"Dissemination of Education for Knowledge, Science and Culture"

-Shikshanmaharshi Dr. Bapuji Salunkhe

# VIVEKANAND COLLEGE, KOLHAPUR

(Empowered Autonomous)

**DEPARTMENT OF STATISTICS**

A Project Report on

# "ANALYSIS OF HOUSEHOLDS ELECTRICITY CONSUMPTION IN KOLHAPUR CITY"

Submitted by –

**Mr. Pankaj Parasharam Jadhav**

**Mr. Darshan Sharad Kanire**

**Miss. Samiksha Mahadev Thorat**

in partial fulfilment for the award of the degree of

MASTER OF SCIENCE

in

STATISTICS

DEPARTMENT OF STATISTICS

VIVEKANAND COLLEGE

KOLHAPUR

**2024-25**

# CERTIFICATE

This is to Certify that,

| Sr. No | Name | Roll No. |
|---|---|---|
| 1 | Pankaj Parasharam Jadhav | 1403 |
| 2 | Darshan Sharad Kanire | 1404 |
| 3 | Samiksha Mahadev Thorat | 1413 |

Have satisfactorily completed the research project work on “**ANALYSIS OF HOUSEHOLDS ELECTRICITY CONSUMPTION IN KOLHAPUR CITY**” as a part of practical evaluation course for **M.Sc. II**, prescribed by the Department of Statistics, **Vivekanand College, Kolhapur (Empowered Autonomous)** in the academic year **2024-25.**

This project has been completed under our guidance and supervision. To the best of our knowledge and belief, the matter presented in this project report is original and has not been submitted elsewhere for any other purpose.

Date: 08/05/2025

Place: Kolhapur

| **Project Guide** | **Examiner** | **Head** |
|---|---|---|
| (Mr. A. A. Pawar) | | (Mrs. V. C. Shinde) |

# ACKNOWLEDGEMENT

It is a genuine pleasure to express our special thanks to Department of Statistics who gave us the golden opportunity to do this wonderful Research Project on the topic **"The Analysis of Household electricity Consumption in Kolhapur City.”**, which helped us in doing a lot of research and we came to know about so many new things.

We would like to express our profound gratitude and deep regards to our Guide Mr. A. A. Pawar for his exemplary guidance, monitoring and constant encouragement.

We are privileged to express our sincere thanks with great respect and gratitude to Mr. R. R. Kumbhar, Principal of College, Mrs. V. C. Shinde, Head of Department of Statistics, Mr. A. A. Pawar Project Guide and all staff members for their aspiring guidance. They all helped with kind of co-operation and constant encouragement. We are grateful to thank them for providing us with all necessary facilities.

Finally, last but by no means least; we would thank our parents for their guidance and support on every step, also to everyone in the Department of Statistics, it was great sharing premises with all of you during last two years. Thanks for all your encouragement.

Yours Sincerely,
M.Sc. II
Department of Statistics

# INDEX

# INTRODUCTION

Electricity consumption in households plays a crucial role in determining the overall energy demand and sustainability of a region or country. Understanding how and why households use electricity is essential for designing policies, promoting energy efficiency, and reducing environmental impacts. Electricity consumption patterns are influenced by a variety of factors, including household size and the awareness of energy-saving practices.

In most households, electricity is used to power essential appliances such as lighting, refrigerators, heating and cooling systems, and electronics. The type and frequency of appliance usage significantly contribute to the amount of electricity consumed. For example, heating and air conditioning can represent a large portion of a household's energy use, especially in regions with extreme temperatures. Additionally, modern technological devices such as computers, televisions, and kitchen appliances also contribute to rising electricity consumption in many homes.

Changes in lifestyle, including the increased use of electronic devices and the shift toward home-based activities, have led to an increase in household electricity consumption in recent decades. However, awareness of the environmental impact and cost associated with excessive energy use has led to a growing interest in energy efficiency measures. Many households are adopting practices such as using energy-efficient lighting, upgrading to energy-saving appliances, and exploring renewable energy options like solar power.

Understanding electricity consumption patterns is vital for both consumers and policymakers. governments and utility companies, analysing consumption data helps in improving grid management, implementing energy-saving programs, and achieving sustainability goals. In this context, surveys and studies focusing on household electricity use are invaluable for gathering data, identifying trends, and creating targeted initiatives to reduce energy waste and promote efficiency.

In this survey, we aim to explore the electricity consumption behaviours of households, identify factors affecting usage patterns. By doing so, we hope to provide insights that can contribute to the development of more efficient energy policies and practices at the household level.

# SCOPE OF THE STUDY

The scope of this study defines the boundaries within which the research has been conducted, ensuring clarity in objectives, data collection, analysis, and interpretation. This study concentrates on understanding and analysing household electricity consumption patterns in Kolhapur city with the aim of identifying key influencing factors, establishing relationships among variables, and providing forecasts for future energy demands. The scope is delineated under the following dimensions:

The analysis will be based on monthly electricity consumption data from a defined historical period (e.g., of past 54 months, depending on data availability). This timeframe allows for a comprehensive examination of consumption trends, seasonal variations, and patterns that influence forecasting accuracy for future energy demand.

The study includes variables such as: Household Income, Household Size, Number of Rooms and Electrical Equipment. These variables are crucial for establishing correlations and understanding the structural contributors to energy consumption.

In addition to quantitative metrics, the study also evaluates:

- The level of awareness among households regarding energy-saving technologies and sustainable practices.
- Attitudes and behaviours related to electricity usage.
- Household satisfaction with the quality, reliability, and affordability of electricity services provided by the utility company.

These aspects provide insight into the human and behavioural dimensions of energy consumption, which are essential for designing effective energy policies and awareness campaigns.

The analysis is intended to support data-driven recommendations for improving residential energy efficiency and planning for sustainable energy management in Kolhapur.

# OBJECTIVES

- Analysis of Monthly Household Electricity Consumption of Kolhapur city and Forecasting for Predicting Future Energy Use.
- Analysing the Relationship Between Household Income and Electricity Consumption.
- Comparative Analysis of the Impact of Household Size, Number of Rooms, and Equipment on Electricity Consumption.
- Evaluate household awareness of energy-saving practices and technologies and Analysing electricity service satisfaction.

# DATA COLLECTION & METHODOLOGY

For the statistical research project on the Topic '**Analysis of households Electricity consumption in Kolhapur City'**, the data is based on both **Primary** and **Secondary** Sources.

The Primary data is collected with the help of Survey by **Questionnaire method** which have questions related to topic.

The Primary data consist sample of size **390** from Kolhapur city. This data collected by dividing the population, that is Kolhapur city, into different sections.

This sample size is determined by Cochron's sample size formula. Then the sample is collected in proportion to total consumers of electricity for each section of population.

| KOLHAPUR CITY | Count | Proportion | Sample size |
|---|---|---|---|
| **Central** | 39160 | 0.2322272 | 90 |
| **East** | 25373 | 0.1504673 | 60 |
| **North** | 32146 | 0.1906326 | 75 |
| **West** | 45756 | 0.2713428 | 105 |
| **Market Yard** | 26193 | 0.1553301 | 60 |
| **Total** | **168628** | | **390** |

The secondary data is collected from 'Mahavitaran, Maharashtra State Electricity Board' (MSEB) Office, Kolhapur. The Secondary data consist monthly electricity consumption of Kolhapur city from April 2020 to September 2024.

