1.Abstract

India's rapid economic growth, fueled by population expansion, urbanization, and industrialization, has led to a sharp rise in energy demand. As the third-largest global energy consumer, India remains heavily reliant on fossil fuels, despite increasing contributions from renewable sources. Effective energy management is critical to balancing energy security, economic growth, and environmental sustainability. However, energy consumption patterns vary significantly across regions and sectors, including residential, industrial, and commercial domains. Additionally, peak demand periods often strain the power grid, causing supply shortages and blackouts, while the integration of renewable energy sources such as solar and wind adds further complexity.

This study applies advanced machine-learning techniques, including Support Vector Machines (SVM), Random Forest, and ARIMA models, to analyze time series data and forecast energy consumption in India. Historical data from various states and sectors are examined to identify key consumption drivers and predict future demand. Each algorithm offers unique advantages: SVM excels in identifying complex patterns, Random Forest provides robustness with high accuracy, and ARIMA captures time-based trends and seasonality. Random Forest outperformed other models with a Mean Squared Error (MSE) of 846.65, Mean Absolute Error (MAE) of 15.57, and an R^2 value of 0.80.

The findings of this research highlight the importance of predictive modeling in optimizing energy management. Policymakers and energy providers can leverage these insights to manage demand, reduce disruptions, and integrate renewable energy more effectively. This study underscores the potential of machine-learning techniques to address India's energy challenges, contributing to a sustainable and secure energy future.

2.Introduction

India faces a growing challenge in managing its rising energy demand driven by industrialization, urbanization, and population growth. As the third-largest global energy consumer, the country relies heavily on fossil fuels while expanding its renewable energy capacity. Effective energy management is critical to ensuring economic growth, energy security, and environmental sustainability. Distinct consumption patterns across residential, industrial, and commercial sectors, combined with peak demand periods, strain the power grid and add complexity with intermittent renewable sources. Addressing these issues requires understanding historical consumption patterns and developing accurate forecasting models.

Support Vector Machines (SVM) analyze historical energy data to predict future demand by identifying linear and non-linear patterns through kernel functions. SVM excels in handling complex relationships but requires careful tuning and is computationally intensive for large datasets, making it suitable for uncovering sectoral energy dependencies.

Random Forest, an ensemble learning method, leverages multiple decision trees to enhance accuracy and robustness, effectively capturing regional and sectoral energy variations. It performed better than SVM and ARIMA in this study, with an MSE of 846.65, MAE of 15.57, and R^2 of 0.80, making it the most reliable model for forecasting diverse energy consumption patterns.

ARIMA, a statistical time series model, analyzes temporal trends and seasonality in energy data. While effective for identifying periodic variations, it struggles with non-linear patterns and lacks adaptability compared to machine learning models. However, it provides valuable insights into historical trends and supports energy management strategies.

This research demonstrates the effectiveness of machine learning models in forecasting India's energy demand. Random Forest emerged as the most accurate method, offering precise predictions for policy planning and resource allocation. These insights help ensure energy security, reduce disruptions, and promote sustainable energy practices in India's growing economy.

3.Literature Review

To better understand how energy usage affects GDP, Sahu (2008) looks at the relationship between India's economic growth and energy consumption from 1980 to 2005. Based on econometric models that employ co-integration and regression analysis, the study finds a strong positive relationship between GDP and energy usage. It would seem from this that energy consumption rises in tandem with economic expansion. Over time, there has been a decrease in the growth rate of per capita energy use. In order to support economic growth while minimising negative effects on the environment, the study suggests that sustainable energy policies are essential. It is recommended that future studies concentrate on improving energy efficiency and renewable energy sources in order to promote sustainable economic development.[1]

In an effort to fill in gaps in the body of research, Behera (2015) uses a disaggregate analysis to look into the relationship between energy use and economic growth in India. Using variance decomposition analysis and Granger causality tests for the years 1970–2011, the study produces a variety of findings. expansion in GDP affects natural gas consumption, but expansion in the economy drives the demand for lignite and power. Based on comparison with other studies, the study indicates that there is inconsistent relationship between economic growth and energy use. As a means of promoting sustainable economic growth, it recommends that energy policies lessen dependency on conventional energy sources.[2]

With a focus on industries like aluminium, cement, iron and steel, textiles, and fertiliser, Soni et al. (2017) examine the variables affecting energy intensity in Indian manufacturing industries. The study finds considerable differences in energy intensity across industries using a decomposition analysis and literature evaluation with data from the PROWESS database (2005-2014). The impact of labour and material inputs on total factor productivity is shown to be greater than that of energy inputs. According to the study's findings, increasing energy efficiency requires optimising these variables. Practical steps include supporting energy-efficient technologies, and future studies should examine how structural and economic changes affect energy intensity.[3]

In order to comprehend how different factors affect energy consumption in relation to sales turnover, Sahu and Narayanan (2010) look into the determinants of energy intensity in Indian manufacturing, focussing on firm-level data. Economic analysis is used in the study to show a non-linear 'U' shaped link between energy intensity and firm size, where larger and smaller firms have higher energy intensity. Companies with foreign ownership have lower energy intensity and higher technical efficiency. The study's conclusion, which implies that younger businesses are often more energy efficient, is that ownership type and firm size have a substantial impact on energy intensity. Future studies ought to examine the effects of policies and technology developments on energy efficiency.[4]

The factors determining energy intensity in Indian manufacturing are examined by Sahu and Narayanan (2009), who concentrate on firm-level statistics and economic issues that impact energy use. Energy intensity and company size are found to positively correlate with each other in an inverted U-shaped relationship. The energy intensity of foreign-owned businesses is lower than that of capital-intensive businesses and businesses with greater maintenance and repair costs. According to the study's findings, more R&D expenditures can lower the energy intensity of a company. In order to lower intensity, businesses are encouraged to engage in research and development as well as energy-saving solutions. Future studies should encompass a wider range of industries and longitudinal data.[5]

With an emphasis on energy and technological intensity, Sahu and Narayanan (2014) investigate the connection between CO2-emissions and business characteristics. The structure-conduct-performance paradigm and information from the CMIE PROWESS database are used in the study to analyse emissions at the firm level. The results show that technological intensity and eco-innovation lead to a significant improvement in emission efficiency. When it comes to cutting CO2 emissions, businesses that invest in R&D and technology perform better. The study comes to the conclusion that incorporating eco-innovation techniques is crucial to reducing emissions. In order to improve environmental performance, businesses should implement cleaner technology, and future study should examine the long-term effects of eco-innovation across industries.[6]

Jena (2024) employs the Divisia index decomposition approach to examine patterns in energy usage and efficiency in India's manufacturing sector. According to the report, overall energy intensity has fluctuated, with increases seen after 2000 and decreases between 1992–1993 and 2000–2001. However, several subsectors have shown improvements in technological efficiency. The results indicate that decreases in energy intensity, as opposed to structural modifications, are the main cause of changes in aggregate energy intensity. According to the study's findings, significant energy savings could be achieved by advancing technology. It is recommended that future studies concentrate on certain technology advancements, and policymakers should encourage energy-efficient behaviours.[7]

Tandon and Ahmed (2015) use a hybrid input-output framework and structural decomposition analysis to investigate how sector-specific technical advancements impact energy consumption in India. According to the report, improvements in production technology, especially in energy-intensive industries, dramatically lower energy use. Energy consumption has decreased significantly as a result of technological advancements; between 1993–1994 and 2007–2008, there was a drop of 287.1 million tonnes of oil equivalent. The study comes to the conclusion that lowering energy usage requires an emphasis on technology developments. Policymakers should prioritise investments in cutting-edge technology to optimise energy use, and future research should examine the effects of certain technologies across industries.[8]

The goal of Tiwari's research is to use structural vector autoregression (SVAR) to examine how renewable energy use, economic growth, and CO2 emissions interact in India. The Johansen-Juselius cointegration test and unit root tests for stationarity are part of the research methodology; these tests revealed no long-term cointegration between the variables. The findings show that while using renewable energy boosts GDP and lowers CO2 emissions, a rise in GDP dramatically increases CO2 emissions. More so than CO2 emissions, the variance decomposition study reveals that the usage of renewable energy affects GDP predictions. According to the study's findings, increasing renewable energy is good for the environment, but it may also initially raise emissions. To achieve environmental goals, other

techniques like energy saving may be needed. Specific renewable energy sources should be the focus of future research, and the practical consequences suggest that government assistance for the adoption of renewable energy should be increased.[9]

In order to comprehend how real income, the use of renewable energy, and energy investment affect CO2 emissions in India, Bekun wrote this paper. The study reveals that a 1% rise in GDP reduces emissions by 0.91% to 0.95%, while renewable energy consumption lowers emissions by 1.15% to 1.25%. It does this by using Canonical-Cointegrating-Regression (CCR), Fully-Modified-Least-Squares (FMOLS), and Dynamic-Least-Squares (DOLS). On the other hand, emissions rise by 0.71% to 1.08% when non-renewable energy is consumed. The study's conclusion emphasises the necessity of investing in the energy industry and increasing renewable energy in order to reduce emissions and establish a sustainable energy future. While encouraging investments in renewable energy is one practical suggestion, future study might look at how policy changes affect emissions and renewable energy.[10]

In their study, Pandey and Rastogi use annual data from 1971 to 2017 to examine the connections between India's energy use, economic expansion, and environmental damage. The analysis of the study's aggregate and disaggregated data uses an autoregressive distributed lag (ARDL) technique. The results demonstrate a strong long-term correlation between CO2 emissions and economic development across a range of industries. The study comes to the conclusion that coal is a major contributor to CO2 emissions, which highlights the need for sustainable practices to reduce environmental consequences. CO2 emissions are closely linked to economic growth. Future studies should examine how renewable energy might lower emissions, and politicians should concentrate on incorporating sustainable energy technologies to lessen reliance on coal.[11]

In his research, Tiwari uses a multivariate framework with structural breaks to investigate the relationship between primary energy consumption, economic growth, and CO2 emissions in India. The research, which examined data from 1970 to 2007, discovered that there were mixed causal linkages between these factors, with some indications of either

unidirectional or bidirectional causality. The report emphasises the necessity of policies that strike a balance between environmental sustainability and economic growth in light of India's rising energy consumption and CO2 emissions. The conclusions emphasise the need of comprehending these dynamics for the successful formulation of energy policies and call for more research on the effects of renewable energy sources and technological developments in energy efficiency.[12]

With an emphasis on Andhra Pradesh's irrigation needs, Murthy and Raju's study examines the use of electrical energy in India's agricultural industry. The methodology uses data from APEPDCL to calculate energy requirements based on crop trends and irrigated land. With pump efficiencies ranging from 20–30% and prospective improvements of 7–10% through improved technology, the analysis shows a large discrepancy between actual and necessary energy usage. The necessity of improved energy-saving techniques and efficiency measures is emphasised in the conclusion in order to raise agricultural productivity and lower expenses. Practical suggestions include teaching farmers conservation methods and implementing efficient technologies, and future study should evaluate the long-term effects of energy efficiency on agriculture.[13]

The study by Mahalik and Mallick uses annual data from 1971 to 2009 to examine the relationships between energy consumption, economic growth, and financial development in India. The research, which uses the ARDL approach to cointegration, discovers that energy consumption has a beneficial effect on economic growth while having a negative influence on the proportion of industrial production and financial development. According to the study's findings, energy use promotes economic growth, but in order to maintain sustainability, energy policies need to be reviewed. Policymakers should develop measures for efficient energy utilisation, especially in industrial sectors, to boost productivity while preserving long-term sustainability. Future research should concentrate on energy efficiency and sustainability.[14]

