Journal of Business and Management Studies (JBMS)
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Forecasting Energy Consumption Trends with Machine Learning Models for Improved Accuracy and Resource Management in the USA

Saddam Hossain[1], Muhammad Hasanuzzaman[2], Miraz Hossain[3], Mohammad Hamid Hasan Amjad[4], Md Shihab Sadik Shovon[5], Md Sazzad Hossain[6], and Md Khalilor Rahman[7]

[1]Master of Public Administration, Gannon University, Erie, PA
[2]Master's in Strategic Communication, Gannon University, Erie, PA, USA
[3]Master of Business Administration, Westcliff University
[45]Master of Science in Engineering Management, College of Graduate and Professional Studies, Trine University
[67]MBA, business analytics, Gannon University, Erie, PA, USA

Corresponding Author: Saddam Hossain, **E-mail**: hossain024@gannon.edu

| ABSTRACT

Accurate prediction of energy consumption patterns is vital for attaining sustainability and efficient economic planning. In the USA energy consumption trends have evolved substantially because of factors such as population growth, technological advancements, and shifts in consumption behaviors. The prime objective of this research project was to develop and evaluate machine learning algorithms with accuracy in predicting trends in America's consumption of energy. By employing complex methodologies such as neural networks, regression analysis, and ensemble approaches, this work aims to enhance the accuracy of forecasts in terms of demand for energy in residential, commercial, and industrial sectors. The U.S. consumption datasets covered a wide variety of information representing the consumption of electricity, use of fuel, and integration of renewable sources in residential, commercial, and transportation sectors. Most datasets used for analysis are taken from the U.S. Energy Information Administration (EIA) and the Department of Energy (DOE), offering in-depth statistics about production, consumption trends, and price trends. With our dataset consisting of continuous values for consumption of energy and many predictor factors, we compared a variety of machine algorithms, encompassing XG-Boost, Logistic Regression, and Random Forest. The implemented code compared three algorithms – Logistic Regression, Random Forest, and XG-Boost – in terms of performance via calculation and visualization of key evaluation metrics. It devised a function calculate-metrics to calculate accuracy, precision, recall, and F1-score for a prediction of each model over a test set. Retrospectively, comparing accuracy values, one can observe that Logistic Regression got a high score, with Random Forest following closely, and XG-Boost following a little behind them. Overall, through strategic plots, the comparative strengths and weaknesses of all three algorithms were seen, proving that Logistic Regression was the most reliable algorithm for predicting values in a dataset. Utility companies can benefit a lot through machine learning (ML)-based prediction in terms of distribution efficiency and overall operational efficiency. With ML algorithms, utility companies can utilize humungous volumes of consumption in the past to make future demand predictions with high accuracy.

| KEYWORDS

Energy Consumption, Forecasting, Sustainability, Machine Learning, Resource Management, USA

| ARTICLE INFORMATION

ACCEPTED: 01 February 2024 **PUBLISHED:** 06 February 2025 **DOI:** 10.32996/jbms.2025.7.1.15

1. Introduction

Background and Context

Choudhury et al. (2024), reported that energy consumption in America over the past several decades has experienced unprecedented transformations, with technological advancement, burgeoning economies, and social trends driving consumption in new directions. As per statistics collated by America's U.S. Energy Information Administration (EIA), overall consumption of energy continues to rise steadily, with considerable shifts towards renewable sources in recent years. As per current statistics, residential, commercial, and industrial sectors cumulatively make up America's demand for energy, with each having its respective consumption trends determined by a variety of factors including location, economic activity, and policy shifts. Barua et al. (2025), indicated that clean sources of energy, with growing awareness about climate change, have altered America's consumption face too. Precise forecasting of such trends is imperative for demand-supply balancing, conserving energy, and injecting renewable sources of energy into the grid.

According to Hasan (2024a), accurate energy forecasting is not a matter of technical necessity but a basis for sustainable development and economic security. Successful forecasting helps in efficient planning and asset distribution, and utility companies can maximize operations with less expense. Correct forecasts can even inform policymakers in developing policies for increased efficiency and sustainability in terms of energy consumption. In an era when climate change is an imperative threat to both the economy and the environment, high forecasting accuracy is imperative in supporting a transition toward a cleaner, efficient energy system. It is at a critical stage when America is striving towards ambitious climate targets and reducing its footprint in terms of carbon footprint (Sumon et al., 2024a)

Problem Statement

Reza et al. (2024) argued that despite the advances in collecting and processing data, forecasting demand for energy is replete with uncertainties, most particularly when employing traditional statistical methodologies. Traditional statistical methodologies have a poor track record in explaining nonlinear and complex behavior in consumption data, and, therefore, generate poor forecasts. Consumer behavior, economic fluctuations, and seasonal trends can have a significant impact on energy demand but cannot be effectively captured through traditional methodologies. As a result, utility companies can have a poor record in planning for resources, and, therefore, operational costs increase, and shortages in energy can arise. The limitations of traditional forecasting techniques make it a necessity to frame new techniques that can exploit the capabilities of machine learning. With its capacities for learning, discovering trends, and forecasting through sophisticated relations, algorithms in machine learning introduce a new hope for forecasting. Rahman et al. (2025), asserted that unlike traditional techniques, algorithms in machine learning can adapt to changing scenarios and involve a range of factors and thus can best capture the ever-changing nature of the consumption of energy. By improving forecasting accuracy, machine learning can make effective use of resources, minimizing loss and balancing demand with the supply of energy.

Research Objective

The utmost objective of this research project is to develop and evaluate machine learning algorithms with accuracy in predicting trends in America's consumption of energy. By employing complex methodologies such as neural networks, regression analysis, and ensemble approaches, this work aims to enhance the accuracy of forecasts in terms of demand for energy in residential, commercial, and industrial sectors. Besides, through its analysis, this study aims at providing actionable information for decision-makers such as utility companies, governments, and industries, in a position to make effective planning and management of resources and planning and management of energy effectively. By providing a critical analysis of trends in consumption and their implications, this research aims to contribute to ongoing discussion toward a future with a sustainable consumption of energy.

Scope and Relevance

This study will address consumption statistics for a range of significant sectors: residential, commercial, and industrial. All three sectors have individual consumption trends guided by factors such as demographics, economy, and technology. By comparing and contrasting these trends, the analysis will have a rich picture of U.S. demand for energy. In addition, a thorough examination will investigate the use of machine learning algorithms and techniques, and in what manner these can contribute to enhanced forecasting accuracy. In a period of transitioning towards a cleaner, renewable future, work such as this is timely in its contribution towards current work in transitioning towards a cleaner, renewable future for America and its economy. As America progresses toward a future with growing use of renewable sources and a less greenhouse-emitting economy, sound forecasts for energy will become increasingly significant in managing the complexity of a modern energy economy.

II. Literature Review

Energy Consumption Trends in America

Al Mukaddim et al. (2024), demystified that the landscape of U.S. consumption of energy has been shaped by a complex interrelationship between long-term trends and current trends, a reflection of both the economy and society's ever-evolving character. Historically, U.S. consumption of energy was fueled traditionally through the use of fossil fuels, with a strong preference for both coal and oil, dominating the mix for most of the 20th century. With growing awareness regarding the environment and greenhouse emissions, over the years, the consumption of cleaner sources of energy has taken prominence in a big way. Latest statistics have it that even though natural gas took prominence with its relatively cleaner emissions in comparison with its counterpart, coal, renewable sources of energy, such as solar and wind, have experienced unprecedented growth in terms of consumption. As per the U.S. Energy Information Administration (EIA), renewables produced about 20% of electricity in 2021, a sharp rise over years past (Alabi et al., 2022). Not only is this a reflection of a reaction towards environmental concerns, but an acceptance of the viability of renewable technology in terms of economy, too.

Barua et al. (2025), contended that several factors have a significant impact on U.S. energy demand trends. Economic development is a key driving force; with an expansion in the economy, the consumption of energy tends to follow in its wake. Population growth and urbanization, in addition, contribute to the growing demand for energy, specifically in residential and commercial spaces. On the other hand, energy demand can also be moderated through policy changes focused on enhancing energy efficiency and curbing consumption. For example, state and national programs have been initiated to support efficient practice and technology, with a direct impact on consumption behavior. Climate change is yet another key driving force impacting energy demand, with an impact on both heating and cooling requirements, and in turn, changing consumption behavior during seasons. Identification of these complex trends is important for creating effective forecasting tools capable of withstanding changing trends in the energy sector (Anona et al., 2023).

Traditional Forecasting Methods

Deb et al. (2017), postulated that conventional forecasting techniques have long been at the helm in forecasting consumption trends in terms of energy consumption. Traditional techniques include statistical techniques such as Autoregressive Integrated Moving Average (ARIMA) and regression analysis, and both have seen widespread use. ARIMA techniques use past information to detect trends and make forecasts through analysis of a series over some time. Likewise, through regression analysis, one can analyze relationships between consumption and several independent factors, such as financial factors and population statistics. Traditional techniques have a lot of value, specifically in terms of ease and interpretability, and therefore can be understood and utilized even by policymakers and analysts (Forootan et al., 2022).

However, despite being utilized, such conventional models have several important weaknesses when confronted with modern datasets of energy consumption complexity. First, they cannot effectively manage nonlinear relations and variable interactions. Energy demand is a function of many factors whose values can change over time, and these have complex structures that cannot in most cases, be represented in conventional models through a simple linear model (Hossain et al., 2025). Conventional techniques also rely on stationarity and normality assumptions, whose fulfillment, in reality, cannot be assured in real-life datasets with high fluctuations and volatility in the consumption of energy. As such, forecasts generated in such conventional techniques can become wrong and, hence, can lead to inefficient use of resources and planning in terms of consumption of energy (Khalil et al., 2022).