# SOFTWARES AND TOOLS

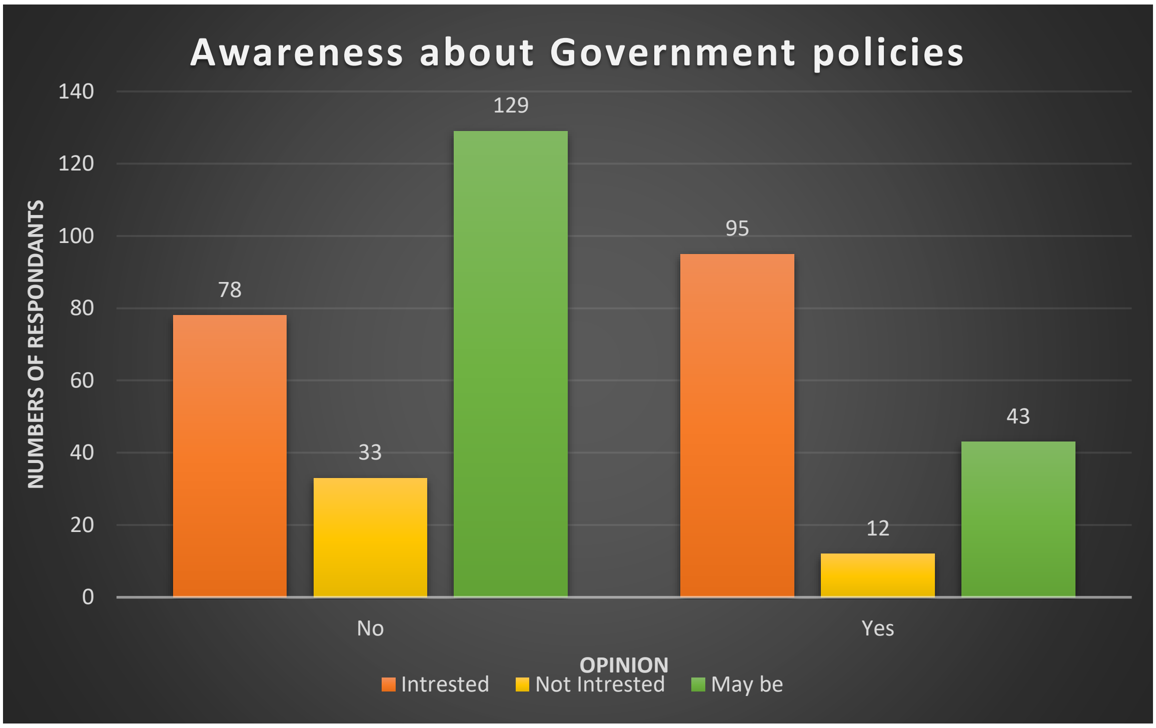
**Statistical methods used:**

- Bar Chart
- Pie Chart
- Time series forecasting
- Chi-square test for Association
- Correlation Heatmap
- Sentiment Analysis (By TextBlob and Model fitting by Logistic regression model)

Data analysis is conducted using software and tools like Excel, Jupyter Notebook, Minitab Statistical Software, and ChatGPT.

# GRAPHICAL REPRESENTATION

- **To analyse impact of awareness on public interests in Government policies: -**

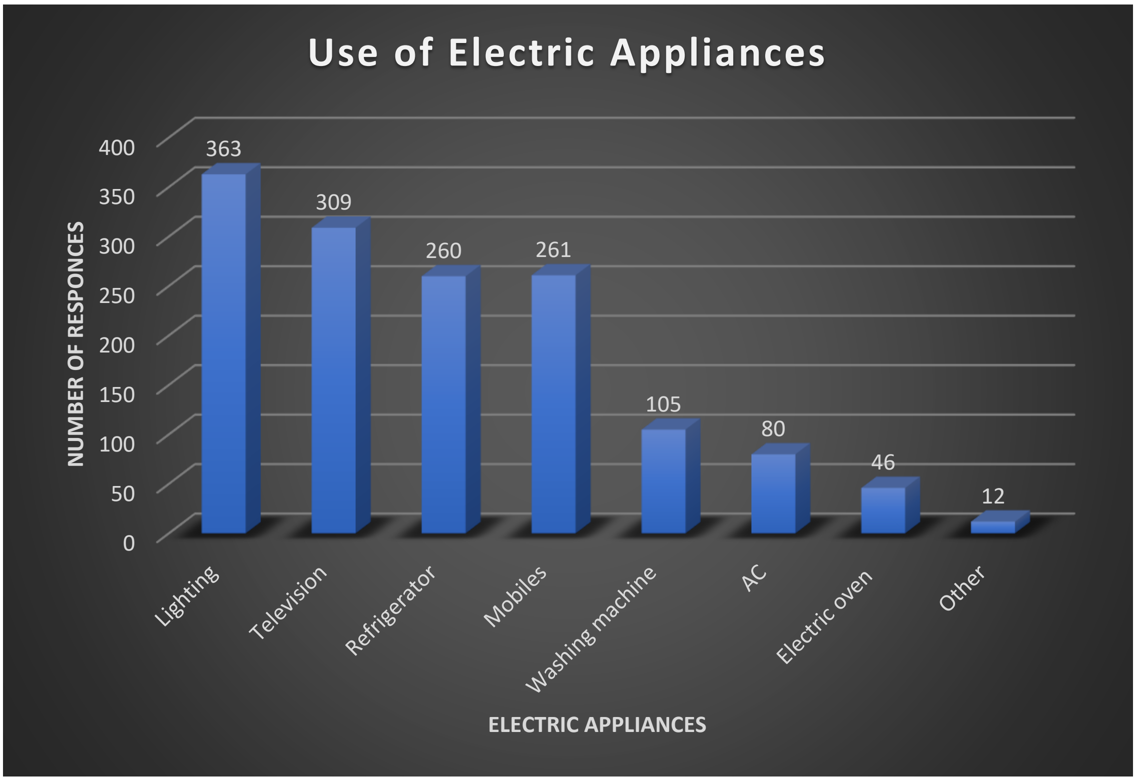



**Conclusion: -**

From above graph, we conclude that out of 390 respondents 150 are aware about government policies and among them 95 are interested in implement government schemes.

We conclude that out of 390 respondents 240 are not aware about government policies and among them 129 may be interested in implement government schemes.