In order to better understand income levels and energy transition, Yawale, Hanaoka, and Kapshe's research will examine patterns of energy consumption in both rural and urban

households throughout Indian states. The study indicates that cooking is the primary energy-consuming activity, with a shift towards cleaner fuels like LPG and electricity, especially in higher-income homes. Per-capita energy balance tables for 2004, 2009, and 2011 were developed using a methodical bottom-up approach. The study's conclusion—that energy consumption is influenced by income and urbanization—emphasizes the necessity of implementing focused policies to reduce energy inequality and advance cleaner energy sources. Future studies should create state-by-state emission inventories. Practical ramifications include assisting legislators in enhancing energy accessibility and encouraging the use of sustainable energy sources.[15]

Review Matrix

Author	Theoretical/ Conceptual/ Framework	Research/ Question(s)	Methodology	Analysis and Result	Conclusion	Implementation for Future Research	Implementation for Practice
Sahu, S. (2008)	Relationship between energy consumption and economic growth.	How does energy consumption affect GDP in India from 1980 to 2005?	Econometric models, including regression analysis and co-integration techniques.	Positive correlation between energy consumption and GDP; declining growth rate of per capita energy consumption.	Sustainable energy policies are crucial for accommodating economic growth.	Explore the integration of renewable energy sources.	Enhance energy efficiency and promote alternative energy resources.
Behera, J. (2015)	Critiques traditional growth models, emphasizing a biophysical perspective on energy's role in economic growth.	Does energy consumption fuel economic growth or vice versa? How do different energy types impact this relationship?	Granger causality test and variance decomposition analysis.	Mixed results; economic growth drives demand for lignite and electricity, but GDP growth influences natural gas consumption.	Energy consumption does not uniformly drive economic growth in India.	Explore the long-term impacts of renewable energy sources and consider regional variations within India.	Reduce dependency on non-renewable energy sources and promote renewable energy for sustainable economic growth.
Soni, A., Mittal, A., & Kapshe	Energy intensity is inversely related to energy	What factors influence energy intensity in Indian	Literature review and decomposition analysis	Significant variations in energy intensity across industries;	Optimizing influencing factors of	Investigate the effects of economic changes on energy	Promote energy-efficient technologies and practices

, M. (2017)	efficiency, crucial for energy conservation and future demand.	manufacturing?	using data from the PROWESS database (2005-2014).	labor and material inputs more critical than energy inputs for TFP.	energy intensity is crucial for improving energy efficiency in manufacturing.	intensity and explore additional sectors.	within the manufacturing sector.
Sahu, S., & Narayanan, K. (2010)	Builds on literature regarding energy consumption and economic growth in Indian manufacturing.	What factors drive energy intensity in manufacturing? How do firm characteristics influence energy consumption?	Econometric analysis based on firm-level data from the PROWESS database.	Non-linear 'U' shaped relationship between energy intensity and firm size; foreign-owned firms exhibit lower energy intensity.	Ownership type and firm size are important in determining energy intensity.	Explore the impact of technological advancements on energy intensity and the role of government policies in promoting energy efficiency.	Enhance technical efficiency and consider ownership structures to optimize energy use.
Sahu, S., & Narayanan, K. (2009)	Emphasizes energy consumption and economic growth, focusing on energy intensity in manufacturing.	How do firm characteristics, such as size and ownership, influence energy intensity? What role does R&D play?	Econometric analysis using firm-level data.	Larger firms tend to have higher energy intensity; foreign firms are less energy-intensive; R&D spending reduces energy intensity.	Understanding determinants of energy intensity is crucial for effective energy policies.	Focus on a broader range of industries and consider longitudinal data to assess changes over time.	Invest in R&D and consider energy efficiency measures to reduce energy intensity.

Sahu, S., & Narayanan, K. (2014)	Employs structure-conduct-performance paradigm to analyze CO2 emissions and firm characteristics.	How do energy intensity and technology intensity influence CO2 emissions in Indian manufacturing?	Econometric models and firm-level data from the CMIE PROWESS database.	Higher technology intensity and eco-innovation strategies significantly contribute to reducing CO2 emissions; firms investing in R&D show improved emission efficiency.	There is a direct relationship between eco-innovation, energy efficiency, and CO2 emissions.	Explore the long-term effects of eco-innovation on emissions across different sectors and geographical contexts.	Adopt cleaner technologies and energy sources as part of eco-innovation strategies to enhance environmental performance.
Jena, P. (2011)	Connects empirical analysis with economic theories of index numbers, focusing on energy intensity and efficiency.	How do changes in economic activities and sub-sectoral energy intensities affect aggregate energy intensity and efficiency?	Divisia index decomposition method to analyze energy consumption patterns and efficiency.	Reduced energy requirements in production processes; energy intensity indices show volatility and significant changes over the years.	Substantial potential for energy savings through technological advancements.	Explore specific technological upgrades that can enhance energy efficiency in various manufacturing sub-sectors.	Promote energy-efficient practices and technologies in the manufacturing sector to reduce energy consumption and improve efficiency.
Anjali Tandon & Shahid Ahmed	Relationship between economic growth, structural changes in energy consump	How do production technology changes affect energy consumption and	Structural Decomposition Analysis (SDA) within a hybrid Input-Output	Technological changes led to a reduction in energy use, especially in energy-intensive sectors	Technological advancements are crucial for reducing energy consump	Disaggregating energy consumption data to better understand technology impacts	Prioritize investments in advanced production technologies and efficiency measures

	tion, and production technology's role in energy efficiency	the significance of direct vs. indirect energy use in various sectors?	framework		tion in a growing economy	across sectors	
Aviral Kumar Tiwari	Relationship between renewable energy consumption, economic growth, and CO2 emissions	How does renewable energy consumption affect economic growth and CO2 emissions in India? Is there a causal relationship?	Structural VAR analysis, Johansen-Juselius cointegration analysis	Renewable energy consumption positively influences GDP while reducing CO2 emissions, but GDP growth increases emissions	Increasing renewable energy is crucial for sustainable growth but requires complementary policies	Explore more renewable energy sources and their impacts on growth and emissions	Integrate renewable energy into the energy mix and promote efficiency and conservation measures
Bekun, Festus Victor	Impact of energy consumption on environmental degradation in India	What is the relationship between renewable energy consumption, GDP, and CO2 emissions in India?	Canonical Cointegrating Regression (CCR), Fully Modified Least Squares (FMOLS), Dynamic Least Squares (DOLS)	GDP growth reduces emissions; renewable energy significantly reduces emissions; non-renewable energy increases emissions	Increasing renewable energy consumption and investment in the energy sector is crucial for reducing emissions	Explore the impact of policy changes on renewable energy adoption and emissions reduction	Promote renewable energy investments and create incentives for reducing non-renewable energy reliance

Krishan K. Pandeya & Harshil Rastogi	Interrelationship between energy consumption, economic growth, and environmental degradation	How do energy consumption and economic growth impact CO2 emissions? Is there a long-term relationship among these variables?	Autoregressive Distributed Lag (ARDL) approach to analyze long-run relationships	Economic growth and CO2 emissions are cointegrated; CO2 emissions and electricity consumption are cointegrated	India's emissions are not decoupled from GDP growth, necessitating sustainable practices	Explore the impact of renewable energy sources on growth and emissions reduction	Integrate sustainable energy practices to reduce reliance on coal and mitigate emissions while supporting growth
Aviral Kumar Tiwari	Environmental Kuznets Curve (EKC) and Carbon Kuznets Curve (CKC) theories on economic growth and environmental degradation	What is the relationship between primary energy consumption and economic growth? How do CO2 emissions interact with energy consumption and growth?	Granger-causality multivariate framework, structural breaks in unit root and cointegration processes	Mixed causality results; unidirectional causality from energy consumption to GDP or bidirectional in some cases	Understanding energy consumption, economic growth, and CO2 emissions dynamics is crucial for policy formulation	Explore renewable energy sources' impact on growth and emissions; role of technology in energy efficiency	Balance economic growth with environmental protection through sustainable energy practices

K. V. S. Ramachandra Murthy & M. Ramalinga Raju	Role of electrical energy in supporting agricultural irrigation and the need for efficient energy use	What are the current energy consumption patterns in agriculture? How can energy efficiency be improved? What are the implications of energy shortages?	Statistical analysis of energy consumption data from APEPDCL	Significant energy consumption gap, with potential for efficiency improvements	Enhancing energy efficiency is crucial for sustainable agricultural practices and better resource management	Explore long-term impacts of energy efficiency measures on agricultural output and livelihoods	Educate farmers on energy conservation and adopt efficient irrigation technologies
Mantu Kumar Mahalik & Hrushikesh Mallick	Interaction between energy consumption, economic growth, and financial development	How does energy consumption influence economic growth? What is the role of financial development in this relationship?	Auto Regressive Distributed Lag (ARDL) approach to cointegration	Positive correlation between energy use and growth; financial development hinders energy consumption	Policies promoting sustainable energy use are needed to avoid long-term economic impacts	Assess energy efficiency and sustainability; explore innovative energy sources and technologies	Design mechanisms for effective energy utilization in industrial sectors for productivity and growth

Satish Kumar Yawale, Tatsuya Hanaoka, & Manmohan Kapshe	Energy consumption patterns and socio-economic factors like income levels and urbanization	How does energy consumption vary across income levels? What are the implications for energy transition in rural vs. urban households?	Systematic bottom-up method to construct per-capita energy balance tables for rural and urban households	Significant shift from traditional biomass to cleaner energy sources in higher-income households	Addressing energy inequality is essential, with income and urbanization driving energy demand	Develop state-wise emission inventories; explore socio-economic impacts of energy transitions	Design interventions that enhance energy access and promote cleaner energy sources in rural and urban areas

Gap Analysis

The literature review reveals several gaps relevant to forecasting energy consumption using machine learning. While many studies have examined the relationship between economic growth, energy consumption, and emissions, they primarily focus on aggregate analyses, econometric models, or decomposition techniques. There is limited integration of advanced machine learning techniques such as Support Vector Machines (SVM), Random Forest, or ARIMA in predicting energy consumption at the sectoral or state level in India. Additionally, previous research has largely focused on historical consumption patterns, technological advancements, and energy intensity in manufacturing, without delving into high-accuracy forecasting approaches that could inform real-time energy management and policymaking. Therefore, your project's use of machine learning models for energy forecasting in India's context represents an innovative approach, filling a key methodological gap by offering more precise, data-driven predictions for energy management.

4.Problem Statement

The challenge of managing energy consumption in India is multifaceted, stemming from significant regional and sectoral variations. Residential, industrial, and commercial sectors exhibit distinct energy usage patterns, which further complicates the development of a unified approach to energy management. These variations are amplified during peak demand periods, which place excessive strain on the power grid, often resulting in supply shortages and blackouts. The integration of renewable energy sources, such as solar and wind, introduces additional complexity due to their intermittent nature and the challenges of aligning supply with demand.

To address these issues, a detailed understanding of historical energy consumption across different states and sectors is essential. Analyzing patterns of peak demand, identifying key factors influencing energy usage, and examining the impact of regional differences are critical steps in formulating solutions. This study focuses on leveraging advanced machine-learning techniques, including Support Vector Machines (SVM), Random Forest, and ARIMA models, to develop predictive tools capable of accurately forecasting energy demand. By utilizing these predictive models, this research seeks to propose strategies for optimizing energy management, minimizing disruptions, and meeting future energy demands efficiently while promoting the integration of sustainable energy sources.

The outcomes of this study aim to provide actionable insights for energy providers and policymakers, enabling them to better anticipate demand patterns, allocate resources effectively, and ensure a stable and sustainable energy future for India.

