

Predicting Global Nuclear Energy Suitability Using Machine Learning Analysis

Saketh Chodisetti

2/16/2023

Abstract

This study addresses the global reliance on non-renewable resources for electricity production in a society seeking cleaner, more reliable alternatives. The research goal is to predict the most suitable countries for new nuclear power plants based on climate factors, existing plants, and the electricity they already generate. We did this by testing various machine learning models, including Logistic Regression, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Decision Tree, and Gradient Boosting, to identify the most efficient classification algorithm. Our results indicate that the KNN model achieved the highest performance metrics. Using our KNN model, Burundi was predicted as the most suitable country for future nuclear development.

1. Introduction

This research investigates the use of machine learning models to predict the most suitable countries for developing new nuclear power plants in the near future. By utilizing specific global power plant data, several supervised learning algorithms were trained and evaluated. As society seeks a transition away from carbon-intensive fossil fuels, nuclear power remains a viable option.

The main problem addressed in this study is that current fuel sources are increasingly unsustainable. To resolve this, we found a model that predicts which countries are most suitable for future nuclear power plant development by using existing trends, climate factors, and energy generation patterns.[1] By looking at a large dataset of over 34,000 plants, and ~190 being nuclear powered, the goal was to build models that can classify which countries contain the conditions that are associated with being environmentally sustainable. We found that 5 countries in the temperate-African region were the most suitable for future nuclear development.

2. Background

As the global electricity demand increasingly rises, much of the world continues to heavily rely on non-renewable energy sources because lack the infrastructure for other wind, hydro, and geothermal. This contributes to environmental degradation and severe climate change. As we seek sustainable alternatives, nuclear power is a seemingly viable option as it has a very high energy output and density, it has relatively low greenhouse gas emissions, and has a small land footprint to accommodate for the lack of infrastructure. However, placing these plants requires careful consideration of climate conditions, existing infrastructure, and current electricity production. [1]

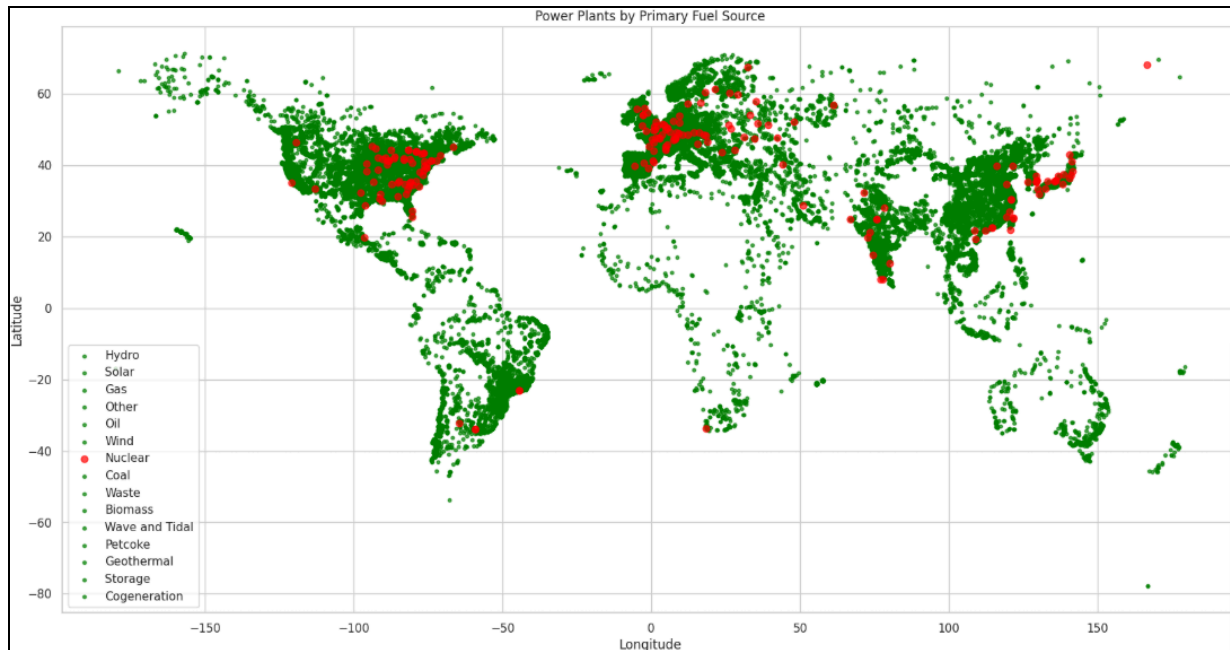
Several approaches have been made in order to attempt to solve energy infrastructure problems. The International Atomic Energy Agency (IAEA) uses the “Milestone Approach,” which is a framework that can evaluate a country’s industrial capabilities and potential. [8] While accurate, this process is more qualitative and not efficient. In our specific methodology, we utilized the Global Power Plant Database provided on Kaggle to use specific technical features for model development.