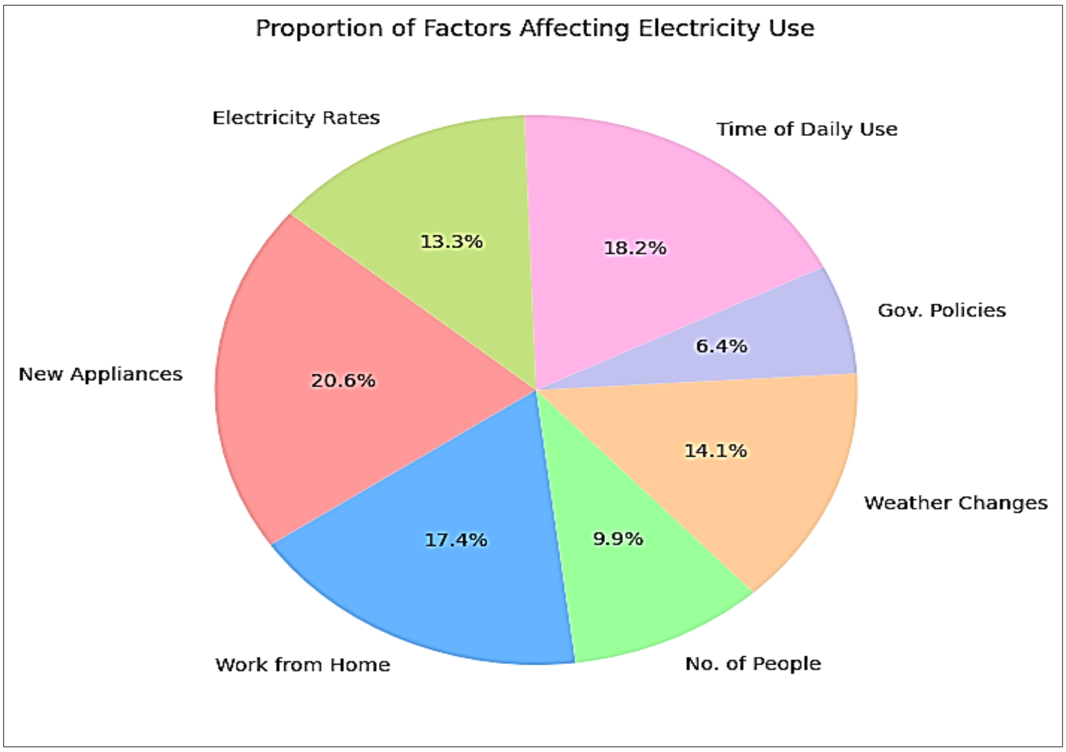
- **To analyse which appliances are used regularly by households.**




**Conclusion:-**

- Lighting and televisions are the highly used electric appliances.
- Refrigerators and mobile devices used significantly.
- Washing machines, ACs, and ovens least used electric appliances.

➢ **The factors affecting household electricity consumption in future: -**




**Conclusion: -**

- New Appliance and Time of Daily use are the biggest factors affecting electricity consumption.
- Government policies have the least direct impact, indicating that awareness and incentives might need to be improved.

# STATISTICAL ANALYSIS

➢ **Analysis of Monthly Average Household Electricity Consumption and Forecasting for Predicting Future Energy Use: -**

**Forecasting Using SARIMA: -**

As electricity demand is influenced by both seasonal variations and long-term usage trends, time series forecasting methods provide a powerful approach to understanding and predicting consumption patterns. Among various models, the Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model is particularly well-suited for data that exhibits both trend and seasonality. SARIMA extends the traditional ARIMA model by incorporating seasonal terms, allowing for better handling of repeated cycles—such as monthly or yearly fluctuations in electricity usage.

This study employs the SARIMA model to forecast monthly household electricity consumption in Kolhapur, evaluate model accuracy, and provide insights for future energy demand management.

**Forecasts are obtained using SARIMA –**

**Best model parameters –**

SARIMA (p, d, q) x (P, D, Q) s

SARIMA (1, 2, 1) x (1, 1, 1) 12

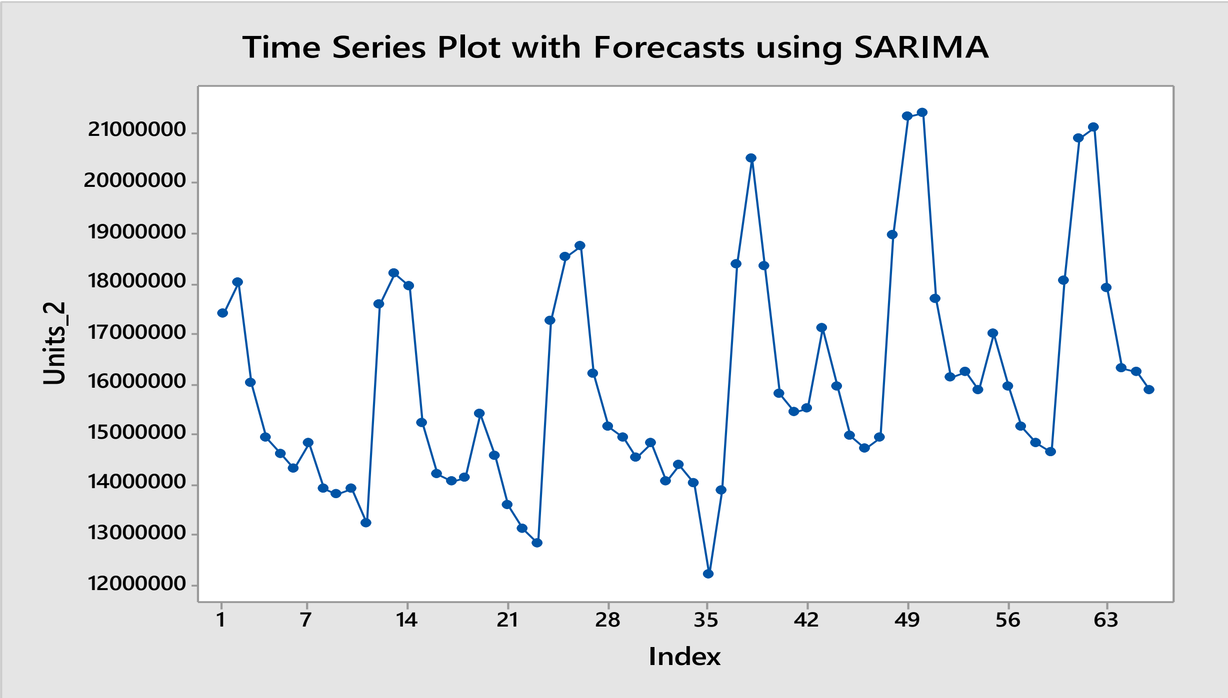
Here,

**Non-seasonal parameters**: AR(1), differencing=2, MA(1)

**Seasonal parameters**: Seasonal AR(1), Seasonal Differencing=1, Seasonal MA(1), seasonal period of 12 (i.e. monthly data)

The model accounts for both trend and seasonal components, making it suitable for data with repeating patterns, such as monthly electricity usage.

**Time Series Plot with Forecasts using SARIMA: -**




## Forecasts are obtained using SARIMA –

## Accuracy Measures: -

MAPE (Mean Absolute Percentage Error): 3.37%
MSE (Mean Squared Error): 492,787,363,807.69
RMSE (Root Mean Squared Error): 701,988.15
MSD (Mean Squared Deviation): 492,787,363,807.69 (same as MSE)

These results confirm that the SARIMA model is highly effective and reliable for short-term electricity demand forecasting in Kolhapur households.