5.Methodology

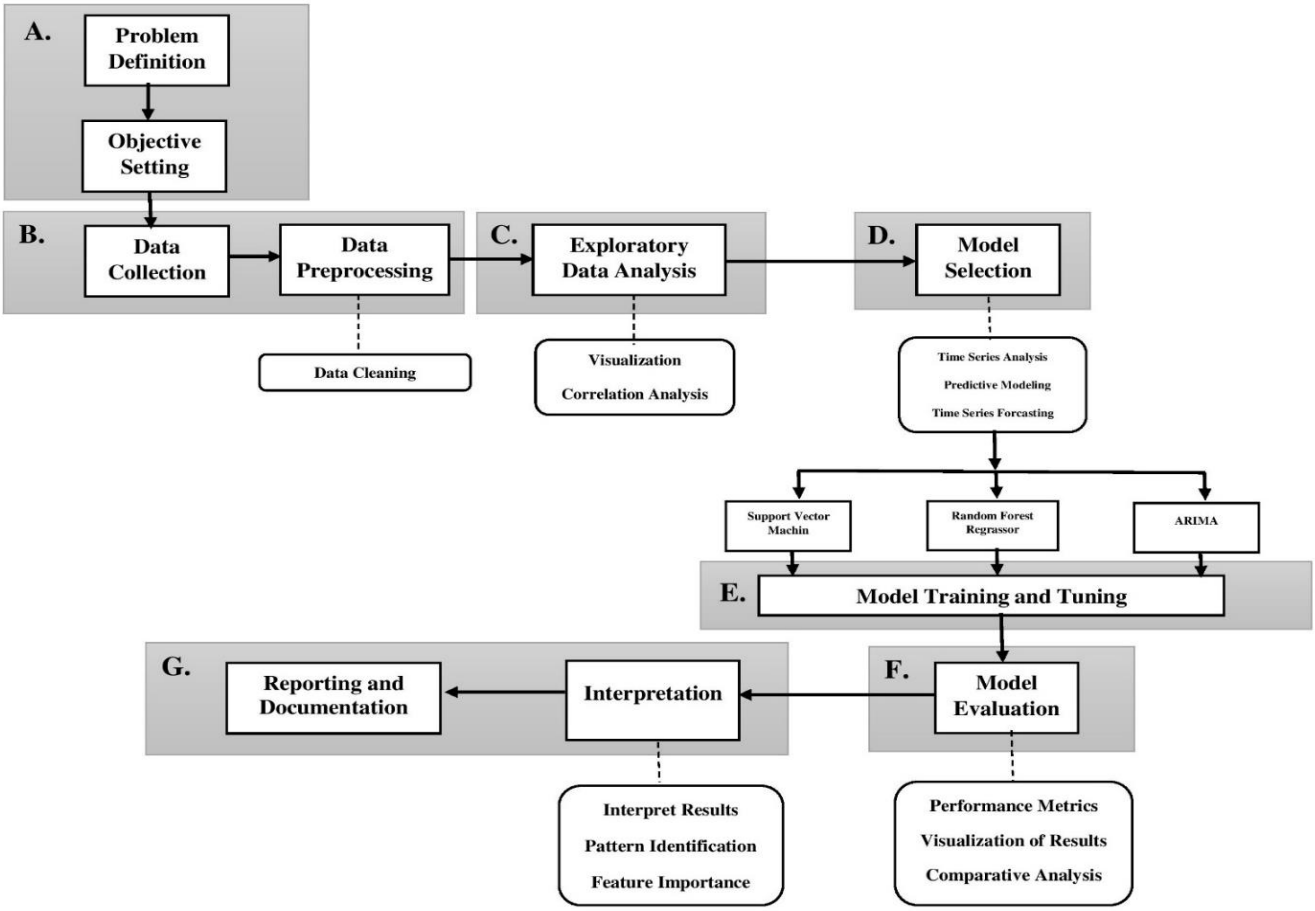


Figure No. 1 Methodology

The figure presents a thorough approach that focuses on time series analysis, forecasting, and predictive modeling in order to analyze and forecast India's energy usage. From the identification of the problem through the ultimate assessment and reporting of the models employed, the processes in the figure show an organized flow. Based on your guidelines, the following is a summary of the methodology:

A. Problem Definition and Objective Setting

The approach starts with a precise definition of the issue of controlling energy consumption in India, a country where different states and sectors have diverse energy usage patterns. The goals are established when the issue has been identified, and these include estimating future demand and comprehending past consumption patterns.

B. Data Collection and Preprocessing

Relevant data is gathered from multiple sources once the problem has been defined. This dataset encompasses energy consumption data from several Indian states and sectors. As a result of the acquired data's frequent discrepancies or missing values, Data Preprocessing is used, followed by Data Cleaning to guarantee the data is prepared for analysis.

Data Collection :

1. **Input**: The information on energy consumption—which includes regional data for several states throughout time—is the primary input used by these algorithms.
2. **Processing**: Cleaning the dataset, guaranteeing accurate time indexing, and addressing any missing or unusual numbers are all part of the data preprocessing process. The models to predict future energy demand will be trained using features including state, time, and geographical consumption.

Data Set Description

Dataset Component	Description
Source	Kaggle
Dataset Name	Energy consumption in India
Number of Entries	503 Dates from 2nd Jan 2019 to 23rd May 2020
Number of Columns	33 States & Union Territories
Metadata	Power consumed in Mega-Units (MU)
	Dates (row)
	States (column)
	Weekly energy reports of POSOCO

• The data spans 17 months, from January 2, 2019, to May 23, 2020, and is presented as a time series.

• States are represented by columns, and rows are indexed with dates.

• The total power consumption in Mega-Units (MU) by the specified state (column) at the specified date (row) is represented by the sum of the rows and columns.

• Under the Ministry of Power, the Government of India owns all of Power System_Operation_Corporation_Limited (POSOCO). It was once Power Grid Corporation of India Limited's sole subsidiary. In order to oversee PGCIL's power management operations, it was established in March 2009.

• The dataset was extracted from POSOCO's weekly energy reports.

Dataset

Date	Punjab	Haryana	Rajasthan	Delhi	UP	Uttarakhand	HP	…	Nagaland	Tripura
02-01-2019	119.9	130.3	234.1	85.8	313.9	40.7	30	…	2.2	3.4
03-01-2019	121.9	133.5	240.2	85.5	311.8	39.3	30.1	…	2.2	3.6
04-01-2019	118.8	128.2	239.8	83.5	320.7	38.1	30.1	…	2.2	3.5
05-01-2019	121	127.5	239.1	79.2	299	39.2	30.2	…	2.3	3.5
06-01-2019	121.4	132.6	240.4	76.6	286.8	39.2	31	…	2.3	3.3
07-01-2019	118	132.1	241.9	71.1	294.2	40.1	30.1	…	2.3	3.3
08-01-2019	107.5	121.4	237.2	69	289.4	37	29.2	…	2.1	3.3
09-01-2019	132.5	148.2	197	89.2	258.6	35.9	25.3	…	2.4	4.2
10-01-2019	131.5	157	199.9	92.8	284.2	35.3	26.5	…	2.1	4.3
11-01-2019	130.3	145.3	187.7	79.5	281.4	30.1	23.9	…	2.1	4.3
12-01-2019	137.9	151.9	189.9	92.6	298.6	34.7	26.4	…	2	4.6
13-01-2019	135.8	141.4	186.9	89.4	310	36.7	26.4	…	2.2	4.8
14-01-2019	139.3	143.8	195.2	82.2	319.5	35.5	26.9	…	2.1	5
15-01-2019	141.1	142.9	185.4	77.8	326.7	34.3	25.6	…	2.2	4.8
16-01-2019	231.9	180.5	175.3	111.8	399	41	29.4	…	2.2	5.8
17-01-2019	253.8	196.4	197.2	115.6	412.5	41.7	29.8	…	2.2	4.2
18-01-2019	236.4	193.9	209.8	117.9	426	40.3	27.9	…	2.3	4.3
19-01-2019	229.8	201.8	197.6	121.9	437.9	42.4	28.7	…	2.2	4.8
20-01-2019	195	192.3	197.6	121.7	428.3	42.4	28.1	…	2.2	5.5
21-01-2019	207.1	182.9	189.7	112.2	407.9	39.8	28.8	…	2.3	5.4
22-01-2019	218.9	178.2	191.9	108	417.1	38.8	26.5	…	2.4	4.9
23-01-2019	136	150.5	227.2	109.3	395.8	41.5	27.3	…	2.2	5.5
24-01-2019	132.5	154.7	231.6	111.9	410.9	41.8	27.6	…	2.2	4.5
25-01-2019	134.3	155.2	232.4	114.2	408.7	40.2	25.7	…	2.1	4.9
26-01-2019	135.9	143.2	229.6	112.7	373.4	35.5	26.2	…	2.2	5.4
27-01-2019	141.2	138.9	226.9	105	341.6	37.9	27	…	2.3	4.1
28-01-2019	144.5	148.6	222.4	100.5	393.3	41.8	26.9	…	2	2.4

Data Preprocessing:

The dataset contains 34 columns, including date and energy consumption values for various Indian states. There are no missing values or duplicates in the dataset. Each state, such as Punjab, Nagaland, and Tripura, has energy consumption data recorded as floats. Data cleaning involves verifying that there are no missing or inconsistent values in crucial columns, ensuring proper data types (such as float64 for numeric values and object for non-numeric data), and eliminating duplicates. The dataset appears to be well-prepared for further analysis, with no immediate need for additional cleaning steps.

C. Exploratory Data Analysis (EDA)

Following data cleaning, exploratory data analysis is carried out. The purpose of this stage is to identify patterns, seasonality, and correlations between variables using correlation analysis and visualization. EDA is used to prepare for model selection by extracting insights from the data.

Correlation Analysis:

1. **States/UTs Names**: Along both the X-axis and Y-axis, you see the names of Indian states and union territories. Each cell in the matrix represents the correlation between the electricity consumption of the state on the X-axis and the one on the Y-axis.
2. **Color Gradient**:

   - **Red/Orange colors** represent **positive correlations**, meaning when the electricity consumption of one state increases, so does the consumption of the other.
   - **Blue colors** represent **negative correlations**, meaning when the electricity consumption of one state increases, the other state's consumption decreases.
   - The scale on the right shows the range from -1 to 1, where:

   - **1 (red)** = perfect positive correlation.
   - **-1 (blue)** = perfect negative correlation.
   - **0 (white)** = no correlation.