3. Dataset

We used a dataset from Kaggle, “Powerplants Database Global”. [5] The dataset contained one file that contained five important features: the country the plant was located in, the exact latitude and longitude of the plant, the electrical generating capacity in megawatts, and the primary fuel source of the plant. Plant locations are shown in Figure 1a, green. The latitude was used as a proxy for climate zones, as shown in Figure 1b.

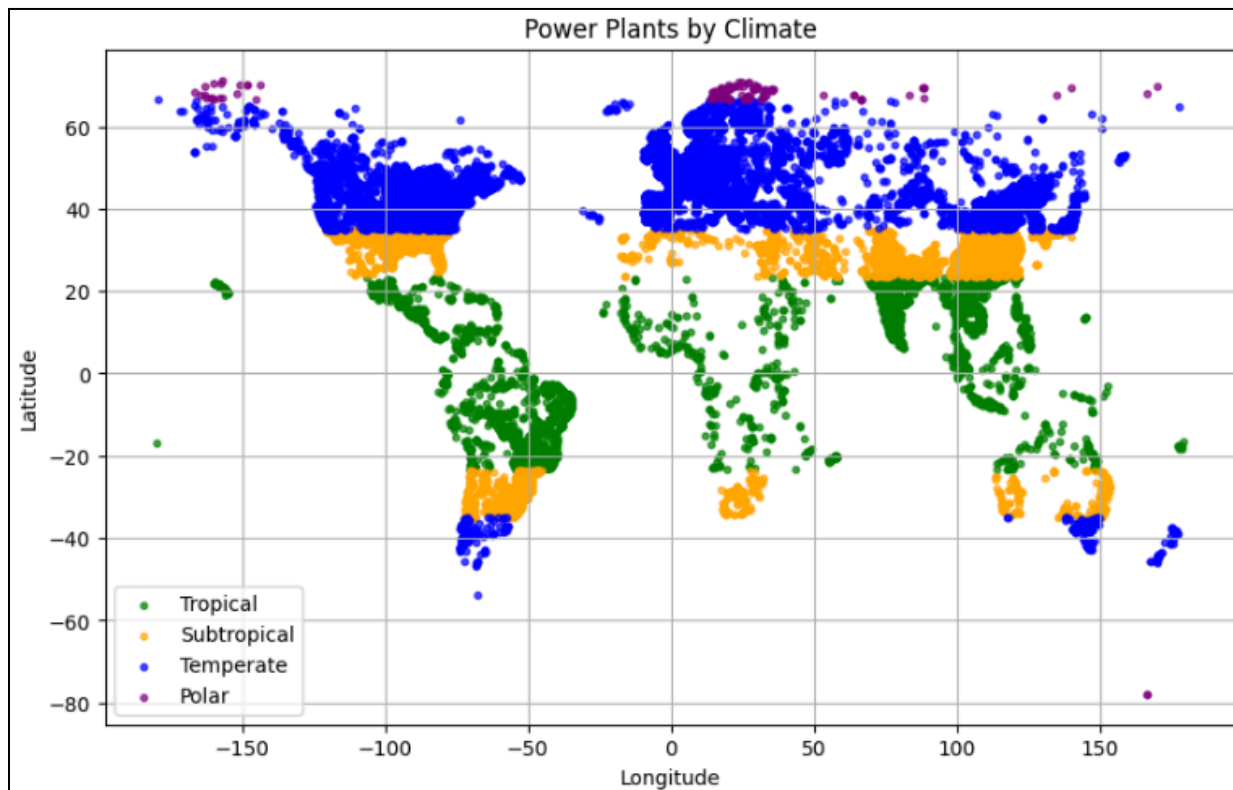
As shown in Figure 1a, the red dots above show only the distribution of nuclear power plants globally, which are mostly located in developed countries like the USA, China, Japan, and across Europe, most of which are clustered in specific latitude ranges. A handful of plants remain in Africa and South America, however. This implies that nuclear plants are concentrated in specific global regions rather than being evenly spread across the world. [9]