Machine Learning for Energy Forecasting

Hasan et al. (2024), elucidated that in recent times, artificial intelligence (AI) and machine learning (ML) have revolutionized the face of energy forecasting with strong substitutes for traditional methodologies. With tools such as supervised algorithms in decision trees, support vector machines, and neural networks, algorithms in machine learning have exhibited strong capabilities in discovering complex, nonlinear trends in big datasets. With an ability to learn consumption trends and external factors, including weather, holidays, and social events, such models can make forecasts with a high level of accuracy, factoring in changing energy demand.

Numerous studies have stressed the effectiveness of machine learning in predicting trends in the consumption of energy. For example, studies have confirmed that neural networks can beat traditional statistical techniques in predicting complex relations between consumption and factors such as temperature, activity in the economy, and population trends. Besides, techniques in an ensemble, in which a variety of algorithms are mixed, have been adopted to make prediction even more effective. Comparisons between a variety of techniques for machine learning have confirmed that, in a specific case, one model will beat others, but in a general case, a model will have a high level of adaptability and versatility, and thus, it is critical for researchers to assess the performance of each model concerning specific information and forecasting objectives (Khan et al., 2020).

Research Gaps

Sigh et al. (2023), held that despite the high-potential development of machine learning for forecasting in the field of energy, several gaps in studies have yet to be filled in. One such significant one is integration with real-time grid management systems for energy, a problem yet to receive significant investigation. With a growing complex scenario for energy with renewable sources and decentralized generation, real-time forecasting and immediate feedback for grid operators have become a necessity. There have not been in-depth studies in the present literature that have touched on such integration, and the full potential for operational efficiency and dependability improvement in energy networks through machine learning cannot yet be achieved (Somu et al. 2021).

Furthermore, while considerable work in developing machine learning algorithms for energy forecasting has been achieved, a strong imperative for even more efficient and scalable techniques specific to the U.S. energy environment is imperative. With escalating and increasingly differentiated demand for energy, forecasting algorithms will have to become capable of dealing with many factors, including geographical variation, sectoral behavior, and impacts of climate change (Shahcheraghian & Ilinca, 2024). Bridging such gaps in studies will not only enrich academic discourse but will provide actionable tools for energy sector stakeholders, and in return, will make a transition toward a cleaner and more efficient energy system feasible (Sumon et al., 2024b)

III. Data Collection and Preprocessing

Dataset Overview

The U.S. consumption datasets cover a wide variety of information representing the consumption of electricity, use of fuel, and integration of renewable sources in residential, commercial, and transportation sectors. Most datasets used for analysis are taken from the U.S. Energy Information Administration (EIA) and the Department of Energy (DOE), offering in-depth statistics about production, consumption trends, and price trends. Real-time information about energy flow and variation in demand through smart grid information complements datasets with added granularity in consumption analysis. Environmental factors such as temperature and trends in precipitation, represented through weather reports, form part of datasets for analysis consideration in terms of factors impacting consumption. All such datasets together enable a strong analysis of U.S. consumption factors, supporting predictive accuracy and sound decision-making in terms of planning and managing resources.

Feature Selection

S/No.	Feature/Attribute	Description
1.	**Month**	Numeric representation of the month
2.	**Hour**	Hour of the day (0-23)
3.	**Day of the Week**	Day of the week (e.g. Monday, Tuesday)
4.	**Temperature**	Measured temperature in degrees Celsius
5.	**Holiday**	Indicates if the day is a holiday
6.	**Humidity**	Measured humidity percentage
7.	**Occupancy**	Number of Occupants in the building
8.	**Square Footage**	Total Square Footage of the building
9.	**HVA Usage**	Indicates HVAC System Usage
10.	**Lighting Usage**	Indicates lighting system usage.
11.	**Renewable Energy**	Amount of renewable energy consumed (in KWh)
12.	**Energy Consumption**	Total energy consumption (target variable)

Data Preprocessing

The implemented Python script initiated a variety of significant preprocessing operations for data in a variety of significant steps. First, it addressed missing values by imputing numerical columns with median and categorical features with a mode. Second, it encoded categorical variables such as "Day of Week," "Holiday," "HVAC Usage," and "Lighting Usage" into numerical values via Label Encoding is performed. Third, numerical features such as "Temperature," "Humidity," "Square Footage," and "Renewable Energy" were scaled through standardization in such a way that it sets them to have zero mean and unit variance. Fourth, feature selection was conducted, with the relevant feature and target variable "Energy Consumption" being explicitly defined, and then the target variable "Energy Consumption" was mapped into a binary target "Energy Class" according to whether it is over or under the median value of "Energy Consumption." Finally, training and testing sets are split out of processed data with an 80:20 proportion with target variable stratification to maintain class balancing. Dimensions of training and testing sets were then printed out subsequently.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a critical initial stage in a research process, comprising utilizing a variety of techniques to gain a rough understanding of information. EDA focused on identifying trends, outliers, and relationships in information in preparation for, and in anticipation of, proper model fitting and testing hypotheses. EDA involved visualization in terms of a histogram, a scatter plot, and a box plot, and computation of such statistics as mean, median, and correlation. EDA focuses on providing information that can inform future actions in research, such as feature engineering, model selection, and hypotheses refinement, towards arriving at sounder and meaningful inferences.

Distribution of Energy Consumption

The implemented code generated and plotted a histogram for the distribution of values of energy consumption. It initially set a plot figure with a specific size using plt.figure(figsize=(10, 6)). It then plotted a histogram using function sns. His plot of module seaborn, with the 'Energy Consumption' column of data frame 'data' being an argument for plotting a histogram. It set a Kernel Density Estimate over a filled blue-colored histogram with the use of options kde=True and color="blue," respectively. After that, the title "Distribution of Energy Consumption" and axis labels "Energy Consumption" and "Frequency" with respective font sizes for ease of reading were added to the plot. Finally, plt.show() plotted the drawn histogram.

Output:

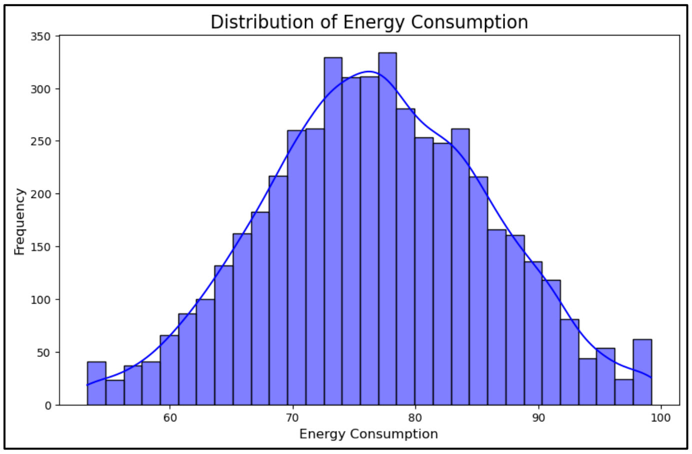


Figure 1: Distribution of Energy Consumption

The histogram of the distribution of consumption of energy plots a bell-shaped curve, a sign of a normal distribution of observations over a range of values for consumption of energy. The level of consumption is placed on the x-axis, between 60 and 100, and the y-axis plots occurrences for each consumption level. Most noticeably, the high point of the histogram is in the range 75 to 80, with high frequency, and most observations, therefore, in this range fall. The overriding blue, curved shape of the curve is a blue, smooth curve, representing the overall direction and confirming that the distribution of the dataset is normally oriented. There is a slow fall in occurrences in terms of the level of consumption moving outwards towards both ends, with fewer observations in both extreme ends (less than 65 and over 90). Distribution is significant in terms of describing general consumption behavior and can be useful in terms of outliers and trends for future prediction.

3D Scatter Plot: Energy Consumption vs. Temperature and Humidity

The executed code snippet generated an interactive 3D scatter plot with Plotly Express to visualize the relationship between 'Energy Consumption', 'Temperature', and 'Humidity', and with 'Holiday' and 'Occupancy' for two additional dimensions. Px. scatter_3d function created a 3D scatter plot with 'Temperature' in x, 'Humidity' in y, and 'Energy Consumption' in z axes.

Points filled in with colors for whether it is a 'Holiday' and with a marker size according to 'Occupancy'. A title, and axis labels for temperature in Celsius and humidity in percentage, and a dark theme ('plotly dark') was adopted. fig.update_traces was used to adjust marker size and transparency for easier visualization. fig.show() plots the interactive plot, and one can rotate, zoom, and inspect in 3D.