**Electricity consumption (Forecasts values) for next 12 months-**

| Period | Forecast | Actual |
|---|---|---|
| Oct-23 | 17199166 | 17085120 |
| Nov-23 | 16246471 | 15941244 |
| Dec-23 | 15639250 | 14952338 |
| Jan-24 | 15296673 | 14701841 |
| Feb-24 | 14626769 | 14923299 |
| Mar-24 | 18541701 | 18938560 |
| Apr-24 | 20035558 | 21328208 |
| May-24 | 20109700 | 21383172 |
| Jun-24 | 17496281 | 17700309 |
| Jul-24 | 16197508 | 16106760 |
| Aug-24 | 15855629 | 16213620 |
| Sep-24 | 15506327 | 15866359 |
| **Oct-24** | **16991470** | |
| **Nov-24** | **15934260** | |
| **Dec-24** | **15141430** | |
| **Jan-25** | **14829790** | |
| **Feb-25** | **14643320** | |
| **Mar-25** | **18041110** | |
| **Apr-25** | **20878430** | |
| **May-25** | **21097410** | |
| **Jun-25** | **17888190** | |
| **Jul-25** | **16290710** | |
| **Aug-25** | **16217280** | |
| **Sep-25** | **15883370** | |

**Conclusions:-**

For the forecasts, these seem to be the best values for the parameters **(p, d, q)** and **(P, D, Q,s)** used in a **SARIMA model**, i.e. SARIMA (1, 2, 1) × (1, 1, 1) 12.

These parameters suggests that model fit well for given data and give forecast values with better accuracy measures.

- **Analysing the Relationship Between Household Income and Electricity Consumption :-**

In this context, it is important to analyse whether a statistical association exists between household income and electricity consumption levels. To investigate this, the Chi-Square Test for Association is used, which helps determine if variations in electricity usage are significantly related to differences in household income groups.

The Relationship Between Household Income and Electricity Consumption is analyzed using **Chi-Square Test for Association**.

**Hypothesis to test-**
**Ho** : There is no significant association between Income and Bill.
**H1** : There is significant association between Income and Bill.

Pearson Chi-Square = 312.939
DF = 16
P-Value = 0.000

## Conclusions:-

Chi-Square Test for Association shows that there is a **strong association** between **Income** and **Bill**. Since the P-value is very small (less than the typical significance level of 0.05), we can confidently reject the null hypothesis and conclude that **Income and Bill are significantly associated.**

- **Comparative Analysis of the Impact of Household Size, Number of Rooms and Equipment on Electricity Consumption.**

**<u>Correlation Heatmap</u>: -**

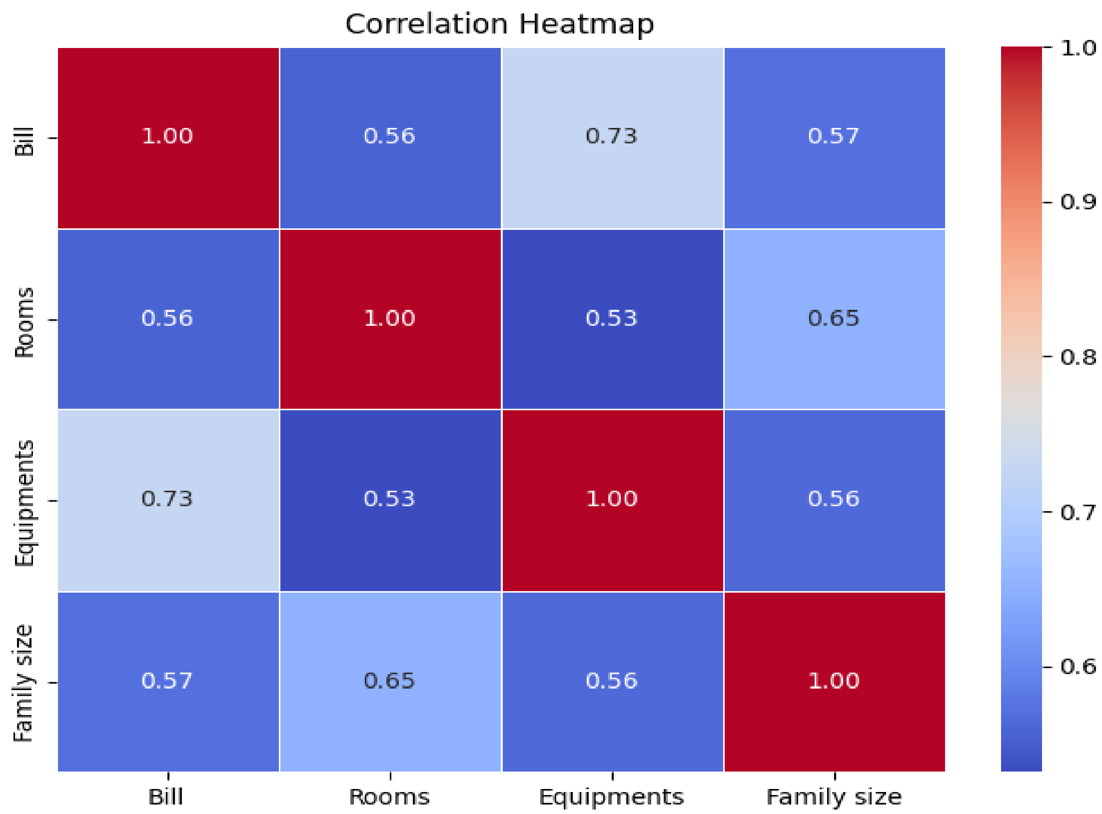
A correlation heatmap is a graphical tool used to show the strength and direction of relationships between numerical variables. Each cell in the heatmap shows the correlation coefficient (typically Pearson's), which ranges from -1 to 1:

+1 indicates a perfect positive correlation

-1 indicates a perfect negative correlation

0 indicates no correlation

The colour gradient (often from blue to red or white to dark) helps in quickly identifying patterns. This visualization is useful for identifying multicollinearity, selecting relevant features for modelling, or simply understanding variable relationships in exploratory data analysis.



This correlation heatmap provides insights into the relationships between four variables: **Bill, Rooms, Equipment,** and **Family size**.

**Conclusions:-**

The analysis reveals a **strong positive correlation (0.73)** between the number of electrical equipment and the electricity bill, indicating that **households with more appliances tend to incur higher energy costs.**

The heatmap suggests that electrical equipment has the strongest influence on the electricity bill, followed by family size and number of rooms. These insights can guide energy-saving strategies targeting equipment usage and household size-related consumption.

---

➢ For Analysing electricity service satisfaction.

## Sentiment Analysis and Tools

Sentiment analysis, also known as opinion mining, is a natural language processing (NLP) technique used to determine the emotional tone behind a body of text. It is widely applied in fields such as market research, social media monitoring, and customer feedback analysis to categorize sentiments as **positive**, **neutral**, or **negative**. Various tools and approaches are used for sentiment analysis, ranging from simple rule-based methods to advanced machine learning models.

**Tools: -**

**TextBlob**:

It is a lightweight Python library that uses a lexicon-based approach. It assigns polarity scores to words and calculates the overall sentiment by averaging these scores.

**VADER (Valence Aware Dictionary and sentiment Reasoner)**:

VADER is another rule-based sentiment analysis tool. It uses a dictionary of lexical features and considers factors like capitalization, punctuation, and degree modifiers to adjust sentiment intensity, making it more sensitive to the way people express emotions online.