Figure 1a: Electricity Generation Powerplant Locations by Primary Fuel Source



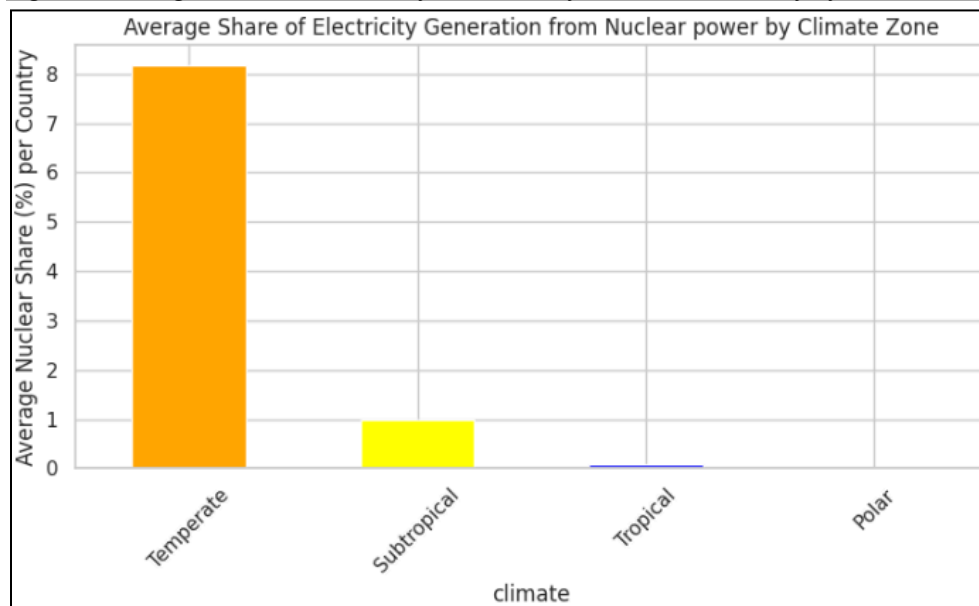
Most plants are located in temperate regions, with tropical and subtropical areas also hosting many facilities, according to Figure 1b. Polar regions, however, have very few power plants, suggesting that global nuclear energy infrastructure tends to be built in climates that are more moderate and stable, such as temperate regions.

Figure 1b: Powerplant Locations by Climate Zone



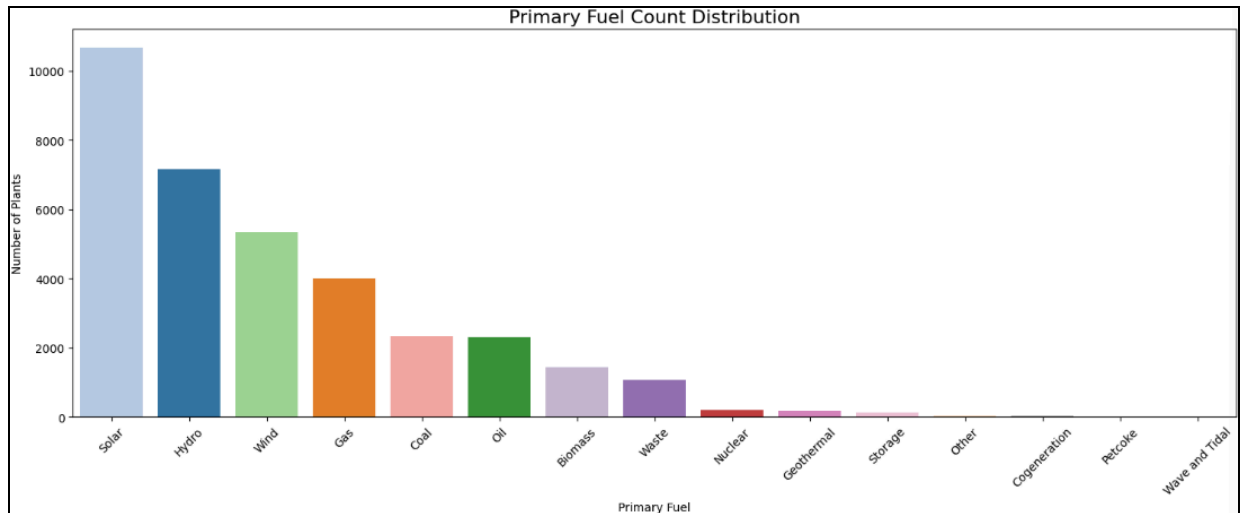
Nuclear electricity generation is overwhelmingly concentrated in temperate climate zones compared to all other zones. As seen in Figure 2, this reinforces the idea that nuclear power is most common in regions with moderate climates, like temperate or subtropical. The vertical axis is the fraction of total energy that is nuclear.