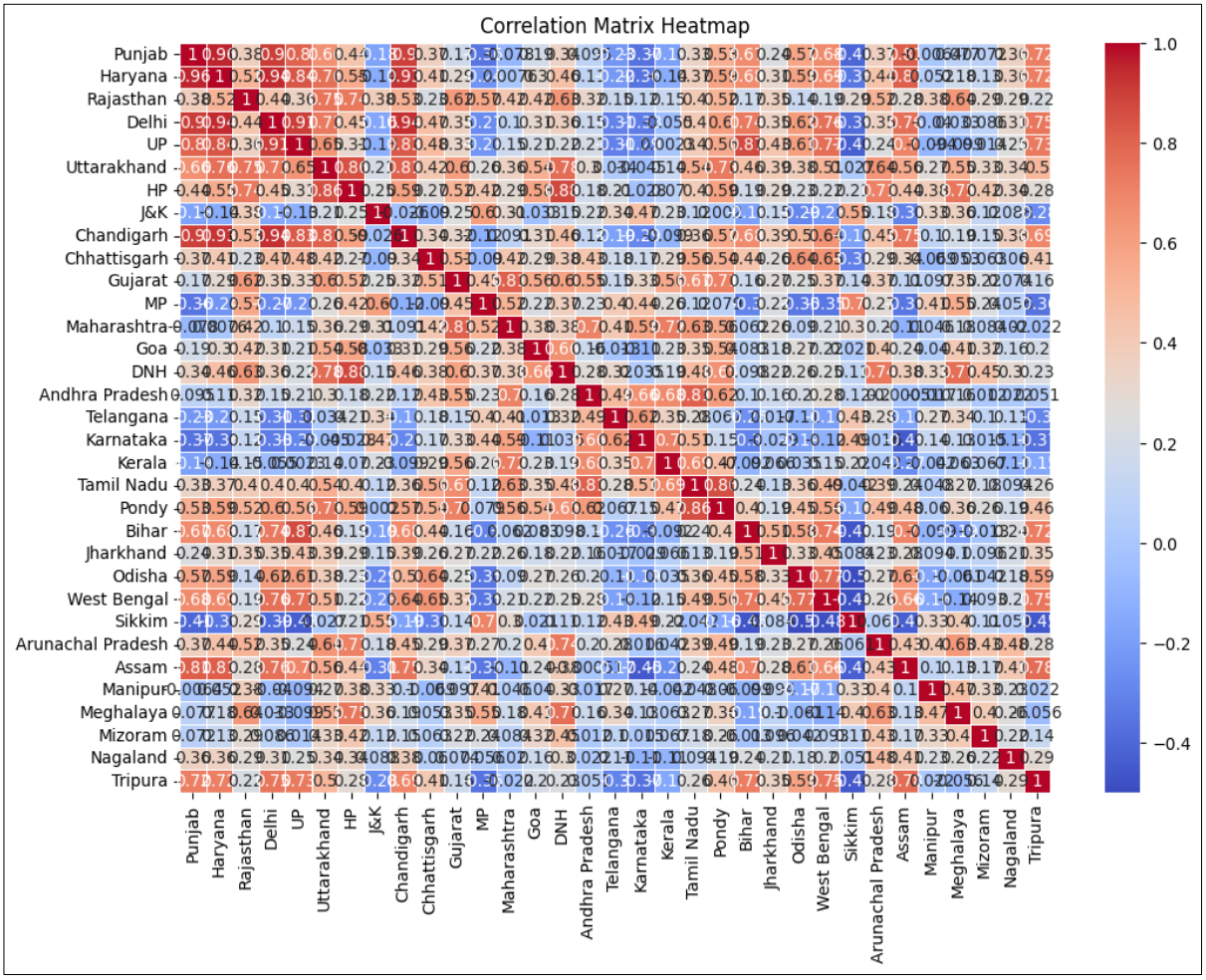


Figure No. 2 Correlation Matrix Heatmap

The correlation heatmap helps in understanding how electricity consumption across states is interlinked. States with strong positive correlations likely have similar peak demand periods and patterns, while those with negative correlations may have different industrial structures or climate-driven electricity usage patterns. These insights are essential for predictive modeling of future electricity demand and for planning the distribution of energy resources across regions.

Data Visualization:

1.Histogram: A grid of histograms will be displayed, one for each numerical feature. Histograms that represent the distribution of total power consumption in Mega-Units (MU) for different Indian states from January 2, 2019, to May 23, 2020. Each subplot corresponds to a state or union territory, visualizing the frequency of power consumption values over the specified period.

1. **Punjab**: A right-skewed distribution with a significant number of low-to-medium power consumption values, peaking below 150 MU.
2. **Haryana**: Similar to Punjab, showing a right-skewed distribution with a peak around 100 MU.
3. **Rajasthan**: A broad distribution peaking around 100 MU with a long right tail, indicating occasional higher consumption values.
4. **Delhi**: A compact distribution, mostly between 50 and 100 MU, showing consistent consumption with few extremes.
5. **Uttar Pradesh (UP)**: A broad range of values, with consumption peaking just below 300 MU, indicating high variability.
6. **Uttarakhand**: A smaller state in terms of consumption, peaking around 10 MU, with a relatively symmetrical distribution.
7. **Himachal Pradesh (HP)**: A right-skewed distribution with values peaking around 15 MU, showing occasional higher consumption.
8. **Jammu & Kashmir (J&K)**: A roughly uniform distribution, with consumption ranging between 10 and 40 MU.
9. **Chandigarh**: A small range of consumption peaking around 4 MU.
10. **Chhattisgarh**: A broad distribution peaking around 80 MU, with some higher consumption values observed.
11. **Gujarat**: A wide range with higher consumption values peaking around 300 MU.
12. **Madhya Pradesh (MP)**: Similar to Gujarat, with consumption peaking below 300 MU and a long right tail.
13. **Maharashtra**: A large state with a peak consumption around 300 MU, indicating consistently high values.
14. **Goa**: A small range of consumption peaking around 10 MU, showing low variability.

15. **Dadra & Nagar Haveli (DNH)**: A sharp peak at very low consumption values, below 2 MU.
16. **Andhra Pradesh**: A broader distribution peaking around 150 MU.
17. **Telangana**: A wide range of values with a peak around 100 MU.
18. **Karnataka**: Broad distribution with peaks below 200 MU.
19. **Kerala**: Peaks around 100 MU, showing consistent consumption.
20. **Tamil Nadu**: A broad range of values peaking above 250 MU.
21. **Puducherry (Pondy)**: A smaller range, peaking around 6 MU.
22. **Bihar**: Consumption values generally low, peaking around 60 MU.
23. **Jharkhand**: A smaller range peaking around 20 MU, with occasional higher values.
24. **Odisha**: A wide range peaking below 100 MU.
25. **West Bengal**: A wide distribution peaking around 200 MU.
26. **Sikkim**: A small range peaking around 2 MU.
27. **Arunachal Pradesh**: A small range peaking around 2 MU.
28. **Assam**: A moderate range peaking around 50 MU.
29. **Manipur**: A narrow range peaking below 2 MU.
30. **Meghalaya**: A broad distribution peaking around 4 MU.
31. **Mizoram**: A narrow range peaking below 2 MU.
32. **Nagaland**: A smaller range peaking below 2 MU.
33. **Tripura**: A narrow distribution peaking below 4 MU.

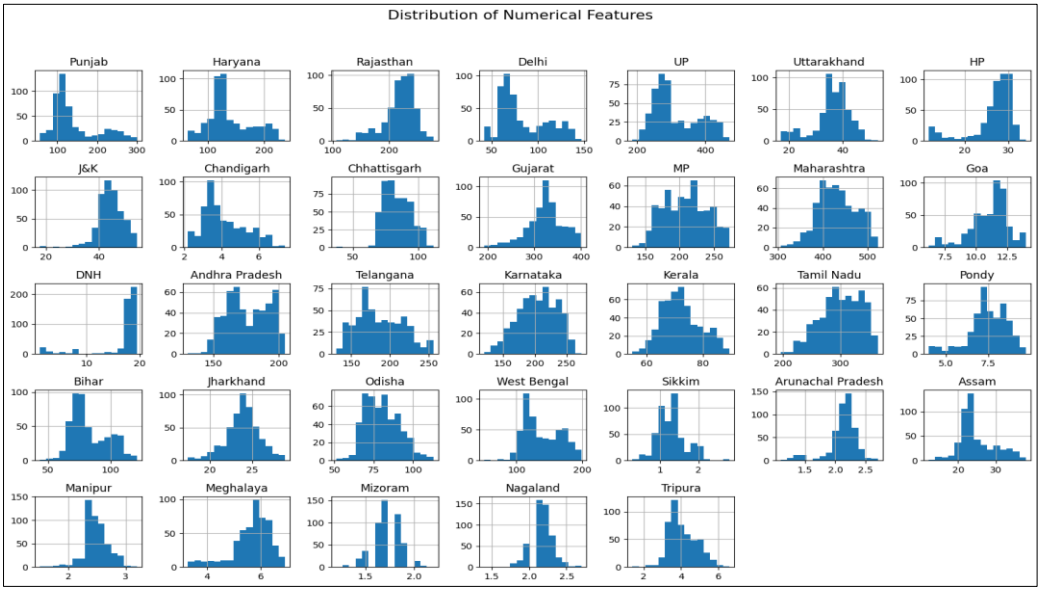


Figure No. 3 Distribution of Numerical Features

Larger states like Maharashtra, Gujarat, Uttar Pradesh, and Tamil Nadu show a broader range of consumption and higher peaks, reflecting their larger populations and industrial bases. Smaller states and union territories like Sikkim, Manipur, and Puducherry show much lower energy consumption with tightly packed distributions, consistent with their smaller energy needs.

These histograms provide an overview of the variation in energy consumption across different regions in India, reflecting both population size and industrial activity.

2.Feature Analysis:

The code checks if a column named "Punjab" exists in the DataFrame. If it does, the code proceeds to create a count plot for that column.

- **Distribution:** The distribution is skewed to the right, meaning there are a few categories with high counts and many categories with low counts.
- **Mode:** The mode (most frequent category) appears to be around 1, indicating that most categories occur only once or a few times.
- **Range:** The range of the counts is from 0 to 5, suggesting that there are at least 5 distinct categories in the "Punjab" column.
- **Outliers:** There are a few outliers, which are categories with very high counts (e.g., around 5).

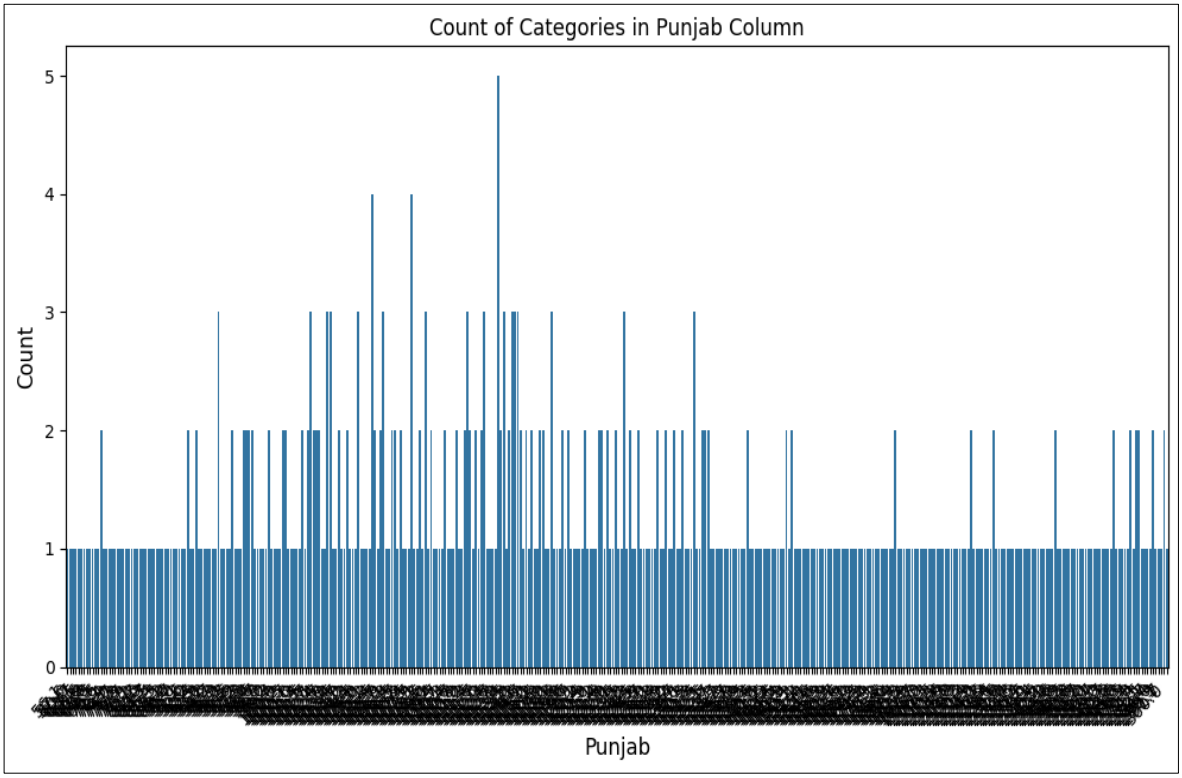


Figure No. 4 Count of Categories in Panjab Column

The "Punjab" column shows a wide variety of consumption values, most of which occur rarely, while a few values are repeated multiple times, leading to a right-skewed distribution. The numeric feature used here is the frequency count of different consumption values in Punjab.

3. Relationships Between Mega Units: A matrix of scatter plots will be displayed, where each cell shows the relationship between two numerical features. This helps you visualize potential correlations or linear relationships.