**BERT (Bidirectional Encoder Representations from Transformers):**

BERT is a deep learning-based model developed by Google. Unlike lexicon-based methods, BERT understands the contextual meaning of words using a transformer architecture. In this analysis, we used a pre-trained BERT model to classify sentiment

Output using **TextBlob**:

| Sentiments | Count |
|---|---|
| Positive | 179 |
| Neutral | 166 |
| Negative | 45 |

Evaluation for –

Using **TextBlob**_Sentiment: **Accuracy: 0.53**

Using **VADER**_Sentiment: Accuracy: 0.52

Using **BERT**_Sentiment: Accuracy: 0.45

Here, TextBlob performed slightly better than VADER and significantly better than BERT in this specific context.

## Conclusion:

The sentiment analysis results provide insight into household perspectives on energy-saving practices and electricity service satisfaction.

Using the **TextBlob** sentiment analysis tool, the majority of responses were identified as **positive (179)**, followed by **neutral (166)** and a smaller number of **negative (45)** responses. Among the three methods evaluated, TextBlob achieved the highest accuracy at **0.53**

### ➢ Modelling in Sentiment Analysis

Modelling in sentiment analysis refers to the process of building computational systems that can automatically classify the sentiment of a given piece of text (e.g., positive, negative, neutral). It involves selecting appropriate techniques and algorithms to analyse and interpret subjective information from textual data.

**Machine Learning-Based Models –**

These models treat sentiment analysis as a text classification task. They require **labelled datasets** for training.

- **Common Algorithms**:
  - Random Forests
  - Support Vector Machines (SVM)
  - XGBoost
  - LightRoom

## Accuracy after using above models:

| Model | Accuracy |
|---|---|
| Logistic Regression | 0.68 |
| Random Forest | 0.66 |
| SVM | 0.66 |
| XGBoost | 0.61 |
| LightRoom | 0.62 |

**Hyperparameter Tuning Using GridSearchCV for given models:**

**Hyperparameter tuning** is the process of optimizing the configuration settings of a machine learning model to improve its performance. Unlike model parameters (learned from data), hyperparameters are set before training and can significantly affect a model's accuracy, overfitting, and generalization. **GridSearchCV** is a method from Scikit-learn that performs an exhaustive search over a specified hyperparameter grid.

Accuracy after Hyperparameter Tuning Using GridSearchCV for given models:

| Model | Accuracy |
|---|---|
| Logistic Regression after Hyperparameter Tuning | 0.687 |
| Random Forest after Hyperparameter Tuning | 0.6608 |
| SVM after Hyperparameter Tuning | 0.6695 |

**Logistic Regression** achieved the highest accuracy of **0.687**, indicating it performed best on the dataset after tuning.

**Improving Logistic Regression Model further:**

Accuracy after applying SMOTE and Retrain Logistic Regression= **0.69**

## Conclusion:

**Logistic Regression** emerged as the most effective model for this sentiment analysis task with accuracy of 0.69, it is due to its strong performance and interpretability, can be recommended for similar text-based sentiment classification tasks in the energy services domain.

# CONCLUSIONS

- ✓ To forecast monthly household electricity consumption, the SARIMA (Seasonal Auto-Regressive Integrated Moving Average) model was applied, since the monthly electricity consumption data exhibit both trend and seasonality.

  The best-fitting model was identified as SARIMA (1, 2, 1) × (1, 1, 1) 12

  Future electricity consumption (Forecast values) for next 12 months was obtained by using this model. The Forecast values are obtained with better Accuracy Measures.

- ✓ To Analyse the relationship between Household Income and Electricity Consumption, using Chi-Square test for association, we find that there is a strong association between Income and Bill. That is Household Income and Electricity Bill are significantly associated.

- ✓ For the comparative analysis of the impact of Household Size, Number of Rooms and Equipment on Electricity Consumption, the correlation heatmap gives a strong positive correlation (0.73) between the number of electrical equipment and the electricity bill. This means that households with more appliances tend to have higher energy bill, followed by family size and number of rooms.

- ✓ The sentiment analysis results provide insight into household perspectives on energy-saving practices and electricity service satisfaction.

  Using the TextBlob sentiment analysis tool, the majority of responses were identified as positive (179)**,** followed by neutral (166) and a smaller number of negative (45) responses. Among the three methods evaluated, TextBlob achieved the highest accuracy at 0.53

  Logistic Regression emerged as the most effective model for this sentiment analysis task with accuracy of 0.69, it is due to its strong performance and interpretability, can be recommended for similar text-based sentiment classification tasks in the energy services domain.

# REFERENCES

Md Rokonuzzaman, Sharmin Jahan, Md Shahidul Haque. Household Electricity Consumption of Middle-Class Family in Chittagong - A Case Study. American Journal of Energy Research, 2016, Vol. 4, No. 2, 35-41

Liga Poznaka*, Ilze Laicane, Dagnija Blumberga, Andra Blumberga, Marika Rosa. Analysis of electricity user behaviour: case study based on results from extended household survey. International Scientific Conference "Environmental and Climate Technologies – CONECT 2014".

Prakash S. Chougule, Suresh V. Patil, Sonam A. Amate, Anjali S. Bhosale. A comparative study of consumption of electricity using statistical measures. EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal Volume: 7 | Issue: 2 | February 2021

A.B. Pawar, M.A. Jawale and D.N. Kyatanavar. Fundamentals of Sentiment Analysis: Concepts and Methodology. Sentiment Analysis and Ontology Engineering, Studies in Computational Intelligence 639

Evaristus Didik Madyatmadja1,*, Felix1, I. Gusti Ketut Edrick1, Joseph William Indarto1,David Jumpa Malem Sembiring. Harmonizing sentiments: Analyzing user reviews of Spotify through sentiment analysis. Journal of Infrastructure, Policy and Development 2024, 8(9), 7101.

William W. S. Wei. Time Series Analysis – Univariate and Multivarite methods. Department of Statistics, The Fox School of Bussiness and Management, Temple University.

# APPENDIX

## QUESTIONARRIE

1. Area of residence -

(राहण्याचे ठिकाण)

- Urban (शहरी)

2. If Urban, select ward -

जर शहरी असाल तर वॉर्ड निवडा -

- A
- B
- C
- D
- E

2. Number of family members -

(कुटुंबातील सदस्यांची संख्या)

___________________(Short-answer text)

3. Type of residence -

(निवासाचा प्रकार)

- Apartment/Flat (अपार्टमेंट/फ्लॅट)
- House / Bungalow (घर / बंगला)
- Semi-detached House (अर्ध-पृथक घर)
- Other… ___________________

4. No. of Rooms in house -

(घरातील खोल्यांची संख्या)

___________________(Short-answer text)

5. No. of Electronic devices in house -

(घरातील इलेक्ट्रॉनिक उपकरणांची संख्या)

___________________(Short-answer text)

6. Monthly Income of family –

(कुटुंबाचे मासिक उत्पन्न )

- Below 25000 (25000 पेक्षा कमी)
- 25000 – 50000
- 50000 – 75000
- 75000 - 1 lakh
- Above 1 lakh (1 लाखांपेक्षा जास्त)

8. What is your average monthly electricity bill?

(तुमचे सरासरी मासिक वीज बिल किती आहे?)