Figure 2: Average Percent of Electricity Generated by Nuclear Per Country by Climate Zone



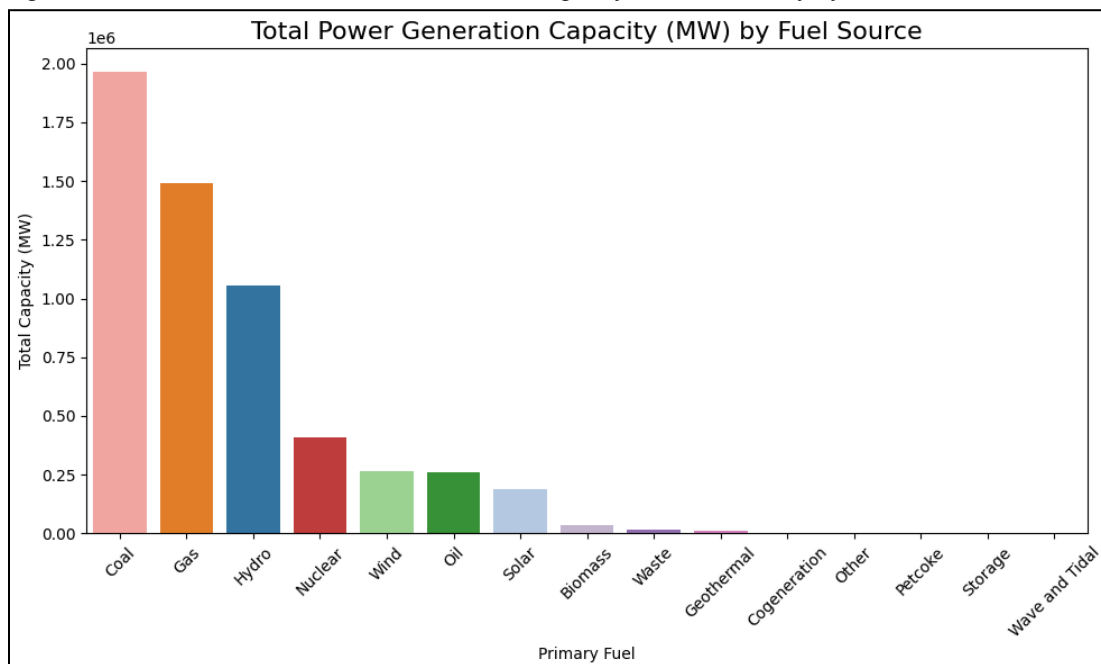
Based on Figure 3, solar energy is the most common source for electricity-generating power plants, followed by hydroelectricity and wind power, based on the number of plants. Nuclear plants make up one of the smallest categories in terms of the number of plants globally. The vertical axis is the total number of plants.