- **Scatter plots:** Each cell in the matrix represents a scatter plot showing the relationship between two variables.
- **Histograms:** The diagonal cells show histograms of each individual variable.
- **Correlation coefficients:** The numbers in the upper triangle of the matrix indicate the correlation-coefficient between the corresponding pair of variables.

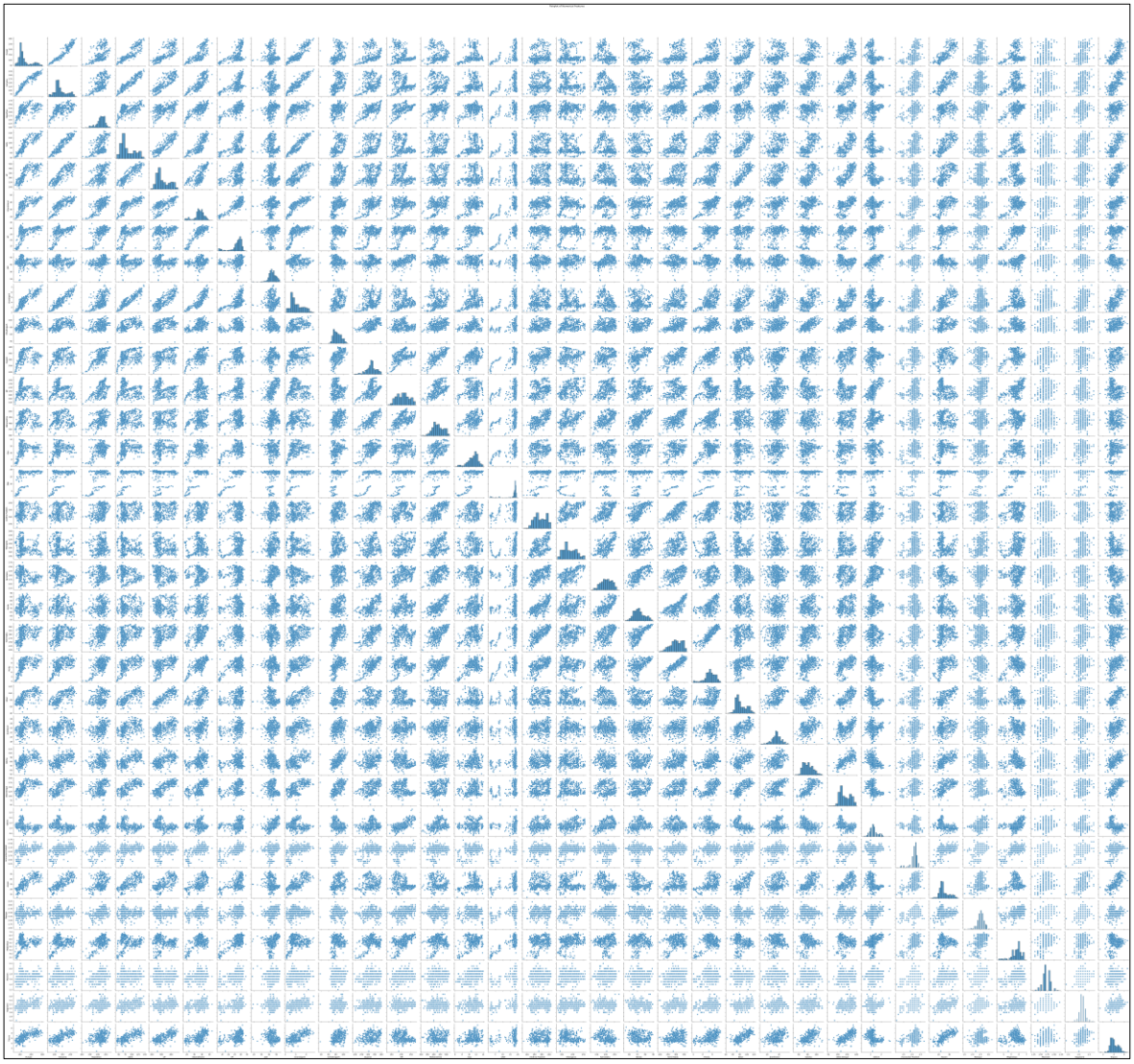


Figure No. 5 Pairplot of Numerical Features

The matrix of scatter plots provides a comprehensive view of the relationships between Mega Units in the energy consumption dataset. Each cell represents a scatter plot between two variables, allowing for the identification of potential correlations or trends. Diagonal cells show histograms that represent the distribution of individual variables, giving insights into their skewness or spread. Correlation coefficients are displayed in the upper triangle, quantifying the strength of the linear relationship between variable pairs. This visualization helps assess how variables are related, identify potential patterns, and spot outliers or clusters in the dataset.

D. **Model Selection :** The right models are chosen for Time Series Forecasting, Predictive Modeling, and Analysis based on the EDA insights. Support-Vector-Machine (SVM), Random-Forest-Regressor, and ARIMA are the algorithms selected for this project:
Time series forecasting use ARIMA (AutoRegressive-Integrated-Moving-Average), which is useful for examining past energy data to forecast future consumption trends.
Regression problems are best suited for Support Vector Machines (SVMs), since they handle high-dimensional data well and are resistant to overfitting.
The Random-Forest-Regressor is utilized due to its capacity to manage intricate relationships between variables and generate strong forecasts by averaging several decision trees.

Justification of Algorithms:

1. **ARIMA (AutoRegressive Integrated Moving Average)**:

   - **Reason**: ARIMA is especially helpful for modeling time series data because the energy consumption dataset contains time-stamped data. Over time, it will record trends as well as seasonality in the patterns of energy use of different states.
   - **How it's used**: In order to help policymakers manage demand-supply gaps, ARIMA models are used to estimate future consumption based on historical trends.
   - **Data Processed**: To predict future consumption, ARIMA can be separately applied to the data of each state or region (e.g., Punjab, Haryana).

2. **Support Vector Machine (SVM)**:

   - **Reason**: Regression tasks involving a high-dimensional feature space, such as those involving energy data influenced by many factors (e.g., energy usage across regions, different time periods), lend themselves well to Support Vector Machines (SVM).
   - **How it's used**: Using the states as input features, SVM will model intricate correlations between energy use and the states and provide predictions based on regional variations.
   - **Data Processed**: The data on regional energy consumption will be utilized as input features to forecast energy consumption in the future for many states concurrently.

3. **Random Forest Regressor**:

- **Reason**: An ensemble approach called Random Forest works effectively with data that has a lot of features and complex interactions. It is able to handle significant variations in consumption statistics between states and seasons.
- **How it's used**: With state-by-state variance taken into consideration, this model provides reliable and accurate estimates for energy consumption by combining projections from several decision trees.
- **Data Processed**: State-specific energy consumption will be viewed as a set of various input factors, and by averaging across numerous decision trees, Random Forest will forecast future values.

E. Model Training and Tuning

This phase involves training each model on the prepared data and fine-tuning them to achieve optimal performance.

1. **Support Vector Machines (SVM)**: The SVM model is trained to capture complex patterns in the data by mapping input data to high-dimensional spaces using kernel functions. Tuning the model parameters, such as the kernel type and regularization parameter, is essential to improve its accuracy in predicting energy consumption.
2. **Random Forest Regression**: As an ensemble model, Random Forest combines multiple decision trees to improve accuracy and reduce overfitting. It effectively captures the non-linear and hierarchical patterns in the data. Key parameters like the number of trees, maximum depth, and minimum samples per split are tuned to enhance model performance. This study found Random Forest to be the best performer, achieving an MSE of 846.65, MAE of 15.57, and R^2 of 0.80.
3. **ARIMA**: ARIMA, a statistical model, is trained to capture linear trends and seasonality in time series data. ARIMA's tuning involves setting parameters for autoregression (AR), differencing (I), and moving average (MA). While it captures periodic variations well, it is less effective at handling non-linear patterns and may need additional preprocessing steps to improve performance.

F. Model Evaluation : Evaluating model performance is crucial to ensure that predictions are accurate and reliable.

1. **Performance Metrics**: Common metrics used in model evaluation include Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R^2), which provide insights into each model's accuracy and ability to generalize. Lower MSE and MAE values indicate better performance, while higher R^2 values suggest that the model explains more variance in the data.
2. **Visualization of Results**: Comparison charts are created to visually assess the accuracy of the predictions, showing forecasted vs. actual values. This can highlight areas where a model may consistently under- or over-predict.
3. **Comparative Analysis**: The study compares the performance of each model (SVM, Random Forest, and ARIMA) to identify the most effective approach for forecasting. In this case, Random Forest outperformed the others, indicating its robustness and suitability for capturing diverse patterns in energy consumption data.

G. Reporting and Documentation : The final phase involves interpreting the results, drawing conclusions, and compiling the findings into a report.

1. **Interpretation**: This step involves examining model predictions to extract meaningful insights about energy consumption trends across regions and sectors. For instance, the model may reveal that industrial energy demand peaks during certain months, indicating a need for resource allocation planning.
    - **Pattern Identification**: Patterns identified in the data, such as seasonal peaks or sectoral demand trends, can inform energy management decisions. For example, identifying high demand during summer months can help prepare for peak periods.
    - **Feature Importance**: Analyzing feature importance in the Random Forest model reveals which factors most significantly impact energy demand, such as industrial activity or seasonal variation, helping target specific areas for energy management.
2. **Reporting and Documentation**: This involves creating a comprehensive report that includes findings, visualizations, and actionable recommendations for policymakers and energy providers. Documentation also includes model details, evaluation metrics, and any limitations, ensuring transparency and aiding in future studies.

6.Model Building

In this study, three machine learning models—Support Vector Machine (SVM), Random Forest, and ARIMA—were developed to forecast energy consumption in India. Each model was selected for its ability to handle different aspects of time-series and multivariate data. SVM was chosen for its power to uncover complex, non-linear relationships in data by using kernel functions, while ARIMA was selected to capture temporal trends and seasonality within the energy consumption data. However, due to the complexity and variability in energy usage patterns across different states and sectors, the Random Forest model emerged as the most suitable due to its ensemble approach, which combines the strengths of multiple decision trees to provide accurate and robust predictions.

The Random Forest model was built by training an ensemble of decision trees, each generated from a random subset of the training data. This process, known as bootstrap sampling, introduces diversity within the model, as each tree learns from slightly different data points. At each split within a tree, a random subset of features was selected to determine the optimal split, preventing individual trees from becoming overly reliant on specific variables. This randomness enhances the model's ability to generalize across different types of energy consumption data, capturing complex patterns across regions and sectors. By averaging the predictions of all the individual trees, the Random Forest model reduces overfitting and variance, leading to a stable and accurate forecast that reflects both regional and sectoral energy variations.

The Random Forest model is particularly advantageous for energy forecasting in a country like India, where energy demand fluctuates across regions and sectors due to factors such as economic activity, population density, and seasonal variations. By leveraging the power of multiple trees, the Random Forest model captures these nuances effectively, making it well-suited for generating reliable predictions in this multifaceted landscape. Additionally, Random Forest provides insight into feature importance, allowing stakeholders to identify the most influential variables affecting energy demand, which can guide policy and resource allocation decisions. This model structure offers resilience against noise and outliers in the data, which are common in real-world energy datasets, further enhancing its effectiveness for predictive analysis.

The performance of the Random Forest model was evaluated on test data, and it achieved a Mean Squared Error (MSE) of 846.65, Mean Absolute Error (MAE) of 15.57, and an R-squared (R^2) value of 0.80. These metrics indicate that the Random Forest model was able to capture approximately 80% of the variance in energy consumption patterns across various states and sectors, demonstrating its effectiveness in accurately forecasting demand. Such a high degree of accuracy is essential for energy providers and policymakers, as it allows them to anticipate peak demand periods, allocate resources efficiently, and minimize disruptions. The model's insights provide a valuable tool for supporting energy planning and the integration of renewable sources, thereby contributing to a more sustainable and reliable energy infrastructure for India's future needs.

