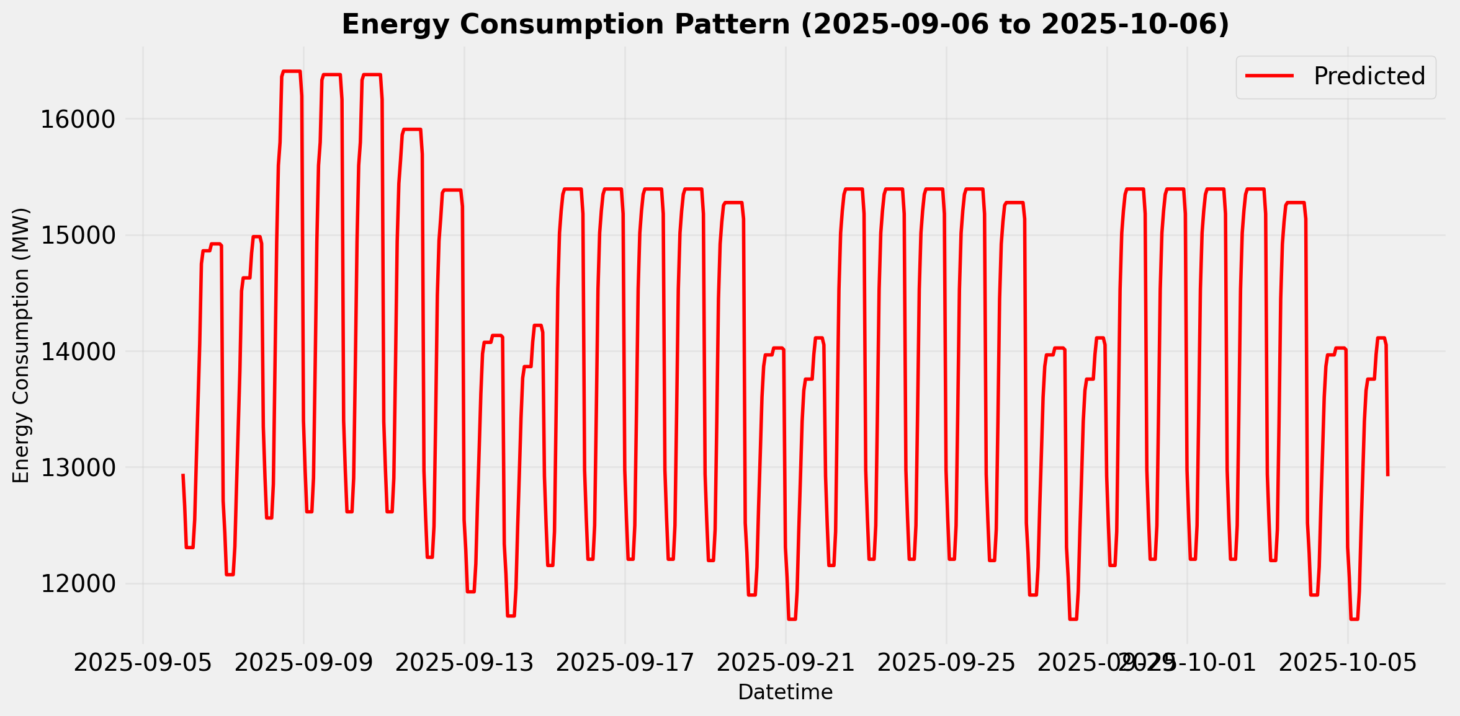
* **Development Environment:** Run the system locally with a virtual environment for testing.
* **Production Environment:** Prepare for containerized deployment with load balancing, monitoring, and logging.
* **Performance and Reliability:** Ensure fast response times, high throughput, error handling, and data validation.
* **Maintainability:** Keep modular design, documentation, and version control to simplify updates and future improvements.