- Below 500 (500 पेक्षा कमी)
- 500-1000
- 1000-1500
- 1500 – 2000
- Above 2000 (2000 पेक्षा जास्त)

9. What is your use of average monthly electricity unit?

(सरासरी मासिक वीज युनिटचा तुमचा वापर काय आहे?)

- 0 – 100
- 100 – 200
- 200 – 300
- 300 – 400
- More than 400 (400 पेक्षा जास्त)

10. Highly electricity consumption month -

(तुमच्या मते, कोणत्या महिन्यात विजेचा वापर जास्त होतो?)

- ○ January (जानेवारी)
- ○ February (फेब्रुवारी)
- ○ March (मार्च)
- ○ April (एप्रिल)
- ○ May (मे)
- ○ June (जून)
- ○ July (जुलै)
- ○ August (ऑगस्ट)
- ○ September (सप्टेंबर)
- ○ October (ऑक्टोबर)
- ○ November (नोव्हेंबर)
- ○ December (डिसेंबर)

11. Do you track your electricity usage?

(तुम्ही तुमच्या वीज वापराचा मागोवा घेता का?)

- ○ Yes (होय)
- ○ No (नाही)

12. Which of the following devices do you use regularly in your household? (Check all that apply)

(खालीलपैकी कोणते उपकरण तुम्ही तुमच्या घरात नियमितपणे वापरता? (लागू होणारे सर्व निवडा))

- □ Lighting (e.g. Bulbs)
- □ Television
- □ Refrigerator
- □ Computer / Laptop / Mobiles
- □ Washing machine
- □ Air conditioner / Heater
- □ Electric Oven / Stove
- □ Other…_____________

13. Do you use any energy-efficient appliances (e.g., LED bulbs, energy-efficient refrigerators, etc.)?

(तुम्ही ऊर्जा-कार्यक्षम उपकरणे (उदा. एलईडी बल्ब, ऊर्जा-कार्यक्षम रेफ्रिजरेटर इ.) वापरता का?)

- ○ Yes (होय)
- ○ No (नाही)

14. Which of the following energy-saving measures do you regularly practice? (Check all that apply)

(खालीलपैकी कोणते ऊर्जा-बचत उपाय तुम्ही नियमितपणे करता? (लागू होणारे सर्व निवडा))

- □ Turning off lights when not in use (वापरात नसताना दिवे बंद करणे)
- □ Using energy-efficient light bulbs (e.g., LED) (ऊर्जा-कार्यक्षम दिवे वापरणे)
- □ Installing solar panels or energy-efficient appliances (सौर पॅनेल किंवा ऊर्जा-कार्यक्षम उपकरणे स्थापित करणे)
- □ Setting the thermostat at energy-efficient temperatures (ऊर्जा-कार्यक्षम तापमानावर थर्मोस्टॅट सेट करणे)
- □ Other…_______________

15. Have you made any upgrades to your home to reduce energy consumption?

(ऊर्जेचा वापर कमी करण्यासाठी तुम्ही तुमच्या घरामध्ये काही सुधारणा केल्या आहेत का?)

- ○ Yes (होय)
- ○ No (नाही)

16. How do you anticipate your household electricity consumption will change in the next 6 months?

(पुढील ६ महिन्यांत तुमच्या घरातील विजेचा वापर कसा बदलेल असा तुमचा अंदाज आहे?)

- ○ Increase (वाढेल)
- ○ Decrease (कमी होईल)
- ○ Remain the same (तसाच राहील)

17. What factors do you think will impact your household's future electricity consumption? (Check all that apply)

(तुमच्या घरातील भविष्यातील विजेच्या वापरावर कोणते घटक परिणाम करतील असे तुम्हाला वाटते?) (लागू होणारे सर्व निवडा)

- □ New appliances (नवीन उपकरणे)
- □ Change in lifestyle (in. Work from home) (जीवनशैलीत बदल)
- □ Changes in number of people in house (घरातील लोकांच्या संख्येत बदल)
- □ Weather changes (हवामान बदल)
- □ Government policies (सरकारी धोरणे)
- □ Other…_______________

18. Are you aware of any programs or government policies aimed at reducing electricity consumption in your area?

(तुमच्या क्षेत्रातील विजेचा वापर कमी करण्याच्या उद्देशाने तुम्हाला कोणतेही कार्यक्रम किंवा सरकारी धोरणे माहीत आहेत का?)

- ○ Yes (होय)
- ○ No (नाही)

19. Do you use any renewable energy sources (e.g., solar power, wind energy) to generate electricity in your household?

(तुमच्या घरामध्ये वीज निर्माण करण्यासाठी तुम्ही कोणतेही अक्षय ऊर्जा स्रोत (उदा. सौर ऊर्जा, पवन ऊर्जा) वापरता का?)

- ○ Yes (होय)
- ○ No (नाही)

20. Would you be willing to invest in renewable energy solutions (e.g., solar panels) if offered government incentives?

(सरकारी सवलती दिल्यास तुम्ही अक्षय ऊर्जा उपायांमध्ये (उदा. सौर पॅनेल) गुंतवणूक करण्यास तयार आहात का?)

- ○ Yes (होय)
- ○ No (नाही)
- ○ Maybe

21. In your opinion, what are the main factors contributing to your electricity consumption? (Check all that apply)

(तुमच्या मते, तुमच्या विजेच्या वापरामध्ये मुख्य घटक कोणते कारणीभूत आहेत?) (लागू होणारे सर्व निवडा)

- □ Number of members in family (कुटुंबातील सदस्यांची संख्या)
- □ Use of appliances (उपकरणांचा वापर)
- □ Time of daily use (रोजच्या वापराची वेळ)
- □ Electricity rates (विजेचे दर)
- □ Heating and cooling needs (हीटिंग आणि कूलिंगच्या गरजा)
- □ Other…_________________

22. Overall satisfaction on electricity service -

(वीज सेवेबद्दल मत) -

- ○ Satisfaction Level
- ○ More Satisfied
- ○ Satisfied
- ○ Neutral
- ○ Less satisfied
- ○ Unsatisfied

23. Opinion on satisfaction of overall electricity service (Needs of improvements)

(Review on electricity bill & power interruptions)-

एकूण वीज सेवेच्या समाधानाबद्दल तुमचे मत (सुधारणेची गरज)

(वीज बिल आणि वीज व्यत्ययांचे पुनरावलोकन) -

________________________(Long-answer text)

## Objective – 1

### ADF test for stationarity: -

```
import pandas as pd

from statsmodels.tsa.stattools import adfuller

# Ensure no missing values

consumption_data = df['Units'].dropna()

# Check if the data has variation

if consumption_data.nunique() > 1:

    # Perform the ADF test

    result = adfuller(consumption_data)

    # Print the results

    print('ADF Statistic: %f' % result[0])

    print('p-value: %f' % result[1])

    print('Critical Values:')

    for key, value in result[4].items():

        print('\t%s: %.3f' % (key, value))

    # Interpretation

    if result[1] <= 0.05:

        print("The time series is likely stationary.")

    else:

        print("The time series is likely non-stationary.")

else:

    print("The data has very little variation or is constant, causing ADF to fail.")
```