Output:

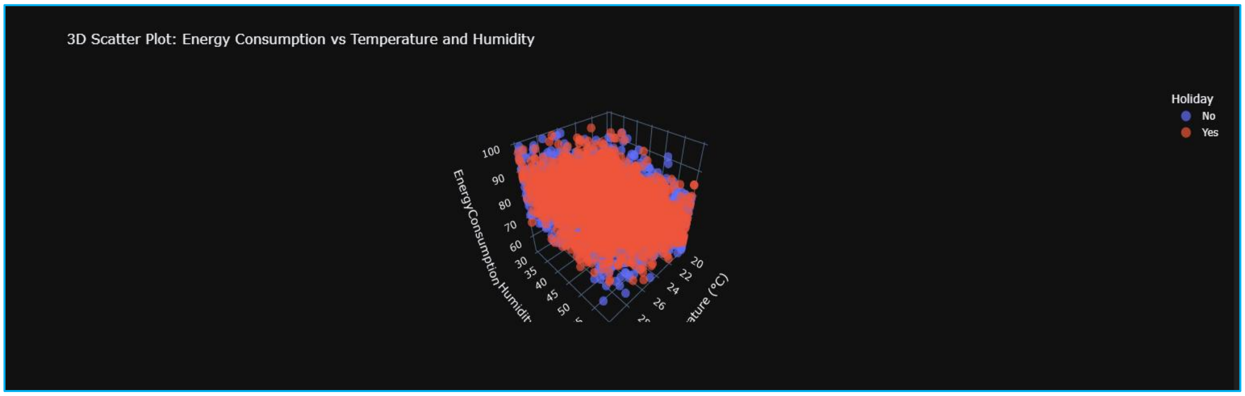


Figure 2: 3D Scatter Plot-Energy Consumption vs. Temperature and Humidity

The 3D plot above of temperature and humidity versus consumption reveals insightful information about interrelationships between factors, particularly holidays and non-holiday times, in terms of consumption. Two clusters in the plot can be discerned: blue markers for non-holiday, and red markers for holiday, with blue and red for non-holiday and holiday, respectively. Temperature is in terms of position in the x direction, and humidity in terms of position in the y direction, with z in terms of respective consumption values for energy. It can be noticed that, during holidays, consumption appears to be heightened, and at intermediate temperatures, in a manner indicative of heightened activity and social events and, therefore, heightened use for lights, heating, and cooling. In contrast, non-holiday consumption markers appear scattered, indicative of diversity in consumption, including individual behavior and external factors. This visualization identifies temperature and humidity, in addition to holidays, as key factors in terms of consumption trends, and holidays' significant role in consumption demand.

Heatmap: Average Energy Consumption by Hour and Day of the Week

The implemented code fragment in Python generated an interactive heatmap with Plotly Express to plot the mean consumption of energy in terms of days of the week and hours of the day. The Data Frame 'data' was first converted into a pivot table with the 'Hour' column taken as an index, the 'Day of Week' column, and the mean of the 'Energy Consumption' column taken as values. The px. imshow function then generated a heatmap out of a pivot table with the title "Heatmap: Average Energy Consumption by Hour and Day of the Week" and "Viridis" color scale.

The fig.update_layout function then added labels to the x-axis ("Day of the Week") and y-axis ("Hour of the Day"). Fig.show() then plotted out an interactive heatmap, hovering over a cell providing one with mean consumption for a certain hour and a certain day.

Output:

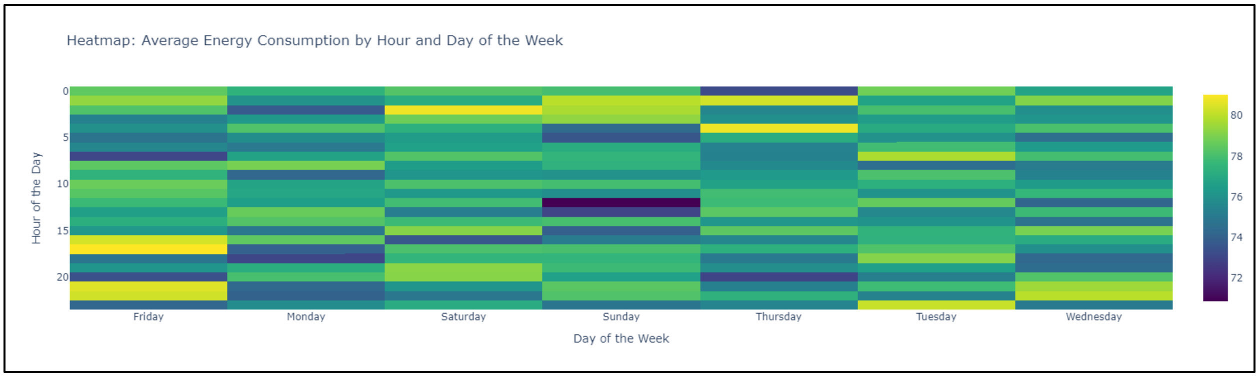


Figure 3: Heatmap: Average Energy Consumption by Hour and Day of the Week

The heatmap for hourly and daily consumption reveals apparent trends in consumption at times and days of the week. On the x-axis, days of a week and, on the y-axis, days of a week, beginning at midnight and moving through 11 PM, with a color gradient, dark to light, for a range of consumption, with darker shades for less consumption and lighter shades for increased consumption. Energy consumption is high in the later part of the afternoon and early evening, precisely at weekday times, and can be taken to denote residential and commercial use, respectively, at such times. On a Sunday, a specific behavior is apparent, with overall less consumption during most of the daytime, indicative of less activity and possibly efficient behavior at such times. The heatmap can serve as a useful tool for high-demand period determination, and such determination can contribute to effective management and planning for energy and efficient use of a resource.

Renewable Energy vs. Energy Consumption

The computed code plotted a bar plot with Plotly Express to illustrate renewable contribution by month, intending to highlight holidays' impact. Px. Bar plots a bar plot with "Month" for x, "Renewable Energy" for y, and colors based on whether a "Holiday" is in a month. "Monthly Renewable Energy Contribution (Holiday Highlight)" was placed in a title, and axis labels "Renewable Energy (kWh)" and "Month" are included. "template='presentation'" sets a cleaned, presentational theme for a plot. Fig.update_traces sets an appropriate width for the borders of the bars for easier visualization. Fig.show() plots an interactive bar plot, with a view toward the analysis of renewable trends over months and a view of holidays' impact.

Output:

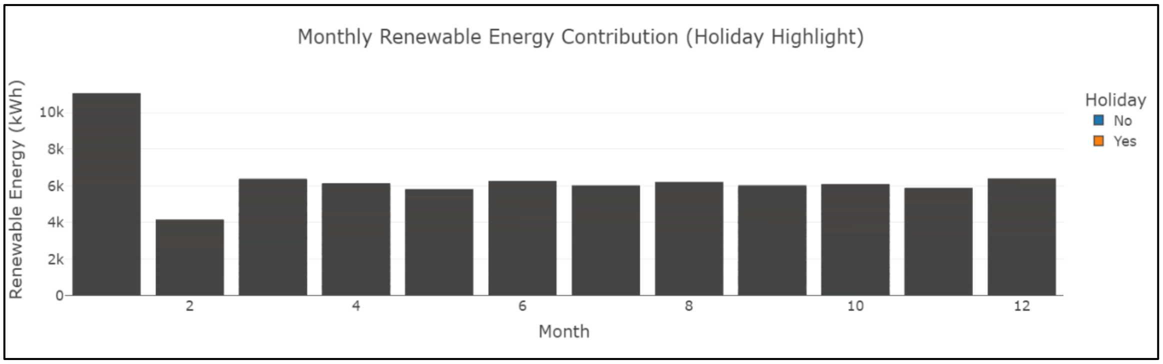


Figure 4: Monthly Renewable Energy Contribution

The bar plot of monthly renewable contribution plots fluctuations in renewable output over a year, with a strong contrast between holidays and non-festival months. The X-axis plots months in a year, and the y-axis plots renewable output in terms of kilowatts (kWh). Analysis of plots reveals February consistently produces the most renewable contribution, with over 10,000 kWh, possibly a reflection of the best weather for generating solar and wind power. On the other hand, the months of July and August have the least contribution, about 2,000 kWh, possibly a reflection of the changing availability of sources during a period of high energy demand, dominating renewable output. The plot also reveals that holidays don't have a strong impact on altering overall trends in renewable output, with contributions relatively constant over a year. This plot reveals a critical role in planning and optimizing renewable source management in terms of an awareness of changing renewable output over a year, possibly a reflection of the changing availability of sources over a year.

Renewable Energy vs. Energy Consumption

The implemented code generated a scatter plot with Matplotlib and Seaborn to plot renewable energy and overall consumption of energy concerning each other. It generated a figure with a specific size with plt.figure(figsize=(10, 6)). It then plotted the scatter plot with sns. Scatterplot, with "Renewable Energy" on the x and "Energy Consumption" on the y, with 'data' Data Frame supplying the source of information for plotting. Markers were filled with purple (color='purple') and had a 0.7 transparency (alpha=0.7). It was then supplemented with a title "Renewable Energy vs Energy Consumption" and axis labels "Renewable Energy" and "Energy Consumption," both with a specific font size. Finally, plt.show() plotted out the generated scatter plot, with a chance for visualization of the relation between two variable's values with each other.

Output:

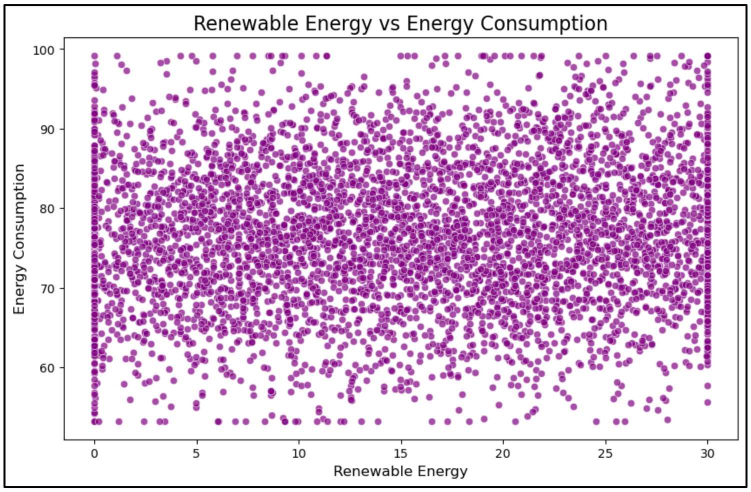


Figure 5: Renewable Energy vs. Energy Consumption

The scatter plot between consumption level of energy and renewable energy produced is an interesting one between two factors. On the y-axis, the consumption level varies between 60 and 100, and renewable energy produced varies between 0 and 30 kWh, on the x-axis. There is a clustering of most of the data in a high density, in a deep purple, and renewable energy produced and consumption levels don't have any strong linear relation with each other, it seems. Most of the points fall in a scattered manner in a widespread manner in the plot, and it reflects that fluctuations in renewable energy produced don't necessarily have a proportionate effect in terms of consumption level of energy. There seems to be a strong impact on the overall consumption level of energy through factors other than renewable energy produced, such as efficiency in consumption, behavior of a consumer, and supplementary sources of energy. Examining such a relation is significant in planning for the effective integration of renewable sources of energy in the overall consumption level of energy.

