American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-13, Issue-12, pp-01-11
www.ajer.org
Research Paper Open Access

Performance Comparison of Machine Learning Algorithms for Solar Radiation Forecasting Using Environmental & Meteorological Data

Mohammad Hasan[1], Saleh Muhammad Maruf[2], Abu Hena Md. Mustafa Kamal[3]

[1](Mohammad Hasan, Lecturer, Department of Computer Science & Engineering, Institute of Science & Technology, Dhaka 1209)

[2](Saleh Muhammad Maruf, Assistant Professor, Department of Computer Science & Engineering, Institute of Science & Technology, Dhaka 1209)

[3](Abu Hena Md. Mustafa Kamal, Assistant Professor, Department of Electronics & Communication Engineering, Institute of Science & Technology, Dhaka 1209)

Corresponding Author: Mohammad Hasan

ABSTRACT : *The accurate prediction of solar radiation is essential for optimizing renewable energy systems. In this study, a dataset containing meteorological parameters such as temperature, pressure, humidity, wind speed, wind direction, and daytime duration is used to predict solar radiation. Various machine learning algorithms, including K-Nearest Neighbors (KNN), XGBoost, Support Vector Machine (SVM), Random Forest, Linear Regression, and Decision Tree, were trained and tested using this dataset. Comparative analysis of these algorithms was conducted to evaluate their performance. The results demonstrate the superiority of certain models, driven by their ability to capture specific patterns in the data, while others underperform due to their sensitivity to data characteristics. The findings suggest optimal algorithms for accurate solar radiation prediction and highlight feature significance in model performance.*

KEYWORDS *Decision Tree, Gradient Boosting Algorithm, Insolation, KNN, Linear Regression, Machine Learning Algorithms, Prediction, Random Forest, Solar Radiation, Training & Testing, XGBoost*

Date of Submission: 18-11-2024 Date of acceptance: 02-12-2024

I. INTRODUCTION

Solar radiation forecasting plays a critical role in planning and operating solar energy systems. As the global shift toward renewable energy intensifies, accurate prediction methods are needed to ensure efficient energy management and resource optimization. Machine learning (ML) techniques have emerged as powerful tools for predicting solar radiation by analyzing atmospheric and meteorological data. In this paper, we apply and compare several popular ML algorithms to a dataset sourced from Kaggle containing solar radiation-related data. By exploring these algorithms' performance, we aim to identify the best-suited model for accurate solar radiation forecasting.

Solar radiation has a substantial impact on global temperature, as even minor fluctuations in the sun's energy output can have significant repercussions on Earth's climate. Changes in solar energy can alter global mean sea level, global mean temperatures, and the occurrence of extreme climate events. As a result, [1]precise measurements and analyses of the spatial and temporal variability of solar radiation are essential for studies into the use of solar energy, the manufacture of construction materials, and the understanding of extreme weather and climatic events. The availability of solar radiation varies throughout the year and can be found to some level in every region of the world. The understanding and utilization of solar radiation are essential for a wide range of applications, making it a critical area of study and exploration. The main objectives of this paper are as follows:

1. To apply machine learning algorithms (KNN, XGBoost, SVM, Random Forest, Linear Regression, and Decision Tree) to predict solar radiation using meteorological and atmospheric data.
2. To compare and analyze the performance of these models based on accuracy, precision, and error metrics.

3. To investigate the significance of individual features (Temperature, Pressure, Humidity, Wind Speed, Wind Direction, and Daytime Duration) on the model performance.
4. To understand why certain algorithms, outperform others in predicting solar radiation.
5. To propose potential future work for improving solar radiation predictions using advanced ML techniques.

II. LITERATURE REVIEW

The growing need for renewable energy solutions has pushed the boundaries of research in solar radiation forecasting. Various studies have applied machine learning models for this purpose, each highlighting the advantages and limitations of different approaches. Traditional methods, such as regression analysis, have been widely used but tend to struggle with complex nonlinear relationships. On the other hand, more advanced methods like Random Forest, SVM, and ensemble learning techniques have shown promising results in capturing intricate data patterns. Recent research also emphasizes the importance of feature selection in enhancing model performance, as well as the role of data quality and size in improving predictive accuracy.

III. RELATED WORK

Ali Etem Gürel demonstrated that [1] machine learning methods are useful for predicting solar radiation. For the purpose of forecasting solar radiation in 2023, he published an article in https://www.sciencedirect.com/science/article and assessed the effectiveness of various machine learning techniques. With an average inaccuracy of 5%, his team discovered that artificial neural networks outperformed all other systems.