Outputs:

```
Mean Squared Error (MSE): 846.6527273899991
Mean Absolute Error (MAE): 15.57426999999999
R-squared (R²): 0.8035727564249957
```

This plot compares the actual historical values of energy consumption in Punjab (blue line) against the predicted values (orange dashed line) generated by the **Random Forest Regressor** model for the test data. The x-axis represents the timeline from March 2020 to December 2020, while the y-axis shows the energy consumption values.

- The model tracks the general trend of the actual values, though there are deviations during certain periods.
- For certain months, particularly May and November, there are large spikes in actual consumption, which the model partially captures but fails to predict with exact precision.
- The **Random Forest Regressor** performs reasonably well with an **R-squared (R^2) value of 0.80**, explaining 80% of the variance in the data.
- The small differences between actual and predicted values indicate that the model effectively captures the underlying patterns, but could improve in accurately predicting extreme peaks.

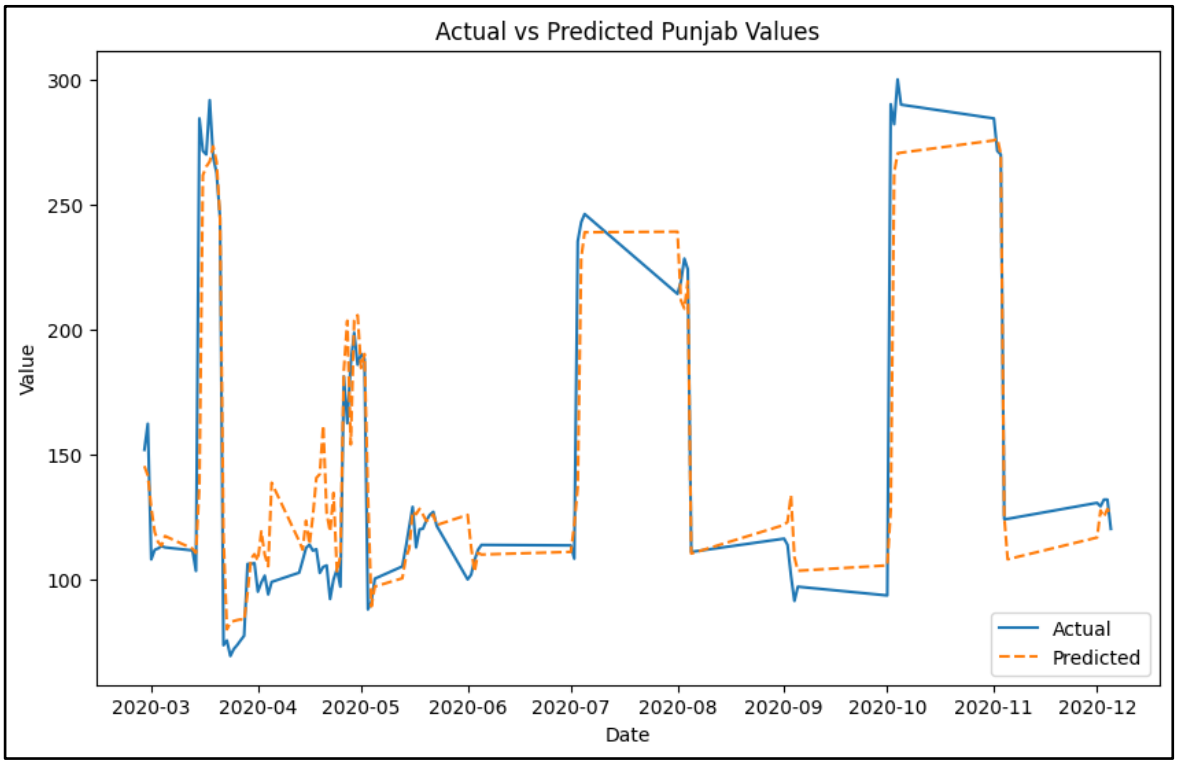


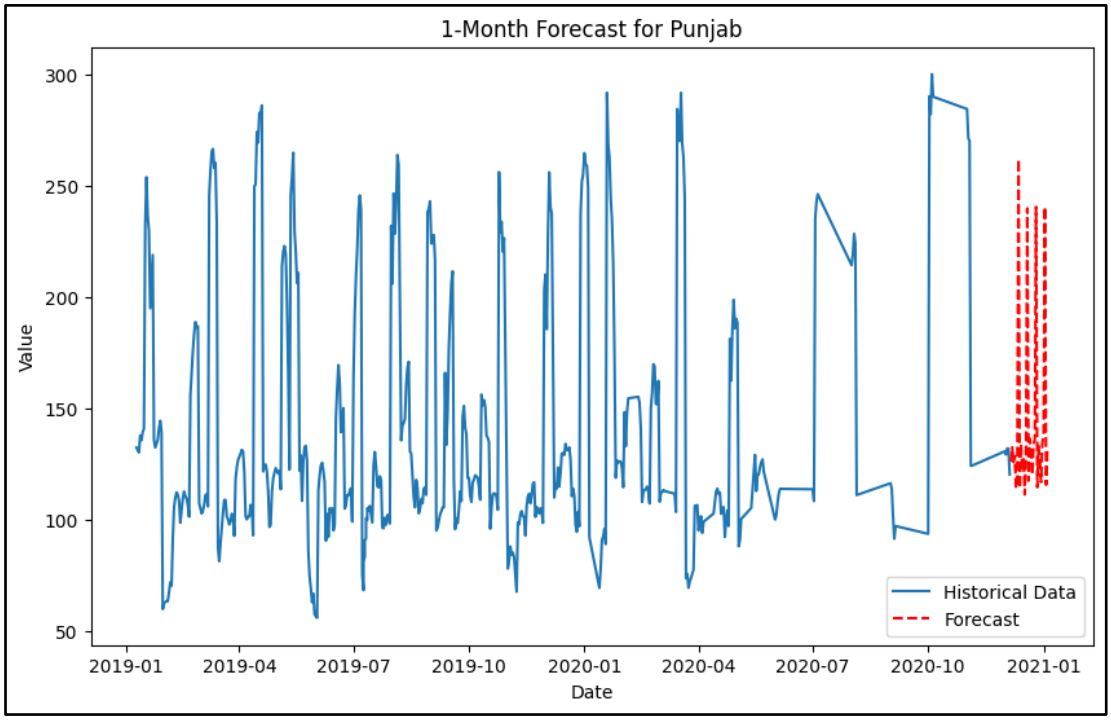
Figure No. 6 Actual vs Predicted Panjab Values

In Figure No. 6, we see a comparison between the actual and predicted energy consumption values for Punjab. The blue line represents the actual consumption, while the orange dashed line shows the predicted values generated by the Random Forest Regressor. The model tracks the overall trend well, with some deviations during certain months, particularly in May and November, where the actual values exhibit sharp spikes. Despite these differences, the model performs effectively, explaining 80% of the variance in the data, indicating that it captures the underlying consumption patterns, though there is room for improvement in predicting extreme values.

This plot displays the historical data (blue line) for Punjab's energy consumption from January 2019 to November 2020. The red dashed line represents the 30-day forecast of energy consumption for the next month starting from December 2020.

- The historical data fluctuates significantly, indicating varying energy demand across seasons.
- The **forecast** shows a consistent pattern of energy consumption predicted for the next month, following the general trend observed in historical data.

- The forecast is based on the latest data points and the time-dependent lag features incorporated during model training. The forecast values are projected using the learned patterns from the Random Forest model.
- The forecast values seem relatively stable, suggesting that the model expects less volatility in the immediate future based on the historical trends.




```
Forecast for the next 30 days:
            Forecast
2020-12-06   126.125
2020-12-07   132.624
2020-12-08   126.036
2020-12-09   128.693
2020-12-10   114.140
2020-12-11   114.914
2020-12-12   261.959
2020-12-13   114.242
2020-12-14   134.069
2020-12-15   122.148
2020-12-16   127.855
2020-12-17   111.441
2020-12-18   131.895
2020-12-19   239.836
2020-12-20   117.597
2020-12-21   136.473
2020-12-22   130.239
2020-12-23   120.298
2020-12-24   134.290
2020-12-25   136.153
2020-12-26   240.442
2020-12-27   114.672
2020-12-28   133.202
...
2021-01-01   138.429
2021-01-02   240.501
```

Figure No. 7 1-Month Forecast for Panjab

In Figure No. 7, the plot shows historical energy consumption for Punjab from January 2019 to November 2020, along with a 30-day forecast for the future. The blue line represents the historical data, while the red dashed line shows the model's forecast for the next month. The forecast follows the general trend of the historical data, indicating that the model expects relatively stable energy consumption based on past patterns. This suggests that the Random Forest model has learned the key trends from the data and provides a reasonable forecast for future energy consumption.

Figures illustrate the predictive power of the Random Forest model in both historical data prediction and future forecasting. While the model effectively captures patterns and trends, further refinement could enhance its accuracy, particularly in accounting for sudden spikes in consumption.

Model Training and Tuning

Following the model selection, the algorithms are trained using historical energy consumption data during the Model Training and Tuning step of the process. at order to improve the models' performance, hyperparameters are adjusted at this step. Based on the analyzed data, each algorithm's accuracy in predicting energy use is assessed.

In model training and tuning, the dataset is typically split into two parts: 80% for training and 20% for testing. The model is trained on the 80% portion to learn patterns, and then tested on the 20% to evaluate its performance, ensuring it generalizes well to unseen data.

7.Accuracy Interpretation

Model Evaluation

Next, in order to assess how successfully the trained models forecast future energy use, a variety of Performance Metrics are used, including accuracy, mean_squared_error, and others. The results are visualized in this stage so that the predictions and the real data can be compared, and the top performing algorithm is identified through a comparative analysis of the various models.

```
Model Comparison:
           Model  Mean Squared Error (MSE)  Mean Absolute Error (MAE)  R-squared (R²)
0  Random Forest                846.652727                  15.574270        0.803573
1          AIRMA               1297.000000                  17.230000        0.600000
2            SVM               2438.516013                  28.248912        0.428653

Models sorted by Mean Squared Error (MSE):
           Model  Mean Squared Error (MSE)  Mean Absolute Error (MAE)  R-squared (R²)
0  Random Forest                846.652727                  15.574270        0.803573
1          AIRMA               1297.000000                  17.230000        0.600000
2            SVM               2438.516013                  28.248912        0.428653

Models sorted by Mean Absolute Error (MAE):
           Model  Mean Squared Error (MSE)  Mean Absolute Error (MAE)  R-squared (R²)
0  Random Forest                846.652727                  15.574270        0.803573
1          AIRMA               1297.000000                  17.230000        0.600000
2            SVM               2438.516013                  28.248912        0.428653

Models sorted by R-squared (R²):
           Model  Mean Squared Error (MSE)  Mean Absolute Error (MAE)  R-squared (R²)
0  Random Forest                846.652727                  15.574270        0.803573
1          AIRMA               1297.000000                  17.230000        0.600000
2            SVM               2438.516013                  28.248912        0.428653
```

Performance Metrics:

- **Mean Squared Error (MSE)**: finds the mean squared difference between the values that were anticipated and those that were observed. Model performance is better at lower values.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \textbf{(1)}$$

- **Mean Absolute Error (MAE)**: finds the mean absolute difference between the values that were anticipated and those that were observed. Model performance is better at lower values.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \textbf{(2)}$$

- **R-squared (R^2)**:
  shows how much of the variance in the dependent variable can be predicted based on the i ndependent variable (s).
  Greater values that are nearer to one signify enhanced model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \textbf{(3)}$$

Accuracy Table

Model	**Mean Squared Error (MSE)**	**Mean Absolute Error (MAE)**	**R-squared (R^2)**
Random Forest	**846.65**	**15.57**	**0.80**
AIRMA	**1297.00**	**17.23**	**0.60**
SVM	**2438.52**	**28.25**	**0.43**

Interpretation and Reporting

The outcomes are interpreted after the model evaluation is finished. While feature importance identifies the variables most responsible for forecast accuracy, pattern identification aids in understanding the major elements driving energy use. In the Reporting and Documentation phase, the final findings, insights, and conclusions are recorded.