### Differencing: -

```
# First Differencing

import matplotlib.pyplot as plt

# Calculate the differenced data

df['Differenced_Units'] = df['Units'].diff()

# Plot the differenced data

plt.figure(figsize=(12, 6))

plt.plot(df.index, df['Differenced_Units'], marker='o', linestyle='-', color='b', label='Differenced Electricity Consumption')

plt.xlabel("Year")

plt.ylabel("Differenced Units Consumed")

plt.title("Differenced Electricity Consumption Over Time")

plt.legend()

plt.grid(True)

plt.show()


# Second Differencing

from statsmodels.tsa.stattools import adfuller

# Import the adfuller function

# Assuming 'consumption_data' is your original time series data

# Second differencing

# Convert 'Units' column to numeric, handling errors by coercing to NaN

# Use 'ignore' instead of 'coerce' for errors argument in older pandas version

# OR Update your pandas library for 'coerce' functionality

second_differenced_data = df['Units'].astype(float, errors='ignore').diff().diff().dropna()  # Second difference
```

```
# Perform ADF test on the second differenced data

result_second_diff = adfuller(second_differenced_data)

print('\nADF Statistic (Second Differenced): %f' % result_second_diff[0])

print('p-value (Second Differenced): %f' % result_second_diff[1])

if result_second_diff[1] <= 0.05:

    print("The second differenced time series is likely stationary.")

    # Proceed with modeling using the second differenced data

else:

    print("The second differenced time series is likely non-stationary. Consider other transformations.")


# prompt: draw the graph for second differencing data

import matplotlib.pyplot as plt

# Plot the second differenced data

plt.figure(figsize=(12, 6))

plt.plot(second_differenced_data.index, second_differenced_data, marker='o', linestyle='-', color='b', label='Second Differenced Electricity Consumption')

plt.xlabel("Year")

plt.ylabel("Second Differenced Units Consumed")

plt.title("Second Differenced Electricity Consumption Over Time")

plt.legend()

plt.grid(True)

plt.show()

import matplotlib.pyplot as plt

from statsmodels.graphics.tsaplots import plot_acf, plot_pacf

# Assuming 'second_differenced_data' is available from the previous code


# Plot ACF and PACF

fig, axes = plt.subplots(1, 2, figsize=(16, 4))

# ACF plot

plot_acf(second_differenced_data, lags=15, ax=axes[0])  # Adjust lags as needed

axes[0].set_title('Autocorrelation Function (ACF)')

# PACF plot

plot_pacf(second_differenced_data, lags=15, ax=axes[1])  # Adjust lags as needed

axes[1].set_title('Partial Autocorrelation Function (PACF)')

plt.tight_layout()

plt.show()


# prompt: fit model Sarima

import matplotlib.pyplot as plt

from statsmodels.tsa.statespace.sarimax import SARIMAX

import pandas as pd # Import pandas for data manipulation


# Load the data from the first sheet if it's not already loaded

try:

    df

except NameError:

    df = pd.read_csv('/content/Electricity consumption.csv')
```

```python
# Convert 'Month' to datetime if not already done
if not pd.api.types.is_datetime64_any_dtype(df['Month']):
    df['Month'] = pd.to_datetime(df['Month'], format='%b-%y')
    df.set_index('Month', inplace=True)


# Assuming 'consumption_data' should be the 'Units' column from your DataFrame
consumption_data = df['Units'] # Define consumption_data here


# Example SARIMA model (replace with your determined values)
model = SARIMAX(consumption_data, order=(1, 2, 1), seasonal_order=(1, 1, 1, 12))  # (p, d, q), (P, D, Q, s)
# Fit the model
results = model.fit()
# Print model summary
print(results.summary())
# Forecast future values
forecast_steps = 12  # Number of steps to forecast
forecast = results.get_forecast(steps=forecast_steps)
forecast
# Get the predicted mean and confidence intervals
predicted_mean = forecast.predicted_mean
confidence_intervals = forecast.conf_int()
# Plot the forecast
plt.figure(figsize=(12, 6))
plt.plot(consumption_data.index, consumption_data, label='Observed')
plt.plot(predicted_mean.index, predicted_mean, label='Forecast')
plt.fill_between(confidence_intervals.index,
                 confidence_intervals.iloc[:, 0],
                 confidence_intervals.iloc[:, 1],
                 color='gray', alpha=0.3, label='Confidence Interval')
plt.xlabel('Year')
plt.ylabel('Electricity Consumption')
plt.title('SARIMA Forecast')
plt.legend()
plt.show()


# prompt: best order best AIC for SARIMA model
import itertools
# Define the p, d, q, P, D, Q, s ranges for your SARIMA model
p = d = q = range(0, 3)  # Try orders up to 2 for simplicity
P = D = Q = range(0, 2)
s = 12  # Seasonal period (e.g., 12 for monthly data)
# Generate all possible combinations of orders
orders = list(itertools.product(p, d, q))
seasonal_orders = list(itertools.product(P, D, Q, [s]))


# Initialize variables to store the best AIC and corresponding model order
best_aic = float('inf')
best_order = None
best_seasonal_order = None
```

```
# Iterate through all possible combinations of orders and fit the SARIMA model
for order in orders:
 for seasonal_order in seasonal_orders:
  try:
   model = SARIMAX(consumption_data, order=order, seasonal_order=seasonal_order)
   results = model.fit()
   if results.aic < best_aic:
    best_aic = results.aic
    best_order = order
    best_seasonal_order = seasonal_order
  except Exception as e:
   # Handle errors (e.g., model not converging) and continue
   print(f"Error fitting model with order {order} and seasonal order {seasonal_order}: {e}")
   continue
# Print the best order and AIC
print(f"Best SARIMA order: {best_order}")
print(f"Best SARIMA seasonal order: {best_seasonal_order}")
print(f"Best AIC: {best_aic}")
# You can now use the best_order and best_seasonal_order to fit your final SARIMA model
import numpy as np


def calculate_metrics(forecast, actual):
  mse = np.mean((forecast - actual) ** 2)
  mape = np.mean(np.abs((actual - forecast) / actual)) * 100
  msd = np.mean(forecast - actual)
  return mse, mape, msd

forecast = np.array([18149522, 16847399, 16548408, 16362725, 17199166, 16246471,15639250, 15296673, 14626769, 18541701, 20035558, 20109700,
17496281, 16197508, 15855629, 15506327])


actual = np.array([18331405, 15809704, 15423735, 15507022, 17085120, 15941244,14952338, 14701841, 14923299, 18938560, 21328208, 21383172,
17700309, 16106760, 16213620, 15866359])


mse, mape, msd = calculate_metrics(forecast, actual)
print(f"Mean Squared Error (MSE): {mse}")
print(f"Mean Absolute Percentage Error (MAPE): {mape}%")
print(f"Mean Signed Deviation (MSD): {msd}")
```