Figure 3: Distribution of Number of Plants Per Fuel Source



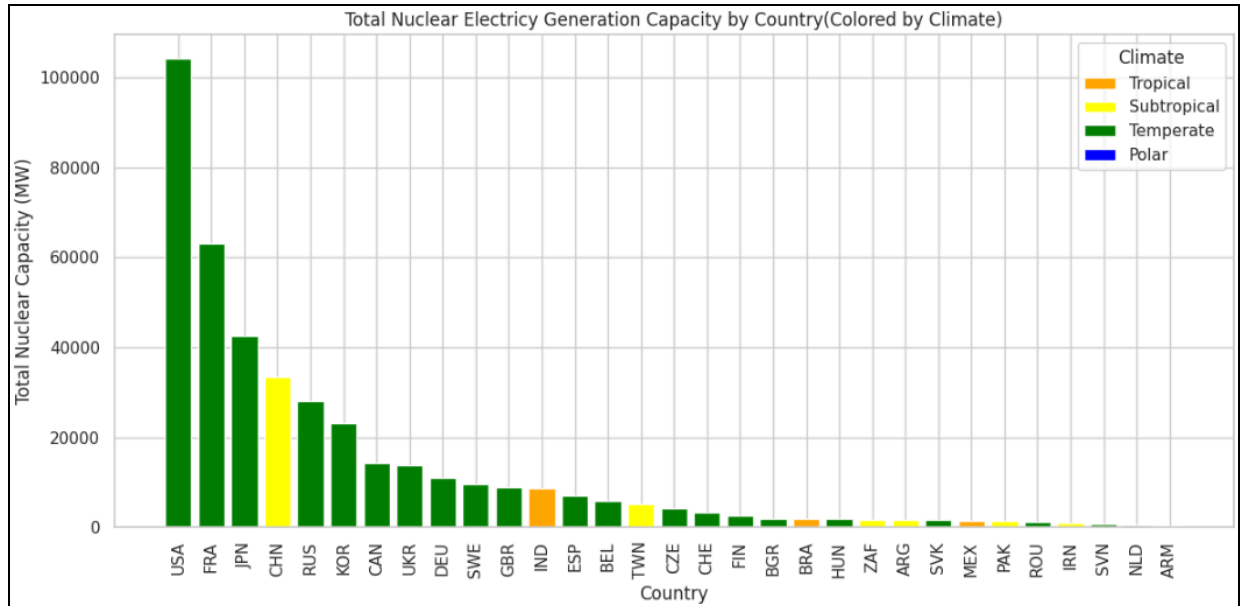
However, according to Figure 4, Coal is responsible for producing the most energy globally despite not having the largest count, followed by gas and hydroelectricity. Nuclear energy generates the 4th most power despite having one of the lowest numbers of plants globally, implying that it is an efficient and viable source of energy. The vertical axis is the total generation capacity in MW.

Figure 4: Distribution of Total Power Generation Capacity in MW Globally by Fuel Source



As seen in Figure 5, the United States has the highest total nuclear power capacity in the world by a large margin, followed by France and Japan. Almost all of the high-capacity nuclear power plants are located in countries within temperate climate zones, with the exception of China. Nuclear plants in the polar regions are virtually nonexistent. [8]

Figure 5: Total Electricity Capacity Generated by Nuclear Power Per Country by Climate Zone



4. Methodology / Models

Data Preprocessing

After uploading the raw dataset, the first step was to clean it up so we could focus on the most important information. We dropped a lot of unnecessary columns: the specific names and unique ID codes of the power plants, details about additional fuel sources, who owns the plants, links to reporting websites, and old data regarding how much electricity the plants generated between 2013 and 2017. Instead, we only kept the essential features needed to map and analyze the plants, including the country name, the electrical capacity (how much power the plant can make, in megawatts), the exact latitude and longitude coordinates, and the primary energy source.[2]

As shown in Table 1, to make sure the data was high-quality and accurate, we took two important steps. First and foremost, we checked and removed any duplicate entries, ensuring that each power plant is only counted once. After removing these redundant rows, we had a clean set of 34,936 unique power plants to work with. Next, we looked for any missing information (null values) and removed rows that were missing critical details, like capacity or location. This ensures the dataset is complete and ready for training our machine learning models. If a feature wasn't present for a significant number of plants, we dropped the feature entirely. This includes the years that the plants were first in operation. [4]

Table 1: First 5 Rows of Cleaned Global Power Plant Dataset

	Country Abbr.	Country	Capacity (MW)	Latitude	Longitude	Primary Fuel Source	Climate
0	AFG	Afghanistan	33.0	32.3220	65.119	Hydro	Subtropical

1	AFG	Afghanistan	10.0	31.6700	65.7950	Solar	Subtropical
2	AFG	Afghanistan	10.0	31.6230	65.7920	Solar	Subtropical
3	AFG	Afghanistan	66.0	34.5560	69.4787	Hydro	Subtropical
4	AFG	Afghanistan	100	34.6410	69.7170	Hydro	Subtropical