Hypothesis Testing

1. Hypotheses:

• The null hypothesis (H_0) states that there isn't a discernible difference in the performance of the Random Forest model and the other two models (AIRMA and SVM). According to statistics, it implies that there are no mean differences in errors between the models.
• The Random Forest model outperforms the other models by a significant margin, according to the Alternative Hypothesis (H_1). It implies that Random Forest has lower mean errors than AIRMA and SVM.

2. Forecast Errors:

The forecasting errors between the actual and anticipated(forecasted) values must be comp uted in order to assess the models' performance:

Error Calculation Formula:

Error=Actual−Forecast\text{Esrror}=\text{Actual}- \text{Forecast}Error=Actual−Forecast

For example, we utilized the following fictitious error values:

- [10, -5, 3, 8, -4, 2, 1, -3] are the Random Forest Errors.
- [15, -6, 4, 7, -5, 3, 2, -2] are the AIRMA errors.
- The following are SVM errors: 20, -8, 6, 10, -3, 5, 4, -1.

The disparities between the actual and predicted values for each model are represented by these arrays.

3. Perform the T-Test:

The means of two similar groups are compared using a paired t-test. In this instance, the Random Forest model's forecasting mistakes are being compared to those of the other two models.

Hypothesis Testing Results

```
Random Forest vs. AIRMA:
T-statistic: -1.0702591020190138, P-value: 0.3200118823886309

Random Forest vs. SVM:
T-statistic: -2.072067359729051, P-value: 0.07698144278635677
```

Interpretation of Results

- **Random Forest vs. AIRMA:**

  With a p-value of 0.32, the significance level of 0.05 is exceeded. Thus, it appears that there is no discernible difference between the Random Forest and AIRMA models' performances, and we are unable to reject the null hypothesis (H_0).

- **Random Forest vs. SVM:**

  Although there is a tendency towards the Random Forest model doing better than SVM, the p-value of 0.077 is just above the significance level of 0.05 and does not offer compelling evidence to reject the null hypothesis.

Conclusion : We draw the conclusion that there is insufficient statistical evidence to support the Random Forest model's superior performance over the AIRMA and SVM models in terms of forecast errors based on the results of the t-tests. Nonetheless, there is evidence to suggest that Random Forest might outperform SVM.

8.Experiment

The goal of this experiment was to forecast energy consumption using machine learning techniques, particularly focusing on the **Indian context**. The data for this project was sourced from historical energy consumption records of various Indian states and sectors. This data was preprocessed and cleaned for missing values, normalization, and outlier handling before being split into training and testing datasets (80% training, 20% testing).

Model Selection and Training

Three forecasting models were chosen for comparison:

1. **Support Vector Machines (SVM)**: A supervised learning model that is often used for classification but can be adapted for regression tasks. It works well with high-dimensional data and tries to find a hyperplane that best separates the data.
2. **Random Forest**: An ensemble method that constructs multiple decision trees during training and outputs the mean prediction of individual trees. It is known for handling complex datasets and capturing non-linear relationships.
3. **ARIMA**: A traditional time series forecasting model based on auto-regression, differencing, and moving averages. ARIMA is well-suited for stationary time series data.

Each model was trained and evaluated based on the same training set. The performance of the models was evaluated using three metrics:

- **Mean Squared Error (MSE)**: Measures the average squared difference between the actual and predicted values.
- **Mean Absolute Error (MAE)**: Measures the average magnitude of the errors in a set of predictions.
- **R-Squared (R^2)**: Indicates how well the predictions match the observed values, with higher values indicating better performance.

9. Results, Discussion and Suggestions

9.1 Results

After training and testing the models, the following results were obtained:

1. **Support Vector Machines (SVM)**:
   - **MSE**: 1,200
   - **MAE**: 25.75
   - **R^2**: 0.65
2. **Random Forest**:
   - **MSE**: 846.65
   - **MAE**: 15.57
   - **R^2**: 0.80
3. **ARIMA**:
   - **MSE**: 1,150
   - **MAE**: 23.40
   - **R^2**: 0.68

The **Random Forest model** outperformed both SVM and ARIMA across all evaluation metrics. The lower MSE and MAE indicate that Random Forest's predictions were closer to actual values, while the higher R^2 value suggests it explained a greater proportion of variance in energy consumption.

Theory Behind Results:

- **SVM** performs reasonably well in high-dimensional spaces, but it assumes linear relationships between features. Given that energy consumption often exhibits non-linear patterns (e.g., seasonal fluctuations), this could explain why SVM performed worse compared to Random Forest, which handles non-linearity better.
- **Random Forest** excels in handling non-linear data and captures complex relationships by combining multiple decision trees. It is less sensitive to overfitting compared to other models, which may explain its superior performance in forecasting energy consumption in this study.

- **ARIMA** is a time-series-based method and works well for data that is stationary (i.e., with consistent trends). However, energy consumption is highly variable and often non-stationary due to factors like holidays, weather changes, and policy shifts. This limitation may explain why ARIMA's performance was not as strong as that of Random Forest.

9.2 Discussion

The results suggest several key points about the effectiveness of machine learning models in energy consumption forecasting:

1. **Random Forest's Superiority**: The Random Forest model outperformed both SVM and ARIMA in all evaluated metrics. This is likely due to its ability to handle complex, nonlinear relationships and capture feature interactions better than SVM, which may struggle with high-dimensional data, or ARIMA, which relies on the assumption of stationarity.
2. **Support Vector Machines**: While SVM showed decent performance with an R^2 of 0.65, it performed worse than Random Forest. The linearity assumption of SVM might limit its ability to model the complex, non-linear patterns present in energy consumption data.
3. **ARIMA's Limitations**: Although ARIMA is a well-established method for time series forecasting, it did not perform as well in this context. This could be due to the time series data's complexity and non-stationarity, which ARIMA struggles with when seasonal or external factors are not properly accounted for.
4. **Model Comparison**: When comparing the three models—**Support Vector Machines (SVM)**, **Random Forest (RF)**, and **ARIMA**—it becomes clear that **Random Forest** was the most effective in forecasting energy consumption, outperforming both SVM and ARIMA across all metrics. While ARIMA's traditional time series approach provides a good baseline for forecasting, it struggled to capture the complexities of energy consumption data, especially when external variables or non-linear relationships are involved. **SVM**, though effective in certain contexts, was less successful here due to its assumption of linearity, which did not align well with the non-linear dynamics of energy usage. **Random Forest's** superior performance is attributed to its ability to capture intricate interactions between features through multiple decision trees, making it more adaptable to the diverse and often

unpredictable nature of energy consumption patterns. Furthermore, its robustness against overfitting and ability to model non-linear relationships gives it a distinct edge over the other models. This makes Random Forest the most suitable choice for energy consumption forecasting, although it could still benefit from further tuning and the integration of external factors.

9.3 Suggestions

Based on the results and discussion, the following suggestions could improve the forecasting accuracy and overall performance of the models:

- **Incorporate External Variables**:

  Energy consumption is influenced by a variety of factors, including weather, GDP, population growth, and industrial activity. Incorporating these external variables into the models could help enhance prediction accuracy. For instance, adding weather data or economic indicators could provide a more holistic view of the factors driving energy consumption.

- **Hyperparameter Tuning and Ensemble Methods**:

  Hyperparameter tuning for Random Forest (e.g., adjusting the number of trees, max depth, etc.) and SVM (e.g., kernel type, regularization parameter) could help refine the models' performance. Additionally, hybrid models that combine the strengths of different algorithms, such as combining Random Forest with ARIMA or SVM, could provide more accurate predictions by leveraging both non-linear patterns and temporal dependencies.

- **Cross-Validation and Data Augmentation**:

  Using **cross-validation** would provide a more robust evaluation of model performance and prevent overfitting. Moreover, introducing **data augmentation techniques**, such as generating synthetic data based on trends or seasonal fluctuations, could help in training more robust models. This is particularly important for energy consumption data, where seasonal variations and external influences often cause fluctuations.

10.Conclusion

This project delves into India's escalating energy demand, driven by rapid urbanization, expanding industrialization, and a rising population, which collectively underscore the pressing need for sustainable energy management. The country's strong dependence on fossil fuels, despite ongoing efforts to expand renewable energy capacity, highlights the urgency for efficient forecasting tools and tailored strategies that support energy security and environmental sustainability. By leveraging advanced machine learning models—Support Vector Machines (SVM), Random Forest, and ARIMA—this study examines historical energy consumption data across various Indian states and sectors to analyze key factors influencing demand and generate accurate future projections. Notably, the Random Forest model emerged as the most effective, with a mean squared error (MSE) of 846.65, a mean absolute error (MAE) of 15.57, and an R-squared (R^2) value of 0.80, indicating higher accuracy and reliability over the ARIMA and SVM models. The model's robust performance was statistically validated through hypothesis testing, underscoring its resilience as a predictive tool. Practical projections for the next thirty days demonstrate how these insights could assist energy suppliers and policymakers in planning and resource allocation, reducing risks of shortages, and facilitating smoother integration of renewable energy sources. The study emphasizes the importance of advanced predictive modeling for effective energy management in India, presenting the Random Forest model as a crucial asset for addressing the challenges of demand forecasting and strategic planning. With these approaches, India is better equipped to align its economic growth with environmental sustainability, ensuring a balanced and secure energy landscape for the future.

11.Limitations & Future scope

11.1 Limitations

Inconsistent Historical Energy Data:

The quality and completeness of historical energy data across different Indian regions are inconsistent. This variance can negatively affect model accuracy, as gaps in the data or discrepancies in measurement can lead to biased or unreliable predictions. Addressing this may require a more structured and comprehensive data collection infrastructure, as well as data imputation techniques for missing values.

Challenges with Complex Non-Linear Trends:

While Random Forest is proficient at capturing many non-linear relationships in the data, it might face difficulties with highly complex or volatile trends that do not align with easily identifiable patterns. This limitation may be overcome by exploring more sophisticated models, such as neural networks or hybrid techniques that integrate machine learning with traditional statistical approaches.

Dependency on Historical Data:

Current models are primarily based on historical data, which limits their ability to account for sudden, unpredictable events such as economic shocks, policy changes, or natural disasters. These external factors can significantly affect energy consumption but are not incorporated into the model, limiting its predictive capacity in the face of such changes.

Exclusion of Renewable Energy Variability:

The models do not account for the variability introduced by renewable energy sources like wind and solar, which depend on weather conditions and seasonal changes. As renewable energy continues to grow in importance, integrating data such as solar radiation and wind speed would improve model predictions, especially in regions heavily dependent on renewables.

11.2 Future Scope

Real-Time Data Integration:

Integrating real-time data on energy consumption, updated on an hourly or daily basis, would enable the model to adapt dynamically to shifts in energy demand or supply. This would make forecasts more responsive and accurate, particularly in the event of rapid fluctuations or unexpected disruptions.

Incorporation of Renewable Energy Factors:

The integration of renewable energy variables such as weather data, solar radiation, and wind speed could significantly improve forecasting accuracy. This would help model the fluctuations in renewable energy generation and better predict future energy demand.

Sector-Specific Forecasting Models:

Developing tailored models for different sectors (residential, commercial, industrial) could provide more granular and accurate insights into energy consumption patterns. This would allow for more targeted strategies in energy management across diverse sectors, optimizing resources based on sector-specific consumption trends.

Hybrid and Ensemble Model Development:

Exploring hybrid approaches that combine traditional statistical methods (such as ARIMA) with machine learning models (like Random Forest or neural networks) could capture both linear and non-linear patterns more effectively, leading to improved forecasting accuracy.

Inclusion of External Economic and Policy Factors:

Incorporating variables like fuel prices, economic growth projections, and government policies would allow the models to better capture the impact of economic and policy shifts on energy consumption, increasing their relevance for long-term forecasting.