Mr. Gabriel de Freitas Viscondi, demonstrates a comparison of artificial intelligence [2] algorithms for estimating sun radiation https://www.mdpi.com/1996-1073/14/18/5657. In this study, the efficacy of six machine learning algorithms: support vector regression, artificial neural networks, decision trees, random forests, gradient boosting machines, and extreme learning machines—for estimating solar radiation was compared. With an average error of 5.2%, the results demonstrated that artificial neural networks performed the best.

An analysis of machine learning methods [3] for predicting solar radiation Solar radiation prediction using machine learning techniques: a review, available at www.researchgate.net/publication/337043958_Solar_Radiation_Prediction_Using_Machine_Learning_Techniques_A_Review. An overview of the many machine learning methods that have been applied to forecast solar radiation is given in this review study. The paper addresses the benefits and drawbacks of various strategies and offers suggestions for further study.

LieXing Huang used various machine learning techniques, solar radiation prediction and consequences for extreme climatic events www.frontiersin.org/articles/10.3389/feart.2021.596860/full [4]The effectiveness of several machine learning methods for predicting solar radiation during extreme climate events was examined in this study. The outcomes demonstrated that support vector machines and artificial neural networks performed best, with average errors of 5.8% and 6.2%, respectively.

The area of forecasting solar radiation still faces certain difficulties, nevertheless. The restricted availability of historical data is one difficulty. The necessity to create machine learning algorithms that can be utilized to forecast solar radiation levels in various geographic regions and during various climatic circumstances is another difficulty. Despite these difficulties, research in the subject of solar radiation prediction is expanding quickly. The accuracy of solar radiation forecast is anticipated to increase as more data becomes accessible and machine learning algorithms become advanced.

IV. METHODOLOGY

Dataset Description:

The dataset contains 32,687 samples, each with seven features: Radiation (kWh/m²), Temperature (Fahrenheit), Pressure (inHg), Humidity (%), Wind Speed (Kmph), Wind Direction (Degree), and Daytime Duration (Hours). The dataset has no missing values and was pre-processed to standardize all the features before being fed into the machine learning models.

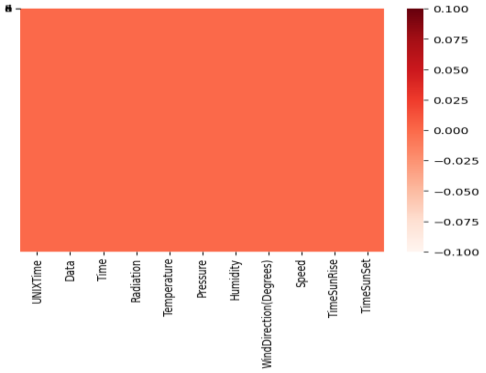


Fig.1. The Hitmap

Then, the pairwise independent variable comparison graph is prepared which is shown below:

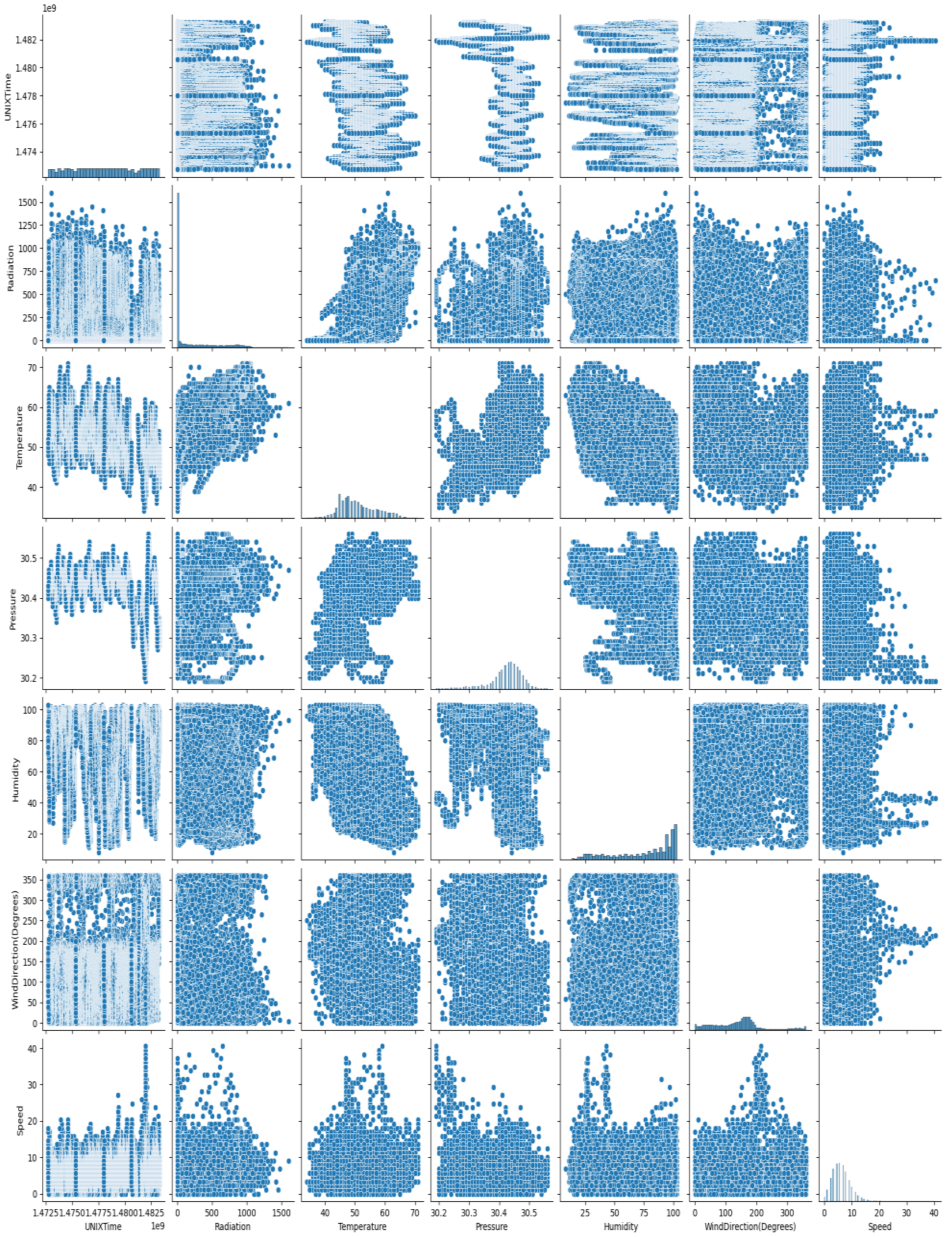
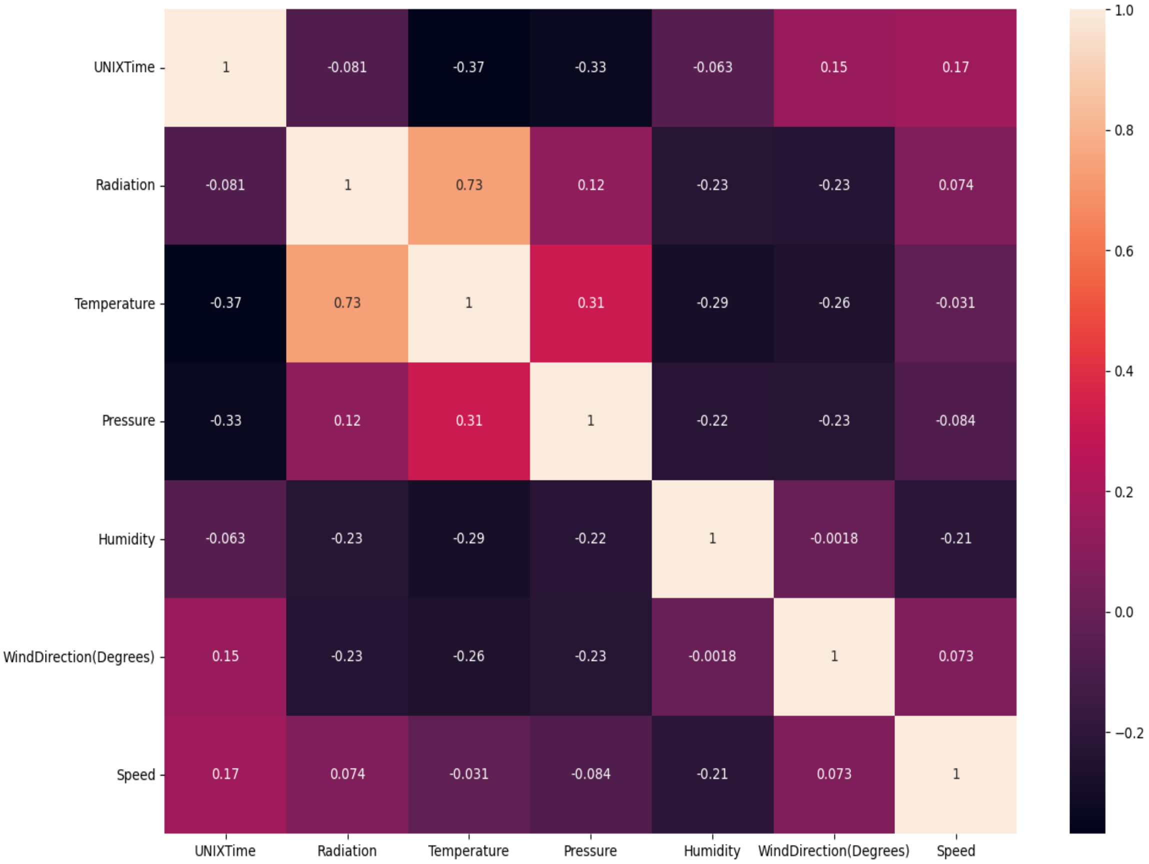