Feature Engineering

We performed feature engineering on this dataset. As my research question describes how climate can affect the future placement of nuclear plants, we engineered a feature for climate zones based on the plants' latitude coordinate shown in the last column of Table 1. As a result, we had another feature that we could use for my model development. [6]

To make the data ready for machine learning models, we used one-hot encoding on the climate column. This transforms text categories such as "Tropical" or "Polar" into separate columns of 1s and 0s, helping the computer understand environmental factors without assuming any numerical order. Next, we encoded the 'primary_fuel' column while dropping one category, the abbreviated name of the countries, to reduce redundant information, which helps prevent the model from becoming too complex or overfitting the data. We dropped all columns related to 'Nuclear' fuel to focus the analysis specifically on non-nuclear energy sources. Finally, we split the data into a train and test dataset, with the test dataset being 30% of the original dataset and the training dataset being the remaining 70% of the data.

Model Development and Evaluation

For our classification models, we implemented a structured approach starting with robust data preprocessing to ensure model stability and performance. Given that several algorithms are sensitive to the magnitude of input features, we applied StandardScaler to normalize the data, ensuring all features contributed equally to the model's calculations.

We began development with established baseline models, specifically Logistic Regression and Decision Trees. Logistic Regression works by fitting a sigmoid function to the data to predict the probability of a binary outcome, fitting a regression line to the data to separate classes. Decision Trees, on the other hand, create a set of binary rules based on feature values to partition the data.

To improve predictive performance, we utilized ensemble methods: Random Forest and Gradient Boosting. In the Random Forest Classifier, multiple Decision Tree models are "bootstrapped" or trained on different random subsets of the data. The final prediction is then determined by taking the majority vote of the results from these trees. Through this process, the final result is more reliable because each Decision Tree model will make errors in different places, and the overall effect will cancel out. Similarly, Gradient Boosting builds trees sequentially, where each new tree aims to correct the errors of the previous ones.

Finally, we explored distance-based and kernel-based models: K-Nearest Neighbors (KNN) and Support Vector Machines (SVM). KNN classifies data points based on the majority class of their nearest neighbors in the feature space, relying heavily on the scaling we performed earlier. SVM works by finding the best dividing line, called a hyperplane, that separates different classes with the largest gap between them. [3]

To properly evaluate our models, we used four distinct metrics: Accuracy, Precision, Recall, and F1 Score.

Accuracy measures the overall proportion of correct predictions, but it can be deceptive if the dataset is imbalanced. Precision specifically calculates the ratio of true positive predictions to all positive predictions made by the model, which is crucial when the cost of a false positive is high. Recall measures the model's ability to identify all relevant instances within a dataset, focusing on minimizing false negatives. Finally, the F1 Score provides a harmonic mean of Precision and Recall, offering a balanced metric that penalizes extreme imbalances between the two.