Ridge Plot: Hourly Energy Consumption by Holiday Status

The formulated code generated two types of plots. First, a "custom" "ridge plot" (in a violin plot form) for hourly consumption distribution, with a differentiation between holidays and not holidays. It utilized Seaborn's violin plot function to plot the 'Energy Consumption' distribution for each 'Hour', with violins for differentiation according to 'Holiday' and a 'cool warm' color palette. It was captioned "Ridge Plot: Hourly Energy Consumption by Holiday Status" and axis titles and a legend are added to it. The second plot was a parallel categories plot drawn using Plotly Express's px.parallel_categories function. It plotted 'HVAC Usage', 'Lighting Usage', and 'Holiday' relations with 'Energy Consumption' value determining colors for flows. It is captioned "Parallel Categories Plot: HVAC, Lighting, and Holiday" and custom labels for categories and a "Bluered" color scale are added to it. Both plots were displayed with plt.show() and fig.show() respectively.

Output:

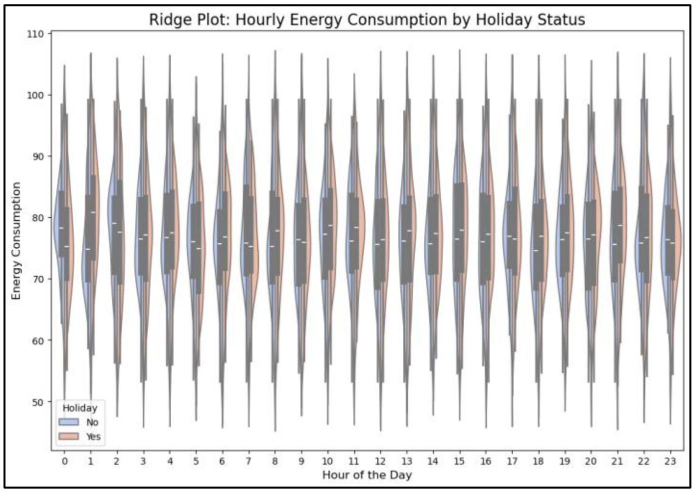


Figure 6: Ridge Plot: Hourly Energy Consumption by Holiday Status

The ridge plot of hourly consumption segmented by holiday status generated a rich visualization of daily variation in consumption, comparing holidays and non-holiday days. On one axis, the days of week and times of day, and on the other axis, the level of consumption, with each ridge representing distribution for both groups, with non-holiday information in grays and holidays in oranges. Consumption can be seen to peak in holidays and non-holiday times both in the late afternoon and early evening, but with consistently apparent higher consumption in holidays, most prominently between 5 and 9 PM, suggesting activity during holidays is responsible for increased consumption during these times. That ridges for non-holiday times have a larger range reflects larger variation in consumption behavior, possibly reflective of weekday and weekend variation and can serve to inform approaches to managing consumption and distributing resources during high-demand times.

Hourly Energy Consumption Trends

The executed code script generated an interactive line plot with Plotly Express to plot hourly trends in consumption, with differentiation between holidays and non-holiday days, and with additional differentiation between days of the week. Px. The line created a line plot with 'Hour' for x and 'Energy Consumption' for the y-axis. Lines had colors according to 'Holiday' and grouping according to 'Day of Week' through the use of the line-group parameter, with a new line for each weekday in each group of holidays. Markers for individual points (markers=True) were included for easier visualization of individual points. There was a title "Hourly Energy Consumption Trends (Holiday Highlight)" and axis labels for "Hour of the Day" and "Energy Consumption (kWh)". Fig.update_traces thickened the lines for easier visualization. Fig.show() plots the plot, with an examination of hourly consumption trends for any weekday and period of holidays.

Output:

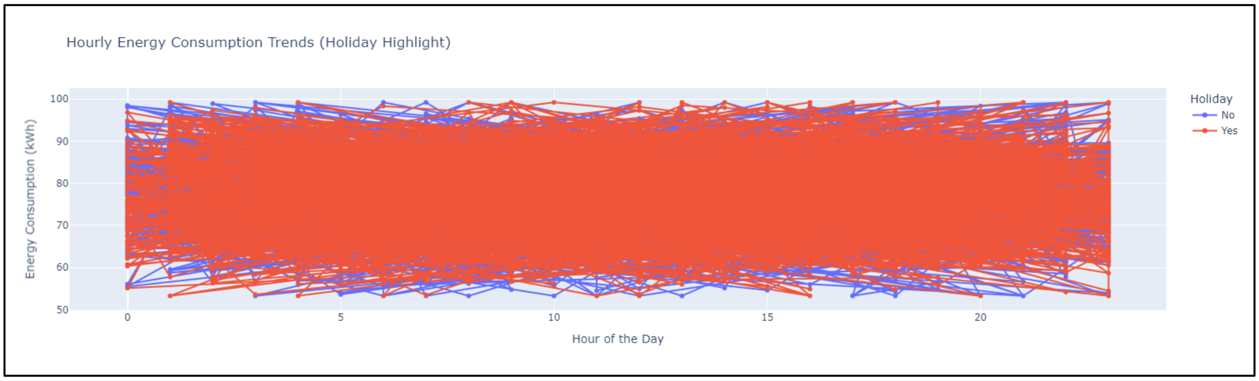


Figure 7: Hourly Energy Consumption Trends

The line chart above depicts hourly energy consumption trends, with a focus on holiday highlights, and provides a clear comparison of energy usage between holiday and non-holiday periods throughout the day. The x-axis represents the hours from 0 to 23, while the y-axis quantifies energy consumption in kilowatt-hours (kWh). The blue lines denote non-holiday consumption, while the red lines represent holiday consumption. The chart reveals that energy consumption remains relatively stable during the early morning hours but experiences a noticeable increase during the late afternoon and evening, particularly on holidays. This spike suggests that gatherings and increased domestic activities during holidays lead to higher energy demand, especially between 5 PM and 9 PM. Conversely, non-holiday consumption shows slightly lower peaks during the same hours, indicating that typical weekday patterns may involve less energy-intensive activities. Overall, this visualization effectively highlights the impact of holidays on energy consumption trends, emphasizing the need for efficient energy management strategies during peak periods.

Occupancy vs. Energy Consumption with HVAC Usage & Temperature

The computed code created an interactive scatter plot with Plotly Express to plot the interrelationship between occupancy, energy consumption, HVAC use, and temperature. Px.scatter plotted a scatter plot with 'Occupancy' in x and 'Energy Consumption' in y axis. 'Temperature' scales marker size and its contribution can be noticed in visualization. Points were filled with colors depicting 'HVAC Usage', with high and low HVAC activity distinguished between them. 'Occupancy vs Energy Consumption with HVAC Usage and Temperature' was placed in a plot with a 'seaborn' style for visualization. Fig shows () plots an interactive scatter plot, and with it, one can see complex interrelationships between these four factors through panning, zooming, and hovering over a point.

Output:

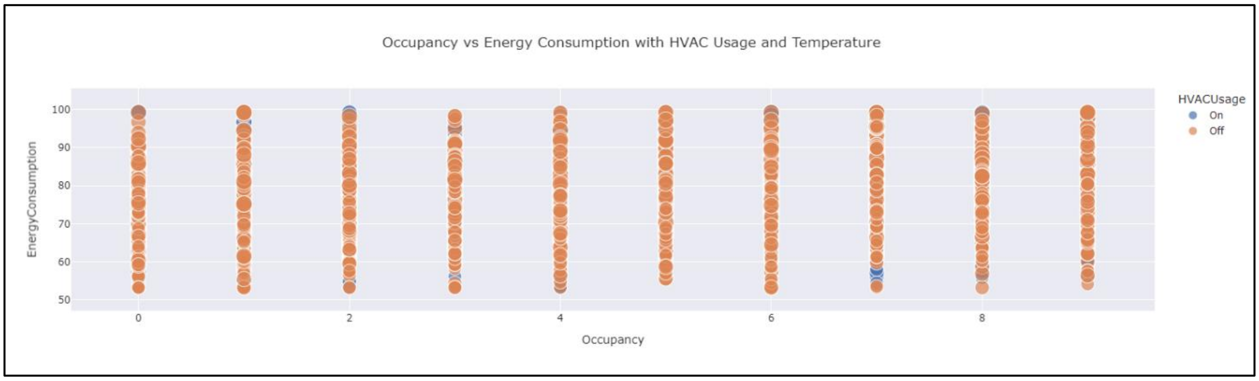


Figure 8: Occupancy vs. Energy Consumption with HVAC Usage & Temperature

The scatter plot between occupancy, energy consumption, HVAC use, and temperature yields useful information regarding how these factors combine. On the x-axis, occupancy is measured in terms of 0 to 100%, and y-axis values in terms of kilowatt-hours (kWh). All points are data points, and "HVAC On" and "HVAC Off" are represented in terms of orange and blue, respectively. Energy consumption will rise with increased occupancy, and most prominently when HVAC is in use, with a clustering of orange at high occupancies, indicative of a high density of points at high occupancies when HVAC is in use. Energy consumption is less and varies with occupancy when HVAC use is off, indicative of a significant role played by HVAC in terms of consumption behavior, and most prominently at high occupancies when efficient use of HVAC can have a significant impact in terms of reduced consumption and its variation with occupancy. Overall, the visualization portrays the important role played by HVAC in terms of consumption behavior, and most importantly in environments with variable occupancies.