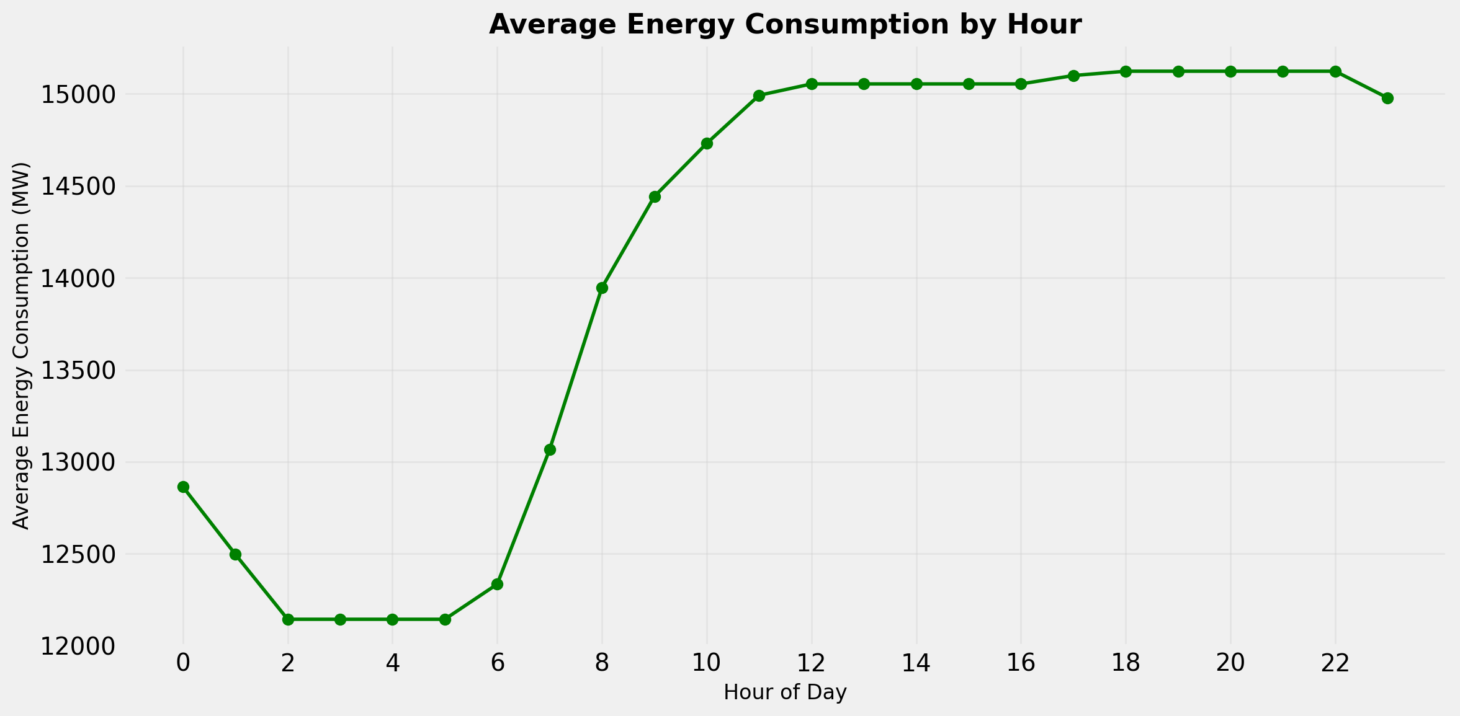
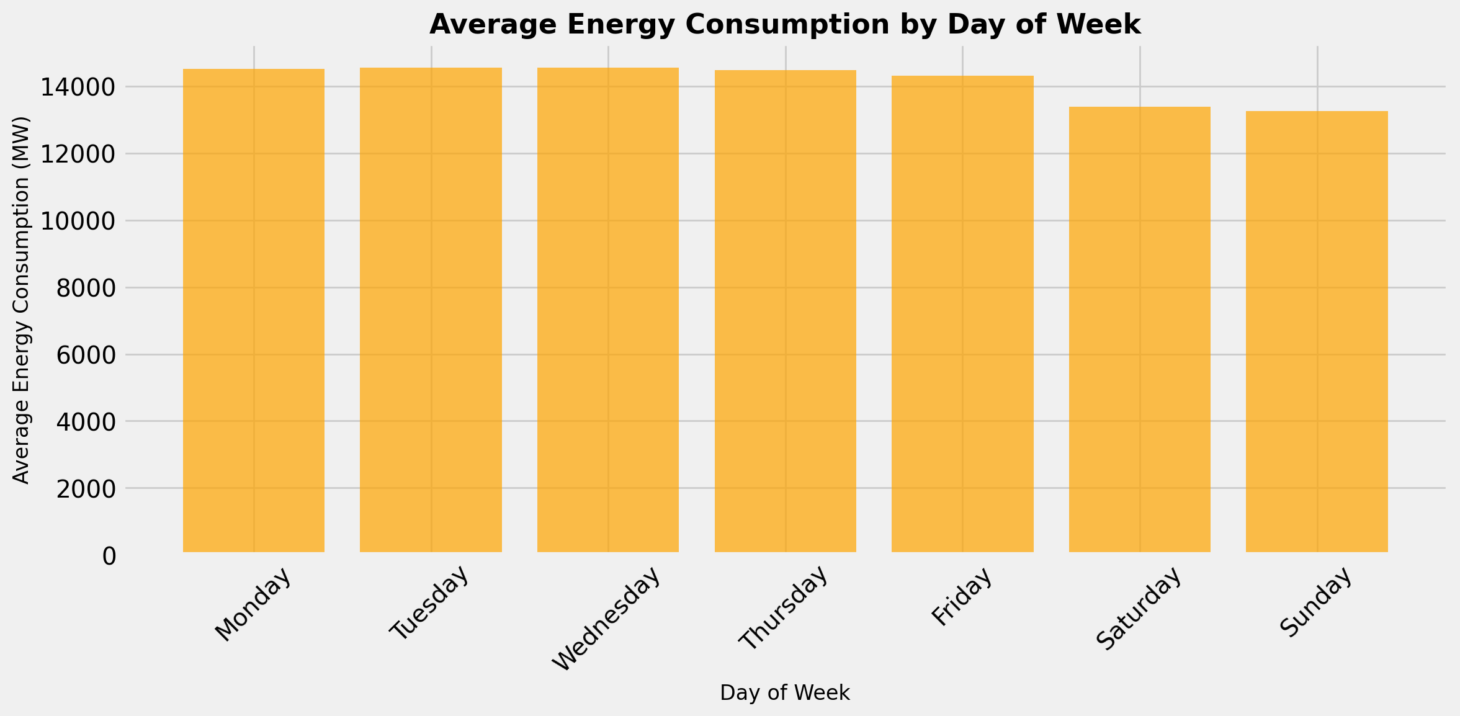
**Figure 1 :** Energy consumption prediction workflow