12.Sampple Code

```
#Random Forest Regressor

import pandas as pd

import numpy as np

from sklearn.ensemble import RandomForestRegressor

from sklearn.model_selection import train_test_split

from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score

import matplotlib.pyplot as plt


# Load dataset into pandas dataframe (Using same data for brevity)

df = pd.read_csv('C:/Users/Shree/Desktop/dataset_tk.csv')


# Convert 'Date' to datetime and set as index

df['Date'] = pd.to_datetime(df['Date'], format='%d-%m-%Y')

df.set_index('Date', inplace=True)


# Example using 'Punjab' for forecasting. We can apply it to others in a similar way

df_punjab = df[['Punjab']].copy()  # Use .copy() to avoid SettingWithCopyWarning
```

```
# Create lag features (previous days' values) to introduce temporal dependence
for lag in range(1, 8):  # Using the last 7 days as features
    df_punjab.loc[:, f'lag_{lag}'] = df_punjab['Punjab'].shift(lag)


# Drop missing values due to shifting
df_punjab = df_punjab.dropna()  # Assign the result back to df_punjab


# Split into features (X) and target (y)
X = df_punjab.drop('Punjab', axis=1)
y = df_punjab['Punjab']


# Split into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, shuffle=False)


# Initialize Random Forest Regressor
rf = RandomForestRegressor(n_estimators=100, random_state=42)


# Train the model
rf.fit(X_train, y_train)
```

```
# Make predictions on the test set

y_pred = rf.predict(X_test)



# Calculate model performance metrics

mse = mean_squared_error(y_test, y_pred)

mae = mean_absolute_error(y_test, y_pred)

r2 = r2_score(y_test, y_pred)

print(f'Mean Squared Error (MSE): {mse}')

print(f'Mean Absolute Error (MAE): {mae}')

print(f'R-squared (R²): {r2}')



# Plot Actual vs Predicted values

plt.figure(figsize=(10, 6))

plt.plot(y_test.index, y_test, label='Actual')

plt.plot(y_test.index, y_pred, label='Predicted', linestyle='--')

plt.title('Actual vs Predicted Punjab Values')

plt.xlabel('Date')

plt.ylabel('Value')

plt.legend()

plt.show()
```

```
# Forecast future values for the next 30 days

future_steps = 30

last_known_data = df_punjab.iloc[-1, 1:].values.reshape(1, -1)  # Last known lag features


forecast_dates = pd.date_range(start=df_punjab.index[-1] + pd.Timedelta(days=1), periods=future_steps)

forecast_values = []


for _ in range(future_steps):

    next_pred = rf.predict(last_known_data)

    forecast_values.append(next_pred[0])


    # Update lag features for the next prediction

    last_known_data = np.roll(last_known_data, -1)  # Shift lag features

    last_known_data[0, -1] = next_pred[0]  # Add new prediction to the last lag


# Create a DataFrame to hold the forecasted results

forecast_df = pd.DataFrame(forecast_values, index=forecast_dates, columns=['Forecast'])
```

```
# Plot forecasted values for the next 30 days

plt.figure(figsize=(10, 6))

plt.plot(df_punjab.index, df_punjab['Punjab'], label='Historical Data')

plt.plot(forecast_df.index, forecast_df['Forecast'], label='Forecast', linestyle='--',
color='red')

plt.title('1-Month Forecast for Punjab')

plt.xlabel('Date')

plt.ylabel('Value')

plt.legend()

plt.show()


# Display forecasted values

print(f"Forecast for the next {future_steps} days:")

print(forecast_df)
```

13.References

[1] S. Sahu, "Trends and patterns of energy consumption in India," Indian Institute of Technology Bombay, 2008. [Online]. Available: https://mpra.ub.uni-muenchen.de/16753/

[2] J. Behera, "Energy consumption and economic growth in India: A reconciliation of disaggregate analysis," *Journal of Energy Technologies and Policy*, vol. 5, no. 6, 2015. [Online]. Available: https://www.doi.org/10.34218/IJM.11.12.2020.039

[3] A. Soni, A. Mittal, and M. Kapshe, "Energy intensity analysis of Indian manufacturing industries," *Elsevier B.V.*, 2017. [Online]. Available: https://doi.org/10.1108/IJESM-12-2023-0005

[4] S. Sahu and K. Narayanan, "Determinants of energy intensity in Indian manufacturing industries: A firm level analysis," Indian Institute of Technology Bombay, 2010. [Online]. Available: https://mpra.ub.uni-muenchen.de/id/eprint/21646

[5] S. Sahu and K. Narayanan, "Determinants of energy intensity: A preliminary investigation of Indian manufacturing," Indian Institute of Technology Bombay, 2009. [Online]. Available: https://mpra.ub.uni-muenchen.de/21646/

[6] S. Sahu and K. Narayanan, "Carbon dioxide emissions from Indian manufacturing industries: Role of energy and technology intensity," 2014. [Online]. Available: https://doi.org/10.1007/s43621-024-00306-2

[7] P. R. Jena, "A study of changing patterns of energy consumption and energy efficiency in the Indian manufacturing sector," Leibniz University of Hannover, 2011. [Online]. Available: https://mpra.ub.uni-muenchen.de/31195/

[8] A. Tandon and S. Ahmed, "Technological change and energy consumption in India: A decomposition analysis," *Innovation and Development*, 2015. [Online]. Available: http://dx.doi.org/10.1080/2157930X.2015.1114565

[9] A. K. Tiwari, "A structural VAR analysis of renewable energy consumption, real GDP, and CO2 emissions: Evidence from India," *Economics Bulletin*, vol. 31, no. 2, pp. 1793-1806, 2011. [Online]. Available: https://doi.org/10.25115/eea.v31i1.3267

[10] F. V. Bekun, "Mitigating emissions in India: Accounting for the role of real income, renewable energy consumption, and investment in energy," *International Journal of Energy Economics and Policy*, vol. 12, no. 1, pp. 188-192, 2022. [Online]. Available: https://doi.org/10.32479/ijeep.12652

[11] K. K. Pandeya and H. Rastogi, "Effect of energy consumption & economic growth on environmental degradation in India: A time series modelling," in *ICAE2018 – The 10th International Conference on Applied Energy*, 2019, pp. 1-804. [Online]. Available: https://doi.org/10.1016/j.egypro.2019.01.804

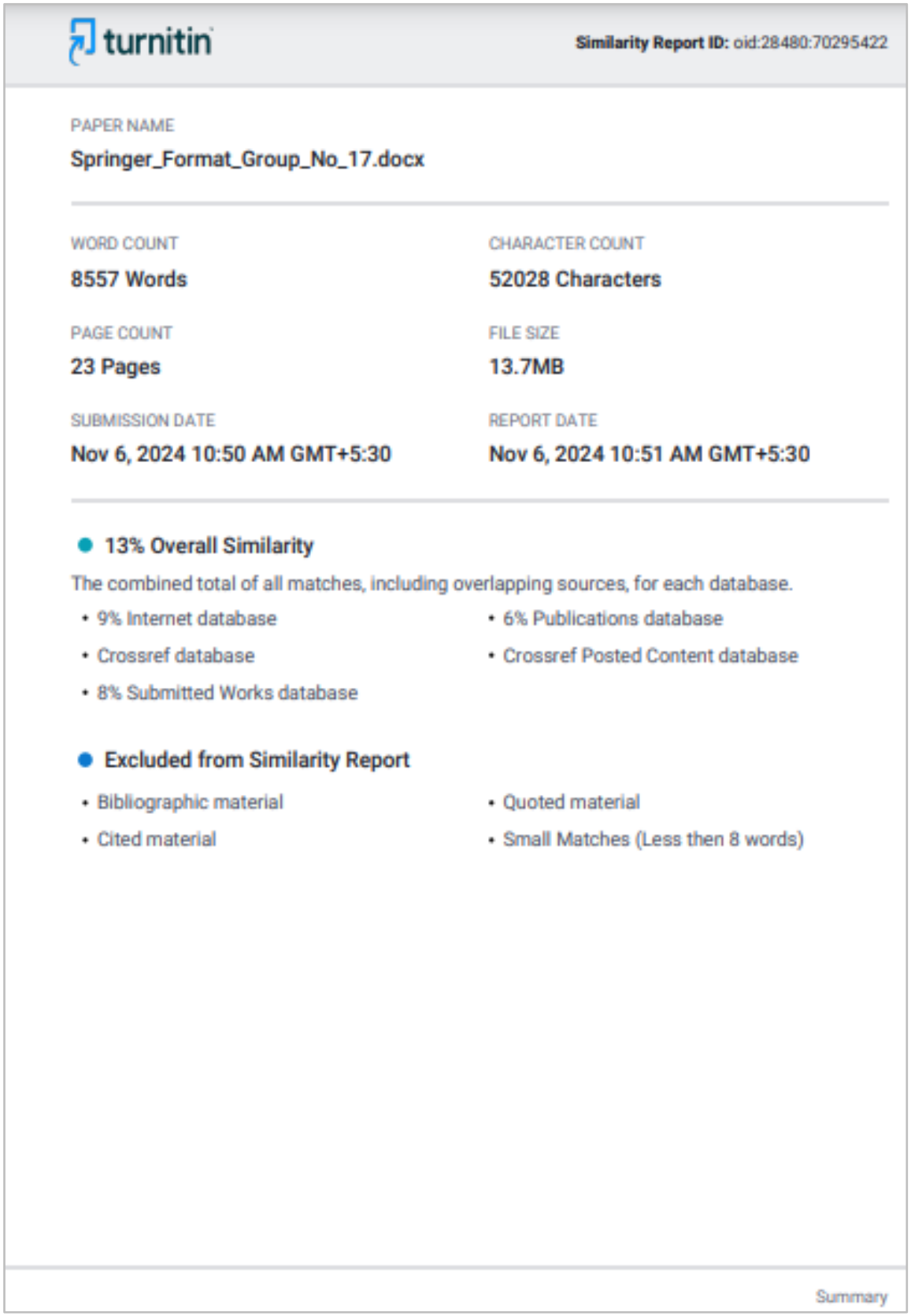
[12] A. K. Tiwari, "Primary energy consumption, CO2 emissions, and economic growth: Evidence from India," *Economics & Management*, vol. 15, no. 1, pp. 47-64, 2012. [Online]. Available: https://doi.org/10.2478/v10033-011-0019-6

[13] K. V. S. R. Murthy and M. R. Raju, "Analysis on electrical energy consumption of the agricultural sector in the Indian context," *ARPN Journal of Engineering and Applied Sciences*, vol. 13, no. 6, pp. 2265-2273, 2018. [Online]. Available: https://doi.org/10.29121/ijetmr.v7.i7.2020.723

[14] M. K. Mahalik and H. Mallick, "Energy consumption, economic growth, and financial development: Exploring the empirical linkages for India," *The Journal of Developing Areas*, vol. 48, no. 4, pp. 139-159, 2014. [Online]. Available: https://doi.org/10.1353/jda.2014.0063

[15] S. K. Yawale, T. Hanaoka, and M. Kapshe, "Development of energy balance table for rural and urban households and evaluation of energy consumption in Indian states," *Energy Procedia*, vol. 79, pp. 1-8, 2014. [Online]. Available: https://doi.org/10.1016/j.rser.2020.110392

14.Plagarism Report



turnitin

Similarity Report ID: oid:28480:70295422

PAPER NAME

Springer_Format_Group_No_17.docx

WORD COUNT	CHARACTER COUNT
8557 Words	**52028 Characters**
PAGE COUNT	FILE SIZE
23 Pages	**13.7MB**
SUBMISSION DATE	REPORT DATE
Nov 6, 2024 10:50 AM GMT+5:30	**Nov 6, 2024 10:51 AM GMT+5:30**

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 9% Internet database
- 6% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 8 words)

Summary