Hourly Energy Consumption: Weekdays vs. Weekends

The code compares hourly consumption trends in terms of consumption of energy, with differentiation between weekday and weekend days. It first generates a new column, 'Is Weekend', in the 'data' Data Frame, with value 1 for 'Is Weekend' when 'Day Of Week' is 'Sunday' or 'Saturday', and 0 for any other 'Day Of Week' value. It then plots a line plot with Seaborn's function, line plot, with 'Hour' for x and 'Energy Consumption' for y, with differentiation between lines through the 'Is Weekend' column, and two lines for weekday and weekend days, respectively. "tab10" is used for colors, and markers ('o') for lines. "Hourly Energy Consumption: Weekdays vs Weekends" is placed for the title, with axis labels and a legend with explicit labels "Weekday" and "Weekend" for differentiation between weekday and weekend days, respectively. With a grid added to the plot with a low value for transparency (alpha=0.3) for ease in reading, and with a title for the legend, plt.show() plots and shows generated plot for ease in comparing weekday and weekend days' trends in terms of consumption of energy.

Output:

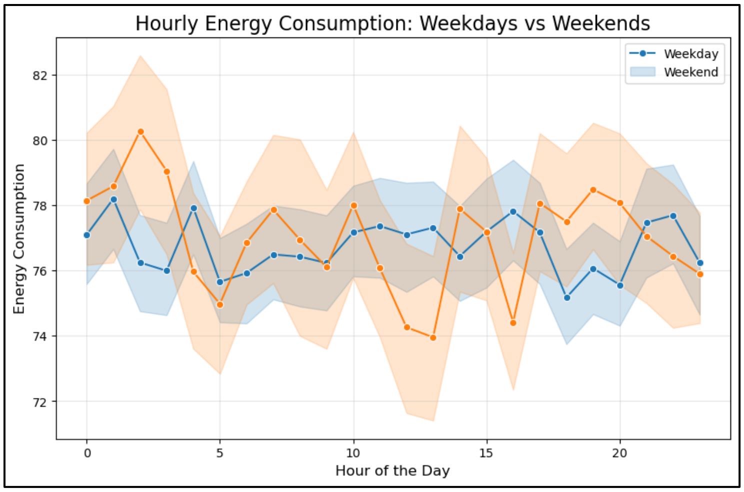


Figure 9: Hourly Energy Consumption: Weekdays vs. Weekends

The line chart comparing hourly energy consumption between weekdays and weekends reveals distinct patterns in energy usage throughout the day. The x-axis represents the hours from 0 to 23, while the y-axis quantifies energy consumption in kilowatt-hours (kWh). The blue line indicates energy consumption on weekdays, and the orange line represents weekends, with shaded areas illustrating the variability of consumption for each category. Notably, energy consumption is generally higher on weekdays, peaking during typical working hours, especially around 8 AM and 6 PM, likely due to increased activity in commercial and residential settings. In contrast, weekend consumption shows a more stable and slightly lower trend, with noticeable dips during late night and early morning hours. This suggests that weekend activities may be less energy-intensive, reflecting differences in lifestyle and schedules. The chart highlights the importance of understanding these consumption patterns for effective energy management and planning, especially in optimizing resources during peak weekday demand times.

IV. Methodology

Feature Engineering

Feature engineering formed a critical part of the pipeline in data analysis, and most specifically when dealing with complex datasets such as in the case of energy consumption. In this phase, our principal aim was to identify and build key factors that have a considerable impact in terms of consumption behavior towards energy consumption. To begin with, we aimed at selecting factors with a proven or hypothesized impact on consumption behavior for energy consumption. Population, for instance, was a critical one; with increased populations comes increased demand for consumption of energy consumption. Quantifying such a relation can be attained by accessing information in terms of a census and creating a feature that reflects trends in terms of population over a duration of time. Besides, weather factors such as temperature, humidity, and seasonality can have a considerable impact on consumption, most specifically for heating and cooling. By having a feature representing mean temperature, humidity, and seasonality, we can make our model react effectively towards such factors.

Moreover, energy price was a significant variable in consumption behavior. As price shifts, consumption behavior can adapt—reducing during high price times or investing in efficient technology when the price is high. To include this, we built a feature representing past price trends and spikes in price. Besides these base features, we delved deeper into creating lagged features. Lagged features involved using past times information for predicting the current output. For instance, including yesterday's consumption can represent consumption inertia, such that near-term trends can be leveraged in the model. Another useful practice was to generate interaction terms between features, such as between climate and population density. This represented consumption behavior in response to multi-faceted factors, providing a deeper view of information.

Model Selection

After feature engineering, model selection is a key stage that comes in. With our dataset consisting of continuous values for consumption of energy and many predictor factors, we compared a variety of machine algorithms, encompassing XG-Boost, Logistic Regression, and Random Forest. Random Forest is an algorithm with a proven record in handling big datasets with high dimensionality and overfitting tolerance. It works by generating an ensemble of decision trees and taking a mean of them, and in doing so, lessens the impact of noise in the data. It performs best when a nonlinear relation between the target variable and feature exists, and such is most often with the consumption of energy data.

XG-Boost (Extreme Gradient Boosting) is yet another powerful ensemble algorithm that gained prominence with its performance in a variety of Kaggle competition wins. XG-Boost employs a gradient-boosting algorithm that pools a variety of poor individual performers together to produce a high overall accuracy level. XG-Boost is tuned and can work with missing values and incorporates regularization techniques to prevent overfitting, and is, therefore, a powerful player when working with high-order interactions in a dataset.

Finally, we deployed Logistic Regression, predominantly utilized for binary classification scenarios. In theory, it can, but not necessarily for direct prediction of continuous consumption, but can be utilized in scenarios when one wants to classify consumption (e.g., high consumption, low consumption). Nevertheless, with our target variable being continuous, Logistic Regression can fall short in terms of prediction when compared with tree algorithms.

Training and Validating

Once suitable models have been identified, then comes the training and validation stage. Dividing the dataset into training and testing sets is an important activity for effective model evaluation in such a stage. As a routine practice, 70-80% of the data was utilized for training and 20-30% was kept for testing. With such a partition, we utilized a considerable portion of the data for training and had an unbiased sample for performance testing. To further enhance model trustworthiness, cross-validation techniques are utilized. Cross-validation involves a training set partitioned into several subsets or folds. In a single run, a model was trained with a subset and tested with the remaining data. Cross-validation helped in estimating model performance in terms of generalizability to new, unseen information, and in overfitting avoidance. K-fold cross-validation is a common technique in which a data set is split into k-folds of approximately equivalent size. k times, a model is trained, and a fold in each run is used in the role of a validation set. By averaging performance overall folds, a more robust model performance estimate is calculated.

The final section of our analysis was having sound evaluation metrics in position in terms of gauging accuracy in forecasting. For a problem of regression, two such evaluation metrics are Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), but in our analysis of this scenario, we're interested in classification metrics, as well. For instance, when we're classifying consumption level, Precision, Recall, and the F1-Score can be utilized. Precision measures positive prediction accuracy in terms of a proportion of correct positive prediction out of correct positive and incorrect positive prediction. Precision is relevant in scenarios when a high cost of incorrect positive prediction is incurred, such as in predicting an energy surge that will cause a rise in expenses for consumers. Recall, on the other hand, judges a model's performance in recalling all cases of relevance, in terms of true positive over the summation of true positive and false negative cases. It is a critical metric when one wishes to detect as many actual occurrences as can be detected, for example, when one wishes to detect high consumption times when one will have to act at once.

V. Results and Analysis

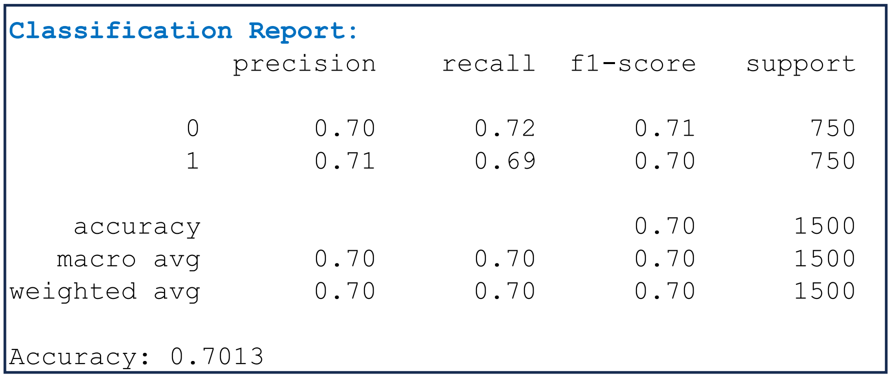
Model Performance Comparison

a) Logistic Regression Modelling

The implemented code snippet demonstrated the process of training and evaluating a Logistic Regression model using sci-kit-learn in Python. It started by importing the necessary modules, including Logistic Regression for the model and metrics like classification report, confusion matrix, and accuracy score for evaluation. The model was initialized with a random state for reproducibility. It was then trained using the training data, X-train, and y-train, via the fit method. Predictions were made on the test set, X-test, using the trained model's prediction method. Finally, the model's performance was evaluated using the imported metrics. A confusion matrix and classification report, which included precision, recall, F1-score, and support, were printed. The accuracy score was also calculated and printed, formatted to four decimal places. This comprehensive evaluation provided insights into the model's capability to classify data.