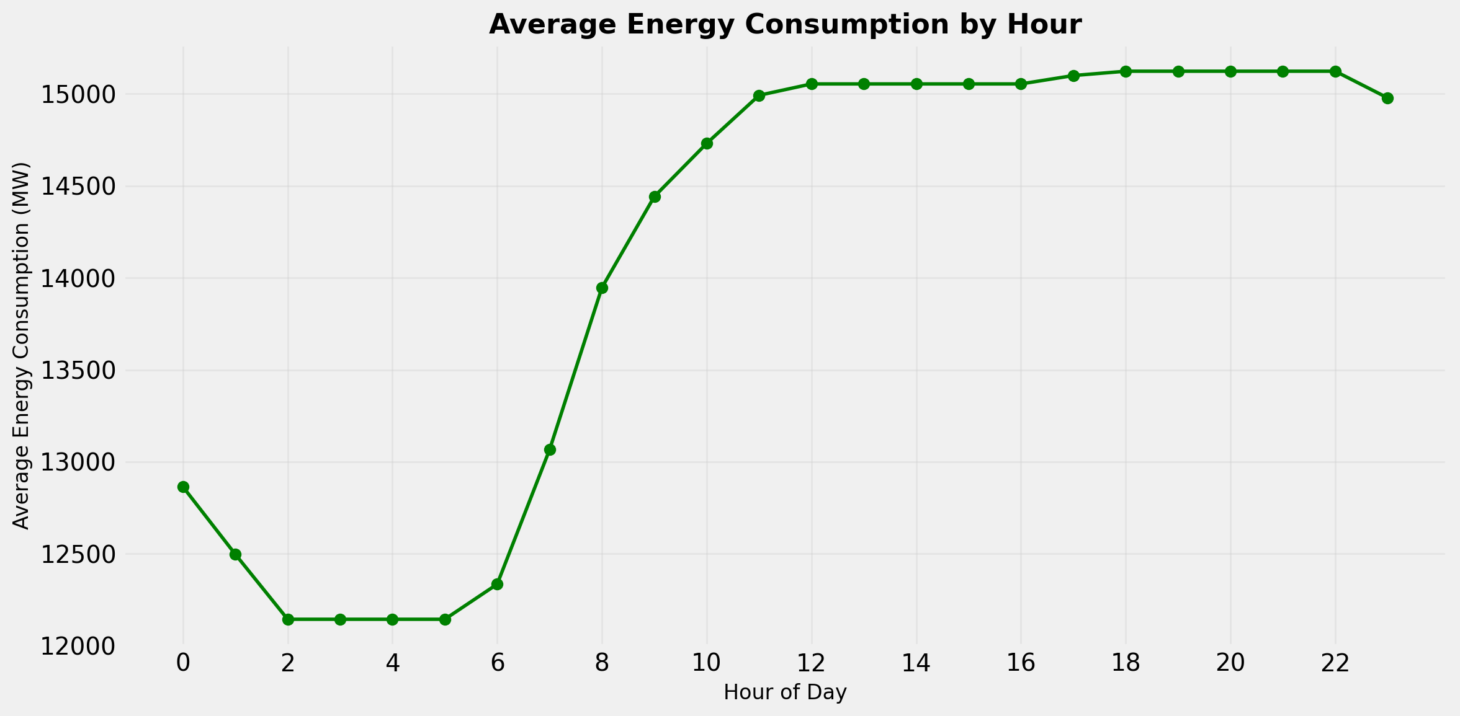
**CHAPTER 3**

**Results**



**Fig 2 : Energy Consumption Pattern**



**Fig 3 : Average Energy Consumption by Day of Week**

**Fig 4 : Average Energy Consumption by Hour**

**CHAPTER 4 Conclusion**

This project successfully delivered an **Energy Consumption Forecasting System** that uses a machine learning model to predict hourly power usage. The core technology is **XGBoost Regression**, which was trained specifically to understand energy patterns based on time factors like the hour of the day. The system is built with a clear, layered structure, meaning the prediction logic is separate from the user interface, making the whole system reliable and easy to maintain.

The system is designed for broad accessibility, offering forecasts through two main channels. First, an interactive **Streamlit Dashboard** allows users to quickly view predictions and spot consumption trends with simple clicks. Second, a **Flask API** provides a set of endpoints, enabling other computer systems to request forecasts programmatically. This dual approach fulfills the goal of providing predictions in both a visual, user-friendly format and a technical, machine-readable format.

Moving forward, the project's next major steps are focused on real-world operational improvements. While the current performance is solid, future work will involve integrating **live data streams** and adding crucial external information like **weather data**. These additions will be key to making the forecasts even more accurate and ensuring the system is robust enough for continuous, real-time decision-making in an operational environment.

**References**

* **G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Communications of the ACM, vol. 18, no. 11, pp. 613–620, Nov. 1975. Relevance: Foundational paper for the Vector Space Model, which enables Vectorization (TF-IDF/Word Embeddings) and the use of Cosine Similarity for content comparison.**
* **G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, Jun. 2005. Relevance: Provides the academic context for the entire architecture, defining and classifying the Content-Based Recommendation System method.**
* **W. McKinney, "Data Structures for Statistical Computing in Python," in Proceedings of the 9th Python in Science Conference, Austin, TX, 2010, pp. 56–61. Relevance: Relates to the practical data management tools (Pandas) used for handling the initial Raw Content Metadata and structuring the Content Vectors and User Profile Vector.**
* **F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011. Relevance: Covers the common software library (scikit-learn) used to implement technical steps like Text Preprocessing, TF-IDF, and Cosine Similarity calculation in the system.**

# Acknowledgement

We wish to express our sincere gratitude to **Dr. Sanjay U. Bokade, Principal**

and **Prof.S. P. Khachane , H.O.D.** of Department Computer Engineering of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on **“Title”**

This project bears the imprint of many peoples. We sincerely thank our project guide **Dr. Shikha Gupta** for his guidance and encouragement in carrying out this synopsis work.

Finally, we would like to thank our colleagues and friends who helped us in completing project work successfully

1. Amey Pawar
2. Anand Prajapati
3. Rosha Prajapati
4. Suraj Prajapati