5. Results and Discussion

Model Development Results

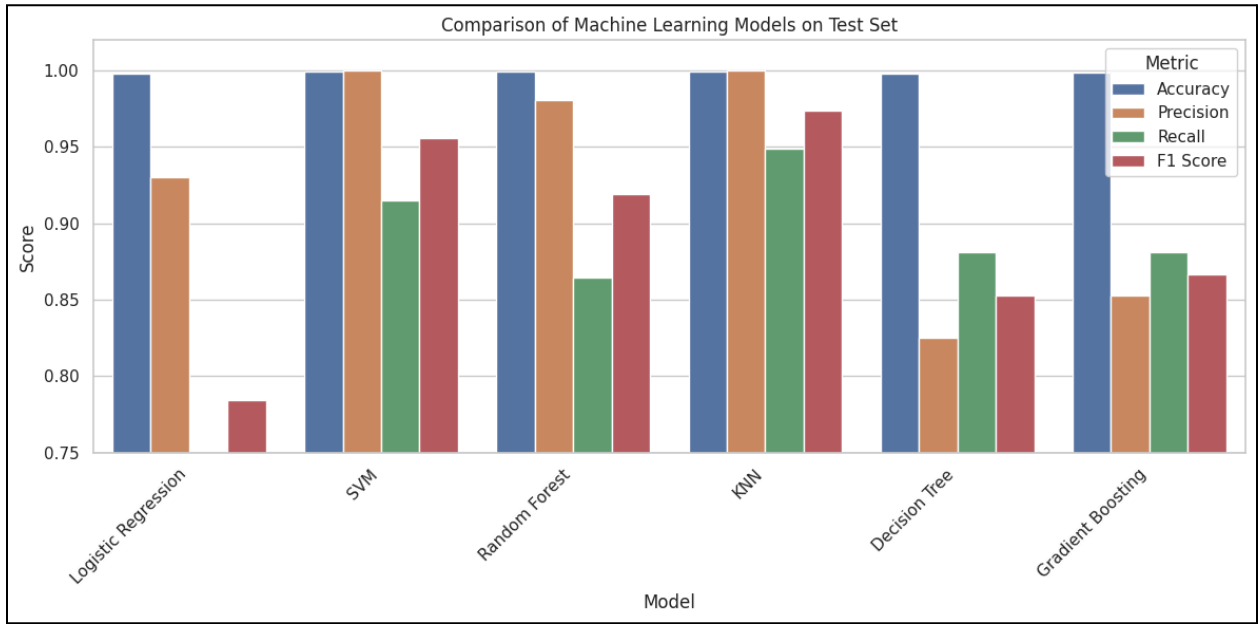
Table 2 shows the performance of different classification models against various metrics (Accuracy, Precision, Recall, and F1 Score). Logistic Regression was the lowest-performing model in terms of F1 Score, followed by the Decision Tree. Based on our evaluation, our two best models were the KNN model and the Support Vector Machine (SVM) model. While the baseline Gradient Boosting model performed well, the KNN model provided the highest overall F1 Score. [7]

Table 2: Model Results on Different Metrics

Model	Metric			
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.998283	0.918367	0.762712	0.833333
Support Vector Machine (SVM)	0.999523	1.000000	0.915254	0.955752
Random Forest Classifier	0.999428	1.000000	0.898305	0.946429
K-Nearest Neighbors (KNN)	0.999618	1.000000	0.932203	0.964912
Decision Tree	0.998378	0.828125	0.898305	0.861789
Gradient Boosting	0.998473	0.852459	0.881356	0.866667

Figure 6 provides a visual comparison of these scores, highlighting the significant lead KNN and SVM maintain over simpler models like Logistic Regression in terms of balanced performance (F1 Score).

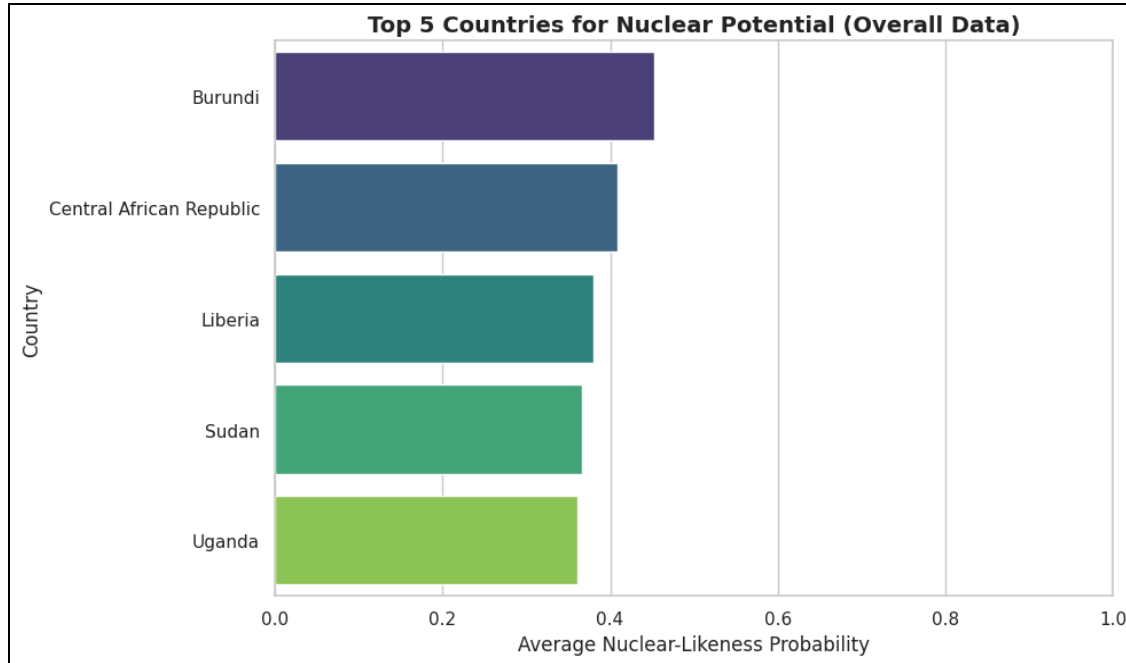
Figure 6: Comparison of Machine Learning Models on Test Set