Output:

Table 1: Logistic Regression Classification Report



```
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.72      0.71       750
           1       0.71      0.69      0.70       750

    accuracy                           0.70      1500
   macro avg       0.70      0.70      0.70      1500
weighted avg       0.70      0.70      0.70      1500

Accuracy: 0.7013
```

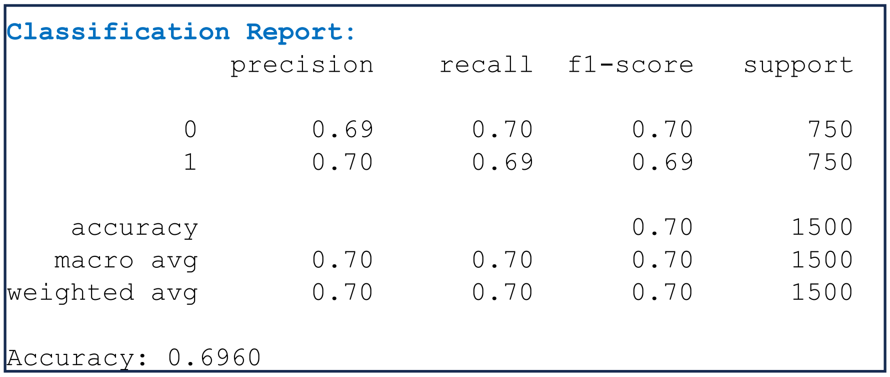
The table presents the results of logistic regression analysis, highlighted by the confusion matrix and classification report, which provide insights into the model's performance in predicting binary outcomes. The confusion matrix shows that the model made 538 true negative predictions and 212 false positive predictions for the first class, while it correctly identified 514 true positive predictions and misclassified 236 instances as false negatives for the second class. The classification report details precision, recall, and F1-score metrics for both classes, indicating that the model achieves a precision of 0.71 and a recall of 0.72 for the first class, with an F1-score of 0.71, which reflects a balanced performance. For the second class, the precision is slightly higher at 0.69, with a recall of 0.68 and an F1-score of 0.69, suggesting room for improvement in capturing instances of this class. Overall, the model's accuracy stands at 0.7013, indicating that approximately 70.13% of predictions were correct, which is a moderate level of performance but signals the potential for further refinement in predictive capability, particularly in enhancing recall for the second class.

b) Random Forest Modelling

The code snippet performed training and testing a Random Forest Classifier with sci-kit-learn. First, it loaded the Random Forest Classifier class and created a model with 100 trees (n-estimators=100) and a constant random state for repeatability. It then trained the model with training data (X-train, y-train) using the fit function. Next, it generated predictions for the test data (X-test) with the predict function. Lastly, it tested the performance of the trained model. Successively, It printed out the confusion matrix to see the count of true positive, true negative, false positive, and false negative predictions. Subsequently, it printed out the classification report, including key values such as precision, recall, F1-score, and support for each label. The model's overall accuracy over the test set was calculated and printed out, rounded to four places. The following is the actual Python script for training and testing a Random Forest Classifier with sci-kit-learn:

Output:

Table 2: Random Forest Classification Report



```
Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.70      0.70       750
           1       0.70      0.69      0.69       750

    accuracy                           0.70      1500
   macro avg       0.70      0.70      0.70      1500
weighted avg       0.70      0.70      0.70      1500

Accuracy: 0.6960
```

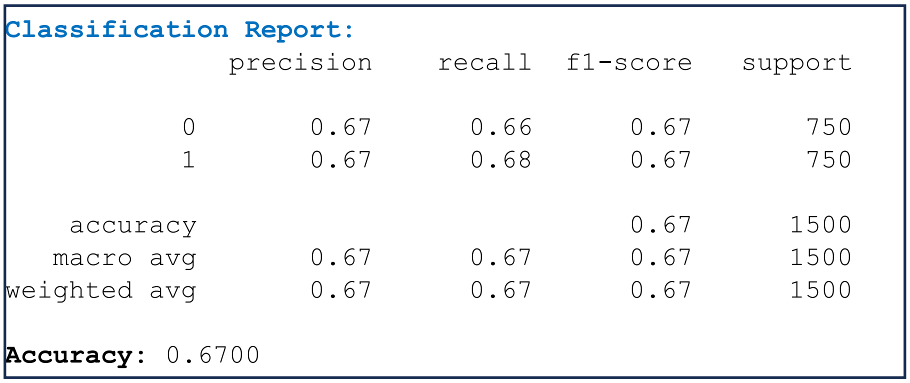
The table above shows a model of a model of a model of Random Forest model, its confusion matrix, and its classification report, both of which give a complete picture of the model's predictive performance for a problem of binary classification. According to its confusion matrix, it predicted 527 cases accurately to be its first-class case and predicted 223 cases inaccurately to be its first-class case's false positive, and 517 correct and 233 incorrect for its second-class case, respectively, for both its cases and incorrect cases for its respective cases. According to its classification report, its first-class case's precision is 0.69, with 69% of its prediction for its first-class case correct, and its recall is 0.70, with 70% of its actual cases for its first-class case detected accurately by its model. For its first-class case, its F1-score is 0.69, a balanced mix of its recall and its precision, and for its second-class case, its precision is 0.69, with a recall of 0.69 and an F1-score of 0.69, respectively. Overall, its accuracy is 0.6960, a 69.60% accuracy in its overall prediction, and its performance is a high level of accuracy, with room for improvement, namely in its ability to reduce its false negatives for its second-class case.

c) XG-Boost Modelling

The code script performed the training and testing of an XG-Boost Classifier. It loaded in the XGB-Classifier function in module xgboost first. It constructed a model with a certain random state for reproduction, disabled the use of a label encoder (which is deprecated in newer XG-Boost releases), and set evaluation to 'logloss' for use with XGB classifiers. It then trained a model with training data (X-train, y-train) through the use of its fit function. It then created a prediction for a test set (X-test) through the use of its prediction function with a trained model. It then tested its performance. Afterward, the code printed out a confusion matrix to assess the count of true/false positive/negative values. Subsequently, it printed out a classification report with precision, recall, F1-score, and support for each label value. Ultimately, it then printed out overall accuracy in the test set, rounded to four places. The following is a full view of the XG-Boost model performance.

Output:

Table 3: XG-Boost Classification Report



```
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.66      0.67       750
           1       0.67      0.68      0.67       750

    accuracy                           0.67      1500
   macro avg       0.67      0.67      0.67      1500
weighted avg       0.67      0.67      0.67      1500

Accuracy: 0.6700
```

The table above presents an XG-Boost model's performance, with key statistics through its classification report and its confusion matrix, describing its performance in a two-class classification problem. As per its confusion matrix, 497 samples have been predicted accurately as a true negative 253 samples predicted in error as a false positive for its first class, and 508 samples predicted accurately and 242 samples predicted in error as a false negative for its second class. As per its classification report, its first-class precision is 0.67, and thus 67% of its prediction for its first-class is correct, and its recall is 0.67, and 67% of actual samples have been captured. For its first-class, its first-class F1-score is 0.67, representing a balanced performance for both its classes' instances. For its second class, its precision, recall, and first-class stand at 0.67, supporting a similar performance for both classes. Overall, its accuracy is 0.6700, and thus 67.00% of all its predictions are correct, representing a fair level of predictive performance with room for improvement, specifically in improving its prediction for both its classes' instances.

Comparison of All Models

The implemented code compared three algorithms – Logistic Regression, Random Forest, and XG-Boost – in terms of performance via calculation and visualization of key evaluation metrics. It devised a function calculate-metrics to calculate accuracy, precision, recall, and F1-score for a prediction of each model over a test set. It then iterated over each trained model, constructed a prediction, calculated the metrics via the function, and stored them in a list. It then converted this list to a Pandas Data Frame for easier manipulation and visualization. It then devised a bar plot for comparative visualization of three machines' performance over four metrics. There is a single bar for a model's performance over a single metric, with a direct comparison of their strengths and weaknesses via a single plot. There were labels, a title, and a legend for ease of visualization, and it concluded with a comparative analysis of the three machines' performance.

Output:

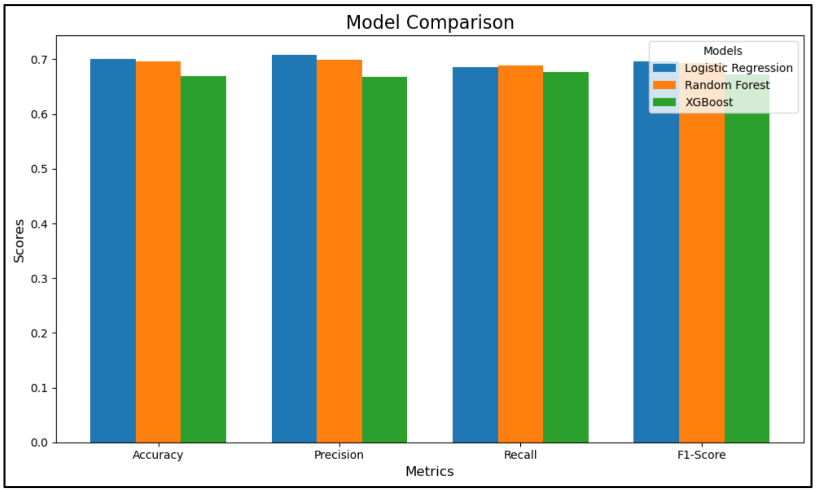


Figure 10: Portrays Model Comparison

The bar plot shows a comparative analysis of three algorithms for machine learning—Random Forest, Logistic Regression, and XG-Boost—based on four significant performance values: Precision, Recall, F1-Score, and Accuracy. All three values for algorithms have been displayed in three colors, with Logistic Regression in blue, Random Forest in orange, and XG-Boost in green. Retrospectively, comparing accuracy values, one can observe that Logistic Regression takes a high score, with Random Forest following closely, and XG-Boost following a little behind them. Comparing values for precision, Random Logistic takes a high value, proving its efficiency in generating correct positive values, but comparing values for recall, all three algorithms have a relatively similar value, with a little edge for them over Logistic Regression. Comparing values for F1-Scores, a similar scenario can be seen, proving that Logistic Regression outdoes both algorithms in most values, proving its solidity in such an analysis. Overall, through such a plot, a fair picture of the comparative strengths and weaknesses of all three algorithms can be seen, proving that Logistic Regression is the most reliable algorithm for predicting values in a dataset.