## Objective – 3

```
import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt


# Load your dataset (replace 'your_data.csv' with actual file)

df = pd.read_csv("O1 Heatmap.csv")


# Print the column names to check for discrepancies

print(df.columns)


# Select relevant numerical variables, correcting any mismatches

# Replace with the actual column names from your DataFrame, as printed above

# Make sure these column names exactly match the output of print(df.columns)

selected_columns = ['Bill','Rooms','Equipments','Family size']

# The columns were renamed to match the actual column names


corr_matrix = df[selected_columns].corr()


# Plot the heatmap

plt.figure(figsize=(8,6))

sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)

plt.title("Correlation Heatmap")

plt.show()
```

## Objective – 4

**# Use of Randaom Forest on Balanced Sentiment Data**

```
import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import LabelEncoder

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, classification_report, confusion_matrix


file_path = "/content/Balanced_Sentiment_Data (1).xlsx"

df = pd.read_excel(file_path)

# Separate features and target

X = df.drop(columns=['Rating'])  # Features (BoW)

y = df['Rating']        # Target (sentiment labels)


# Encode target labels (e.g., Positive → 2, Neutral → 1, Negative → 0)

label_encoder = LabelEncoder()

y_encoded = label_encoder.fit_transform(y)

# Split data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(

    X, y_encoded, test_size=0.2, random_state=42

)
```

```
# Train a Random Forest classifier

model = RandomForestClassifier(n_estimators=100, random_state=42)

model.fit(X_train, y_train)

# Predict on the test set

y_pred = model.predict(X_test)

# Evaluate model

accuracy = accuracy_score(y_test, y_pred)

report = classification_report(y_test, y_pred, target_names=label_encoder.classes_)

conf_matrix = confusion_matrix(y_test, y_pred)


# Output results

print(f"Accuracy: {accuracy:.2f}\n")

print("Classification Report:")

print(report)

print("Confusion Matrix:")

print(conf_matrix)
```

# Use of SVM and XGBoost on Balanced Sentiment Data

```
pip install xgboost scikit-learn

from sklearn.svm import SVC

from xgboost import XGBClassifier

from sklearn.metrics import accuracy_score

# Train SVM

svm_model = SVC(kernel='linear', random_state=42)

svm_model.fit(X_train, y_train)

y_pred_svm = svm_model.predict(X_test)

accuracy_svm = accuracy_score(y_test, y_pred_svm)

print(f"SVM Accuracy: {accuracy_svm:.2f}")

# Train XGBoost

xgb_model = XGBClassifier(use_label_encoder=False, eval_metric='mlogloss', random_state=42)

xgb_model.fit(X_train, y_train)

y_pred_xgb = xgb_model.predict(X_test)

accuracy_xgb = accuracy_score(y_test, y_pred_xgb)

print(f"XGBoost Accuracy: {accuracy_xgb:.2f}")
```

**# Use of Hyperparameter tuning for Random Forest with GridSearchCV (For Improving Accuracy)**

```
from sklearn.model_selection import GridSearchCV

from sklearn.ensemble import RandomForestClassifier

# Define the parameter grid

param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2]
}

# Initialize Random Forest

rf = RandomForestClassifier(random_state=42)

# Grid search with 5-fold cross-validation

grid_search = GridSearchCV(
    estimator=rf,
    param_grid=param_grid,
    cv=5,
```

```
    scoring='accuracy',
    n_jobs=-1,
    verbose=1
)

# Fit to training data
grid_search.fit(X_train, y_train)
# Best estimator and accuracy
print("Best Parameters:", grid_search.best_params_)
print("Best CV Accuracy:", grid_search.best_score_)
# Evaluate on test set
best_rf = grid_search.best_estimator_
y_pred_best_rf = best_rf.predict(X_test)
accuracy = accuracy_score(y_test, y_pred_best_rf)
print("Test Set Accuracy with Best RF:", accuracy)
```

---

→

```
# Logistic Regression Tuning with GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import accuracy_score, classification_report
# Define the model
lr = LogisticRegression(max_iter=1000)
# Grid of hyperparameters to search
param_grid = {
    'C': [0.01, 0.1, 1, 10],
    'penalty': ['l2'],  # 'l1' only works with 'liblinear' solver
    'solver': ['liblinear', 'saga']
}
# Run grid search
grid = GridSearchCV(lr, param_grid, cv=5, scoring='accuracy', verbose=1, n_jobs=-1)
grid.fit(X_train, y_train)
# Best model
best_lr = grid.best_estimator_
# Evaluate on test set
y_pred_lr = best_lr.predict(X_test)
print("🔍 Best Parameters:", grid.best_params_)
print("✅ Test Set Accuracy:", accuracy_score(y_test, y_pred_lr))
print("\n📋 Classification Report:")
print(classification_report(y_test, y_pred_lr))
```

```
# Ensemble Logistic + SVM + RF with VotingClassifier.
from sklearn.ensemble import VotingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score
# Instantiate individual models
log_clf = LogisticRegression(C=10, penalty='l2', solver='liblinear', max_iter=1000)
svm_clf = SVC(C=10, kernel='rbf', gamma='scale', probability=True)  # probability=True needed for soft voting
```

```
rf_clf =
RandomForestClassifier(n_estimators=100,
max_depth=20, min_samples_split=5,
min_samples_leaf=1, random_state=42)


# Combine into VotingClassifier
voting_clf = VotingClassifier(
    estimators=[
        ('lr', log_clf),
        ('svm', svm_clf),
        ('rf', rf_clf)
    ],
    voting='soft'  # use 'hard' if you don't want
probability averaging
)
# Fit ensemble model
voting_clf.fit(X_train, y_train)
# Predict and evaluate
y_pred_ensemble = voting_clf.predict(X_test)
print("✅ Ensemble Accuracy:",
accuracy_score(y_test, y_pred_ensemble))
print("\n📋 Classification Report:")
print(classification_report(y_test,
y_pred_ensemble))
# This is just below your best solo model
(Logistic Regression @ 68.7%) — but it's
more balanced across classes.
# Best Overall Model: Tuned Logistic
Regression with Accuracy=68.7%
# Improve Logistic Regression model Further:
# To Apply SMOTE and Retrain Logistic
Regression
pip install imbalanced-learn
from imblearn.over_sampling import SMOTE
from sklearn.linear_model import
LogisticRegression
from sklearn.metrics import
classification_report, accuracy_score


# Apply SMOTE on training set
smote = SMOTE(random_state=42)
X_train_sm, y_train_sm =
smote.fit_resample(X_train, y_train)
# Check class balance (optional)
from collections import Counter
print("Resampled class distribution:",
Counter(y_train_sm))
log_clf_smote = LogisticRegression(C=10,
penalty='l2', solver='liblinear', max_iter=1000,
random_state=42)
log_clf_smote.fit(X_train_sm, y_train_sm)
# Predict on test set
y_pred_smote = log_clf_smote.predict(X_test)
# Evaluate
print("✅ Accuracy (SMOTE):",
accuracy_score(y_test, y_pred_smote))
print("\n📋 Classification Report (SMOTE):")
print(classification_report(y_test,
y_pred_smote))


# SMOTE did its job!
# Matching the best Logistic Regression score,
but with a more balanced treatment of Neutral
and Positive classes.
# Same overall accuracy, No performance
drop,
# And now the model has learned with
balanced exposure to all classes.
# finalize & save this model
# Finalize and save SMOTE-balanced Logistic
Regression model so it’s ready for
deployment, sharing, or integration into an
app!
```