University of Hertfordshire UH

Name: Vankatasaisireesha Bolisetty
Student ID: 23036961

1. Project Overview

1.1 Project Title Machine Learning-Based Prediction and Optimization of Renewable Energy Production

1.2 Summary of Project Topic and Background

The growing demand for sustainable energy solutions has led to an increased reliance on renewable energy sources such as solar and wind. However, the unpredictable nature of weather conditions poses challenges in accurately forecasting energy production and demand. This project aims to apply machine learning techniques to forecast renewable energy production using a dataset that combines information from the European Network of Transmission System Operators for Electricity (ENTSOE), Red Electric España, and OpenWeather.

The project involves preprocessing and integrating energy generation data with weather conditions and time features. Three machine learning models—Random Forest, XGBoost, and LSTM—will be implemented to evaluate their performance in predicting total energy load. The project emphasizes the importance of accurate forecasting in optimizing energy distribution and supporting grid stability.

1.3 Research Question

How can advanced machine learning models such as Random Forest, XGBoost, and LSTM be used to improve the accuracy of renewable energy production forecasts using time-series weather and load data?

1.4 Project Objectives

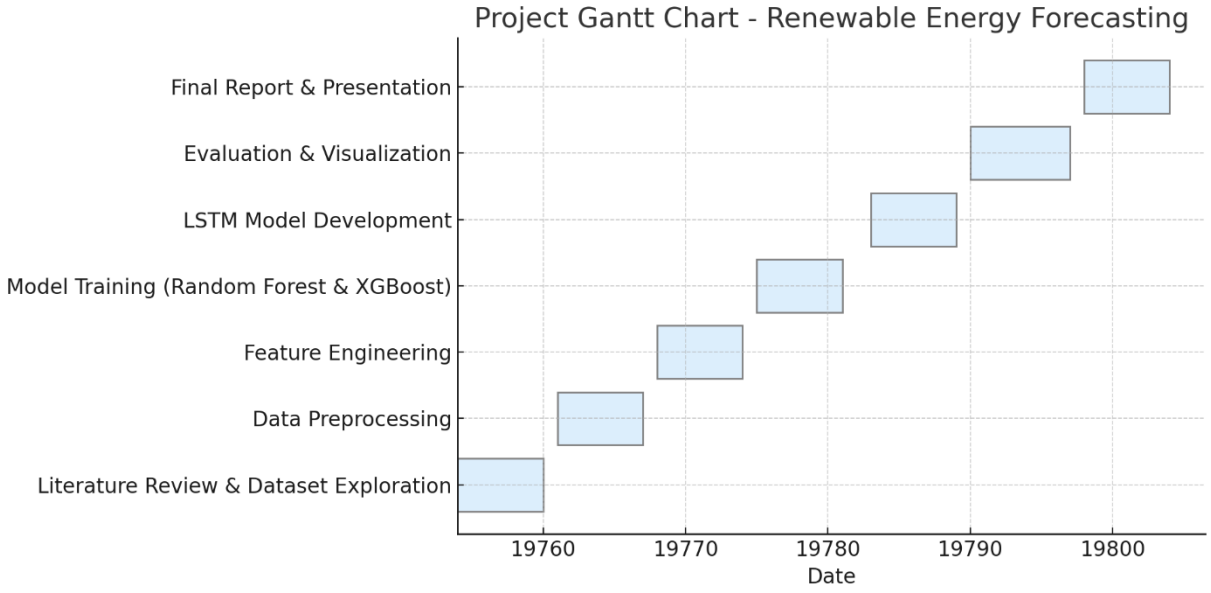
- Collect and preprocess historical energy and weather data.
- Engineer relevant features such as lag values, cyclical time features, and temperature-based inputs.
- Train and evaluate machine learning models including Random Forest, XGBoost, and LSTM.
- Compare model performance using RMSE and R^2 metrics.
- Visualize model predictions and residuals.
- Document the workflow, findings, and challenges in a final report.
- Create a presentation summarizing key outcomes and recommendations.

1.5 References

1. Hyndman, R.J. and Athanasopoulos, G., 2018. Forecasting: Principles and Practice. 2nd ed. OTexts. Available at: https://otexts.com/fpp2/
2. Makridakis, S., Spiliotis, E., and Assimakopoulos, V., 2020. The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 36(1), pp.54–74. Available at: https://www.sciencedirect.com/science/article/pii/S0169207019301128
3. Rahman, M.M. and Saha, T.K., 2022. Machine learning-based forecasting of renewable energy production: A review. Renewable Energy Reports, 8, pp.229–248. Available at: https://www.sciencedirect.com/science/article/pii/S2352484722000051

2. Project Plan: Task List and Timeline

Week	Dates	Task	Description
1	Feb 1–Feb 7	Literature Review & Dataset Exploration	Review research papers and examine the Kaggle dataset structure
2	Feb 8–Feb 14	Data Preprocessing	Handle missing values, format timestamps, and merge datasets
3	Feb 15–Feb 21	Feature Engineering	Create lag features, normalize weather metrics, extract time components
4	Feb 22–Feb 28	Model Training (Random Forest & XGBoost)	Train models and tune hyperparameters
5	Mar 1–Mar 7	LSTM Model Development	Build and train LSTM using TensorFlow/Keras
6	Mar 8–Mar 15	Evaluation & Visualization	Generate evaluation metrics and comparison plots
7	Mar 16–Mar 22	Final Report & Presentation	Compile project documentation and prepare viva slides



3. Data Management Plan

3.1 Dataset Overview The dataset used in this project is hosted on Kaggle under the title "Energy Consumption, Generation, Prices and Weather" and integrates multiple data sources including the European Network of Transmission System Operators for Electricity (ENTSOE), Red Electric España, and OpenWeather. This composite dataset provides comprehensive coverage of hourly energy-related metrics, including generation by different sources (e.g., solar, wind, fossil, nuclear), electricity load, and corresponding weather attributes such as temperature, humidity, wind speed, and cloud cover.

https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather?resource=download

The dataset spans several years and includes detailed time-indexed records that are ideal for time series analysis and predictive modeling. The high-resolution temporal granularity supports the construction of lag-based and seasonal features, essential for training machine learning models.

3.2 Data Collection & Processing

- Tools used for data collection and preprocessing include Python libraries such as Pandas and NumPy for data wrangling, Jupyter Notebook and Google Colab for interactive development, and Matplotlib/Seaborn for exploratory data analysis.
- The preprocessing steps involved cleaning and aligning datasets from different sources based on timestamp. Irrelevant or redundant columns were dropped to reduce noise and dimensionality. Missing values were imputed or rows were removed depending on the context and importance of features. Lag features and cyclical time features were generated to help models capture trends and seasonality.

3.3 Version Control

- All code versions and updates are maintained in a private GitHub repository: https://github.com/sireesha1010/Renewable_energy_production

- The repository is updated weekly with commits and includes checkpoints of notebooks, cleaned datasets, and charts to ensure traceability and reproducibility.

3.4 Storage and Backup

- Source code is maintained in the GitHub repository and includes scripts for preprocessing, feature engineering, and model training.
- The raw and processed datasets are stored locally and backed up regularly to Microsoft OneDrive to prevent data loss. The backup system is updated alongside major project milestones.
- Access to project files is restricted to the student and supervisor to maintain project integrity.

3.5 Ethics and Licensing

- The dataset is publicly available and distributed under a Creative Commons Zero (CC0) license, which permits unrestricted academic use.
- The dataset does not include any personally identifiable information (PII) or sensitive data.
- This project complies with the University of East London's ethics policy for research involving public datasets and does not require ethical clearance.