Insights on U.S. Energy Trends

Analyzing U.S. consumption trends reveals significant information regarding fluctuations in demand in sectors and regions, with a multi-faceted interplay of factors contributing to them. Perhaps one of the most salient observations is the extreme range in consumption between sectors, such as residential, commercial, industrial, and transportation sectors. For instance, residential demand peaks in summertime when high temperatures cause increased use of air conditioning, and winter demand peaks in colder regions of the country when heating is paramount. Meanwhile, the industrial sector will have relatively steady requirements for most of the period, with spikes during times of boom in the economy when production accelerates. Besides, geographical variation comes through; warmer-climate states, for instance, Florida and Texas, have a larger per capita level with high intensity for cooling, and milder-climate states, such as Washington and Oregon, have a lesser overall consumption but can have spikes in terms of winter demand. Geographic variation then comes together with localized policies, sources of energy, and availability of renewable sources, and a mosaicked picture of consumption profiles for the country is developed.

External factors, such as climate fluctuations and policy, have a profound impact on consumption trends and patterns in terms of energy use. Climate change, for one, has increased extreme weather events, both impacting demand and supply for energy. Higher temperatures promote increased consumption of electricity for cooling, and extreme winter storms can cause spikes in unanticipated heating demand. In addition, shifts in both state and federal policies have a significant role in shaping consumption behavior in terms of energy use. For example, incentives for renewable use, such as tax credits for solar panels, have motivated residential and commercial entities to transition towards cleaner sources of energy, impacting the overall demand for fossil fuels in general. Policies for curbing carbon emissions have also motivated industries to develop and implement cleaner, more efficient technology, impacting consumption trends in industries. On top of that, the current transition towards electric cars (EVs) is transforming transportation demand for energy, with estimates suggesting that increased use of EVs could have profound implications for the consumption of electricity, particularly in urban settings. As such factors develop, it is important to understand their impact on the demand for energy in terms of effective decision-making in terms of both energy policy and management of resources.

Scenario-Based Forecasting

Scenario-based forecasting is a tool for strategic planning for future demand for energy under new circumstances, allowing decision-makers to assess consequences and make preparations in anticipation. One such plausible scenario is a heightened use of renewable sources, such as wind and solar, for energy use. As technology brings down renewable energy expenses and new states implement policies for clean energy, a rapid transformation in demand for energy can be envisaged. In such a scenario, traditional use of fossil fuels can be curtailed, and electricity demand can go up with electrification in sectors including transportation and warming. It can make for a localized grid, with new options for storing energy and smart grid technology taking a dominant role in balancing demand and supply. Besides, the integration of electric cars in the economy for energy will necessitate high investments in charging infrastructure, altering demand and peak consumption times.

Conversely, an economic recession brings about a reverse scenario that can cause a fall in demand for energy in a variety of sectors. During the recession, production in industries tends to slow down, and less consumption of energy for processes and production in industries takes place. Home and commercial sectors can even face less demand with tightening of consumption and less consumption of cooling and heating systems. Enterprises can even postpone energy-intensive expansion and modernization, and overall consumption is reduced even further. In such a case, one can observe the role played by economic factors in predicting energy, and any fluctuations in GDP, employment, and confidence have a direct impact on demand for energy. By forecasting such scenarios, energy professionals can comprehend the likely impact of economic factors and renewable integration in defining future consumption, and industries and policymakers can develop strong strategies that respond to changing trends in the energy sector. All such planning is important in offering security and sustainability in an uncertain environment.

VI. Practical Applications

Implications for Energy Suppliers

Utility companies can benefit a lot through machine learning (ML)-based prediction in terms of distribution efficiency and overall operational efficiency. With ML algorithms, utility companies can utilize humungous volumes of consumption in the past to make future demand predictions with high accuracy. With predictive powers, providers of energy can schedule production in such a manner that availability and loss can both maximize and minimize, respectively, such that demand and availability can closely resemble one another. For instance, when ML predicts a spurt in demand in a certain period, utility companies can make a proactive realignment in terms of producing less and, when not enough, utilize additional power from renewable sources when possible. Not only will this maximize customer satisfaction through fewer outages, but it will also allow cleaner sources of energy to become a part of the grid, improving sustainability.

Moreover, optimizing grid stability is significant for utility companies, and ML can make a big contribution in such a scenario. With real-time analysis of data, utility companies can identify impending grid infrastructure vulnerabilities or bottlenecking and correct them even before such larger issues arise. For example, ML algorithms can scan for trends in maintenance activity and past outages and make predictions regarding when and where failures will occur, and preventive maintenance and timely upgrades can then follow suit. Apart from that, with increased accuracy in demand forecasting, utility companies can save operational costs incurred in overproducing during off-peaking and underproducing during peaking, respectively. Not only will such an optimization save operational expenses, but it will also lessen the environmental impact of unnecessary production, in harmony with overall sustainability objectives.

Policy Recommendations for the U.S. Government

For regulatory bodies, information gained through advanced ML techniques can inform efficient and sustainable policies in the energy sector. Policymakers can utilize analysis through ML to identify trends and correlations in consumption, and in such a manner, implement programs with a specific target towards resolving specific concerns in the energy sector. For example, through consumption and demographics analysis, regulators can identify communities that require additional infrastructure or incentives for efficiency in consumption. With a model backed by data, such information aids ineffective use of resources, with an optimized impact of policies and a less agonizing path towards a cleaner future in terms of energy.

Furthermore, integrating national planning with machine learning will require a whole-of-government, whole-of-economy, and whole-of-society collaboration, including governments, energy providers, and technology developers working together. Policymakers will have to drive the creation of information-sharing and collaboration frameworks between and amongst numerous stakeholders in the energy economy. By creating partnerships that maximize the potential of ML, governments can become ever more effective at forecasting trends in energy, experimenting with new policies for effectiveness, and creating flexible regulatory frameworks for managing new and emerging issues. In addition, investments in training and educational programs in ML and analysis of data in the energy economy will allow workers to effectively utilize these tools, and allow America to remain at the cutting edge of energy innovation.

Sustainability and Optimizing Resources

Machine learning algorithms have a strong potential for efficiency driving and footprint reduction in sectors and industries. With consumption information processing, Machine Learning algorithms can identify trends and provide actionable insights for residential, commercial, and industrial consumption optimizations. For instance, smart home technology powered with

ML can dynamically adjust heating and cooling in real-time about occupancy trends, cutting consumption loss in a significant manner. Outside residential, ML can go to community and city scales for distribution optimizations, driving efficient consumption, and demand-side management methodologies for convincing consumption at peak times.

Enhancing demand response programs is yet another significant application of machine learning for energy sustainability. With a marriage of ML algorithms and demand response programs, a utility can make a more reliable prediction of peak demand times and encourage consumers to restrict consumption during such times. Not only does it stabilize the grid, but it reduces the use of peaker plants, fueled with fossil fuels, during high-demand times, not an environmentally friendly practice and a source of added greenhouse gas emissions. Energy planning for storing can also gain with the incorporation of ML, with consumption behavior examined and forecasts generated for when to discharge and charge energy storing systems, with stored renewable energy utilized effectively, balancing demand and supply and conserving use of non-renewable sources of energy. Overall, the integration of machine learning with programs for energy efficiency and resource planning is critical in achieving sustainability and minimizing the overall greenhouse footprint of the energy sector.

VII. Discussion and Future Directions

Challenges in Energy Forecasting

Energy forecasting entails numerous obstacles that must be addressed in an endeavor to make it reliable and accurate. Reporting restrictions and discrepancies in reporting energy information are one of its biggest obstacles. Most providers and utility companies have numerous approaches for reporting and collecting information, and discrepancies between them can occur, and such discrepancies can cause inaccuracies in forecasts. For example, discrepancies in reporting peak demand, such as its measurement, and renewable energy output, such as its calculation, can yield incomplete and untruthful datasets. In dataset combination, such discrepancies become a significant problem, and in dealing with trends over a period, such discrepancies can become even more complex to work with. Besides, the energy sector is under external factors such as the economy, legislation, and weather, and such factors can cause uncertainty in forecasting algorithms. As a result, not only is high-quality information challenging but getting information in a uniform and similar format for reporting in the sector is challenging too.

In addition to information-related concerns, real-time forecasting continues to represent a critical challenge in an increasingly variable energy environment. With the increased integration of renewable sources, demand, and supply fluctuations can become a reality with increased and unpredictable frequency. For instance, solar and wind output is weather-sensitive, and quick weather shifts can have a significant impact on output volumes. Consequently, traditional forecasting methodologies employing past information can fall short of keeping pace with such quick fluctuations, and grid instability and ineffectiveness can become a reality. To counter such an issue, providers of energy must develop new, complex real-time forecasting methodologies capable of acting with quick adaptability to changing scenarios. That could involve inserting sophisticated algorithms in machines capable of reading information from a variety of sources, including weather forecasts, grid sensors, and consumption behavior trends, and providing timely and reliable forecasts.

Limitations of the Study

While the report is full of information about methodologies for prediction and its use, its weaknesses have to be taken into consideration. One such significant weakness is in terms of biases in data collection and model generalizability. For training a model for a prediction, data could not necessarily represent a larger population for prediction. For instance, in case a dataset is urban in its composition, a model trained with such a dataset will underpredict in rural settings, whose consumption behavior can vary immensely. Biases can even arise in terms of a model's generalizability in case a lot of technological and/or regulative development took place in between, and one must, therefore, interpret the report with caution, knowing its observations cannot necessarily apply in any region and scenario universally.