Model Predictions

Figure 7 provides the KNN model’s predictions of the top five countries with the highest potential for future nuclear development, which are Burundi, Central African Republic, Liberia, Sudan, and Uganda, in that order. These countries showed patterns in climate and energy infrastructure that are similar to places where nuclear power already works well. Burundi ranked the highest overall, even though it currently has no nuclear plants, which suggests it may have strong potential for future development. Overall, these results show that several African countries could be promising locations for expanding nuclear energy in the future.

Figure 7: Predicted Top Countries for Nuclear Development



6. Conclusion

Based on our analysis, the K-Nearest Neighbors (KNN) model turned out to be the most effective algorithm for this project. It had the strongest overall performance and gave the most reliable predictions about which countries could be good candidates for future nuclear development. While other models like SVM, Random Forest, and Gradient Boosting also performed well, none of them matched the balance of accuracy and consistency shown by KNN.

After choosing KNN as the best model, we used it to estimate nuclear-likelihood probabilities for every country in the dataset for countries that do not have pre-existing nuclear plants. The results showed that Burundi, Central African Republic, Liberia, Sudan, and Uganda were the top five countries with the highest potential for nuclear power development. These countries shared climate and energy characteristics similar to places where nuclear power already works successfully, even though many of them currently have little or no nuclear infrastructure.

Overall, this study shows that machine learning can be a helpful tool for identifying where clean and reliable energy sources, like nuclear power, could be expanded in the future. The findings point to several African countries as promising locations and highlight how data-driven methods can support global efforts toward more sustainable energy.

7. Future Goals

In the future, this research could be expanded by including more detailed environmental, economic, and political data to make the predictions even more accurate. Using larger or more updated datasets, as well as testing

more advanced machine learning models, could also improve the results. These enhancements would help create an even clearer picture of which countries are best prepared for future nuclear energy development.

Acknowledgments

I would like to thank my mentor, Mr. Gary Mavko, for their help and guidance throughout this project and for providing essential resources and dedication towards my research.

References

1. Byers, Logan, et al. "A Global Database of Power Plants." *World Resources Institute*, 4 May 2018, www.wri.org/research/global-database-power-plants.
2. *Data Preprocessing: A Complete Guide with Python Examples* | Datacamp, www.datacamp.com/blog/data-preprocessing. Accessed 16 Feb. 2026.
3. "Getting Started with Classification." *GeeksforGeeks*, GeeksforGeeks, 8 Nov. 2025, www.geeksforgeeks.org/machine-learning/getting-started-with-classification/.
4. "How to Preprocess Data in Python." *Built In*, builtin.com/machine-learning/how-to-preprocess-data-python. Accessed 16 Feb. 2026.
5. King, Alistair. "Nuclear Energy Datasets." *Kaggle*, 17 June 2024, www.kaggle.com/datasets/alistaiking/nuclear-energy-datasets/data?select=power_plant_database_global.csv.
6. *Koppen Climate Classification* | Definition, System, & Map | Britannica, www.britannica.com/science/Koppen-climate-classification. Accessed 16 Feb. 2026.
7. "Machine Learning | Google for Developers." *Google*, Google, developers.google.com/machine-learning/crash-course. Accessed 16 Feb. 2026.
8. *Nuclear Power in a Clean Energy System – Analysis* - IEA, www.iea.org/reports/nuclear-power-in-a-clean-energy-system. Accessed 16 Feb. 2026.
9. "Pyplot Tutorial#." *Pyplot Tutorial - Matplotlib 3.10.8 Documentation*, matplotlib.org/stable/tutorials/pyplot.html. Accessed 16 Feb. 2026.