Another notable limitation is that deeper, more complex deep learning algorithms have to be developed in a position to make long-term forecasts with added accuracy. Conventional machine algorithms can make effective forecasts on a short-term basis but can break down with complex, nonlinear consumption trends that characteristically unfold over long-term horizons. Deep learning algorithms, with hierarchical representations of information, have the potential for added nuance and predictive powers, but at a high computational and high-data demand for training. Hence, added work must go towards researching deep learning methodologies in energy forecasting, to develop algorithms that can generalize effectively between scenarios and make reliable long-term forecasts.

Future Research Directions

Looking ahead, several exciting future avenues for energy forecasting include a range with high potential for model accuracy and efficiency improvement. One such direction involves a union between IoT technology smart grid analysis and machine learning forecasting. With smart meters and smart devices becoming widespread, colossal amounts of real-time information about consumption, production, and grid performance become available in real-time. By utilizing such information, one can develop increasingly real-time forecasting models that dynamically react to real-time demand and supply fluctuations in energy. With IoT smart grid analysis and machine learning forecasting, utility companies can make smarter decision-making processes, and with timely and efficient management of energy, a smarter and more efficient use of infrastructure can become a reality. Besides, a union between IoT analysis and machine learning can make predictive maintenance of grid infrastructure a reality, with reduced downtime and operational costs.

Another exciting path for future work is in utilizing reinforcement learning for dynamic energy management. Unlike traditional supervised learning, reinforcement learning learns best through experimenting and then corrects its actions, and is therefore best utilized in scenarios in which real-time decision-making is critical. In the case of energy management, reinforcement algorithms can learn to maximize consumption and production strategies for energy in real time based on changing demand and supply scenarios. For example, such algorithms can learn to modulate the consumption of energy in an industrial scenario in real-time according to price signals or learn to manage effectively storing and recalling energy, maximizing the use of renewable sources of energy. By studying such new methodologies, future work can make future energy systems more robust flexible, and capable of dealing with new energy environment complexity.

VIII. Conclusion

The utmost objective of this research project was to develop and evaluate machine learning algorithms with accuracy in predicting trends in America's consumption of energy. By employing complex methodologies such as neural networks, regression analysis, and ensemble approaches, this work aims to enhance the accuracy of forecasts in terms of demand for energy in residential, commercial, and industrial sectors. The U.S. consumption datasets covered a wide variety of information representing the consumption of electricity, use of fuel, and integration of renewable sources in residential, commercial, and transportation sectors. Most datasets used for analysis are taken from the U.S. Energy Information Administration (EIA) and the Department of Energy (DOE), offering in-depth statistics about production, consumption trends, and price trends. With our dataset consisting of continuous values for consumption of energy and many predictor factors, we compared a variety of machine algorithms, encompassing XG-Boost, Logistic Regression, and Random Forest. The implemented code compared three algorithms – Logistic Regression, Random Forest, and XG-Boost – in terms of performance via calculation and visualization of key evaluation metrics. It devised a function calculate-metrics to calculate accuracy, precision, recall, and F1-score for a prediction of each model over a test set. Retrospectively, comparing accuracy values, one can observe that Logistic Regression got a high score, with Random Forest following closely, and XG-Boost following a little behind them. Overall, through strategic plots, the comparative strengths and weaknesses of all three algorithms were seen, proving that Logistic Regression was the most reliable algorithm for predicting values in a dataset. Utility companies can benefit a lot through machine learning (ML)-based prediction in terms of distribution efficiency and overall operational efficiency. With ML algorithms, utility companies can utilize humungous volumes of consumption in the past to make future demand predictions with high accuracy.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Aderibigbe, A. O., Ani, E. C., Ohenhen, P. E., Ohalete, N. C., & Daraojimba, D. O. (2023). Enhancing energy efficiency with AI: a review of machine learning models in electricity demand forecasting. Engineering Science & Technology Journal, 4(6), 341-356.

[2] Alabi, T. M., Aghimien, E. I., Agbajor, F. D., Yang, Z., Lu, L., Adeoye, A. R., & Gopaluni, B. (2022). A review of the integrated optimization techniques and machine learning approaches for modeling, prediction, and decision-making on integrated energy systems. Renewable Energy, 194, 822-849.

[3] Al Mukaddim, A., Mohaimin, M. R., Hider, M. A., Karmakar, M., Nasiruddin, M., Alam, S., & Anonna, F. R. (2024). Improving Rainfall Prediction Accuracy in the USA Using Advanced Machine Learning Techniques. *Journal of Environmental and Agricultural Studies*, *5*(3), 23-34.

[4] Anonna, F. R., Mohaimin, M. R., Ahmed, A., Nayeem, M. B., Akter, R., Alam, S., ... & Hossain, M. S. (2023). Machine Learning-Based Prediction of US CO2 Emissions: Developing Models for Forecasting and Sustainable Policy Formulation. *Journal of Environmental and Agricultural Studies*, *4*(3), 85-99.

[5] Barua, A., Karim, F., Islam, M. M., Das, N., Sumon, M. F. I., Rahman, A., ... & Khan, M. A. (2025). Optimizing Energy Consumption Patterns in Southern California: An AI-Driven Approach to Sustainable Resource Management. *Journal of Ecohumanism*, *4*(1), 2920-2935.

[6] Chowdhury, M. S. R., Islam, M. S., Al Montaser, M. A., Rasel, M. A. B., Barua, A., Chouksey, A., & Chowdhury, B. R. (2024). PREDICTIVE MODELING OF HOUSEHOLD ENERGY CONSUMPTION IN THE USA: THE ROLE OF MACHINE LEARNING AND SOCIOECONOMIC FACTORS. *The American Journal of Engineering and Technology*, *6*(12), 99-118.

[7] Deb, C., Zhang, F., Yang, J., Lee, S. E., & Shah, K. W. (2017). A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74, 902-924.

[8] Forootan, M. M., Larki, I., Zahedi, R., & Ahmadi, A. (2022). Machine learning and deep learning in energy systems: A review. Sustainability, 14(8), 4832.

[9] Hasan, M. R. (2024). Revitalizing the electric grid: A machine learning paradigm for ensuring stability in the USA. *Journal of Computer Science and Technology Studies*, *6*(1), 141-154.

[10] Hasan, M. R., Shawon, R. E. R., Rahman, A., Al Mukaddim, A., Khan, M. A., Hider, M. A., & Zeeshan, M. A. F. (2024). Optimizing Sustainable Supply Chains: Integrating Environmental Concerns and Carbon Footprint Reduction through AI-Enhanced Decision-Making in the USA. *Journal of Economics, Finance and Accounting Studies*, *6*(4), 57-71.

[11] Hossain, M. S., Mohaimin, M. R., Alam, S., Rahman, M. A., Islam, M. R., Anonna, F. R., & Akter, R. (2025). AI-Powered Fault Prediction and Optimization in New Energy Vehicles (NEVs) for the US Market. *Journal of Computer Science and Technology Studies*, *7*(1), 01-16.

[12] Khalil, M., McGough, A. S., Pourmirza, Z., Pazhoohesh, M., & Walker, S. (2022). Machine Learning, Deep Learning, and Statistical Analysis for forecasting building energy consumption—A systematic review. Engineering Applications of Artificial Intelligence, 115, 105287

[13] Khan, P. W., Byun, Y. C., Lee, S. J., Kang, D. H., Kang, J. Y., & Park, H. S. (2020). Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources. Energies, 13(18), 4870.

[14] Rahman, M. K., Dalim, H. M., Reza, S. A., Ahmed, A., Zeeshan, M. A. F., Jui, A. H., & Nayeem, M. B. (2025). Assessing the Effectiveness of Machine Learning Models in Predicting Stock Price Movements During Energy Crisis: Insights from Shell's Market Dynamics. *Journal of Business and Management Studies*, *7*(1), 44-61.

[15] Reza, S. A., Chowdhury, M. S. R., Hossain, S., Hasanuzzaman, M., Shawon, R. E. R., Chowdhury, B. R., & Rana, M. S. (2024). Global Plastic Waste Management: Analyzing Trends, Economic and Social Implications, and Predictive Modeling Using Artificial Intelligence. *Journal of Environmental and Agricultural Studies*, *5*(3), 42-58.

[16] Singh, S., Bansal, P., Hosen, M., & Bansal, S. K. (2023). Forecasting annual natural gas consumption in the USA: Application of machine learning techniques-ANN and SVM. Resources Policy, 80, 103159.

[17] Shahcheraghian, A., & Ilinca, A. (2024). Advanced machine learning techniques for energy consumption analysis and optimization at UBC Campus: Correlations with meteorological variables. Energies, 17(18), 4714.

[18] Somu, N., MR, G. R., & Ramamritham, K. (2021). A deep learning framework for building energy consumption forecast. Renewable and Sustainable Energy Reviews, 137, 110591.

[19] Sumon, M. F. I., Osiujjaman, M., Khan, M. A., Rahman, A., Uddin, M. K., Pant, L., & Debnath, P. (2024). Environmental and Socio-Economic Impact Assessment of Renewable Energy Using Machine Learning Models. *Journal of Economics, Finance and Accounting Studies*, *6*(5), 112-122.

[20] Sumon, M. F. I., Rahman, A., Debnath, P., Mohaimin, M. R., Karmakar, M., Khan, M. A., & Dalim, H. M. (2024). Predictive Modeling of Water Quality and Sewage Systems: A Comparative Analysis and Economic Impact Assessment Using Machine Learning. *in Library*, *1*(3), 1-18.