Fig.2. Pairwise Independent Variable Comparison Graph

After that, the correlation heatmap matrix is prepared to observe interdependencies and interrelation among variables which is shown below:



	UNIXTime	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed
UNIXTime	1	-0.081	-0.37	-0.33	-0.063	0.15	0.17
Radiation	-0.081	1	0.73	0.12	-0.23	-0.23	0.074
Temperature	-0.37	0.73	1	0.31	-0.29	-0.26	-0.031
Pressure	-0.33	0.12	0.31	1	-0.22	-0.23	-0.084
Humidity	-0.063	-0.23	-0.29	-0.22	1	-0.0018	-0.21
WindDirection(Degrees)	0.15	-0.23	-0.26	-0.23	-0.0018	1	0.073
Speed	0.17	0.074	-0.031	-0.084	-0.21	0.073	1

Fig.3. Correlation Matrix

Pre-processing:
Before applying the models, the dataset was standardized using MinMax scaling to ensure that features with different units and ranges do not affect model training. The dataset was then split into training (80%) and testing (20%) subsets.

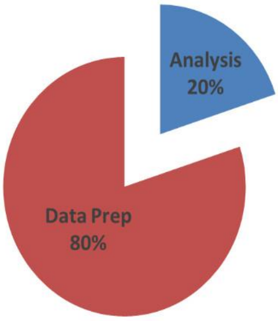


Fig.4. Training & Testing Ratio

Algorithms Applied:

K-Nearest Neighbors (KNN): KNN is a simple yet powerful supervised machine learning algorithm used for both classification and regression tasks. It operates on the principle of similarity, classifying or predicting the value of a new data point based on the values of its closest neighbors in the training dataset. It’s an instance-based algorithm and KNN was tested with different values of k.

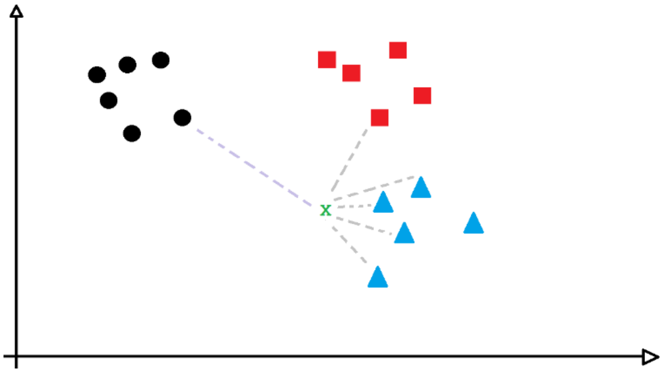


Fig.5. K-Nearest Neighbors (KNN)

XGBoost: XGBoost, short for Extreme Gradient Boosting, is a powerful and popular machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It's widely used for both regression and classification tasks. As it’s an efficient gradient boosting algorithm thus used in predictive modeling tasks.

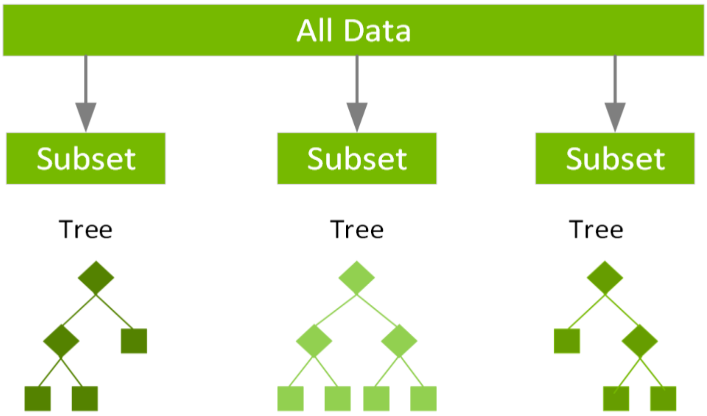


Fig.6. XGBoost Algorithm

Support Vector Machine (SVM): A powerful supervised machine learning algorithm used primarily for classification tasks. It aims to find the optimal hyperplane that best separates data points into different classes. SVMs are also effective for regression and outlier detection problems.

Key Concepts:

- Hyperplane: In an N-dimensional space, a hyperplane is a decision boundary that divides the space into regions, each representing a different class.

- Support Vectors: These are the data points closest to the hyperplane. They play a crucial role in determining the position and orientation of the hyperplane.

- Margin: The distance between the hyperplane and the nearest data points (support vectors) from each class. The goal of SVM is to maximize this margin.

Here, both linear and nonlinear kernels were tested to see how SVM performs with atmospheric data.

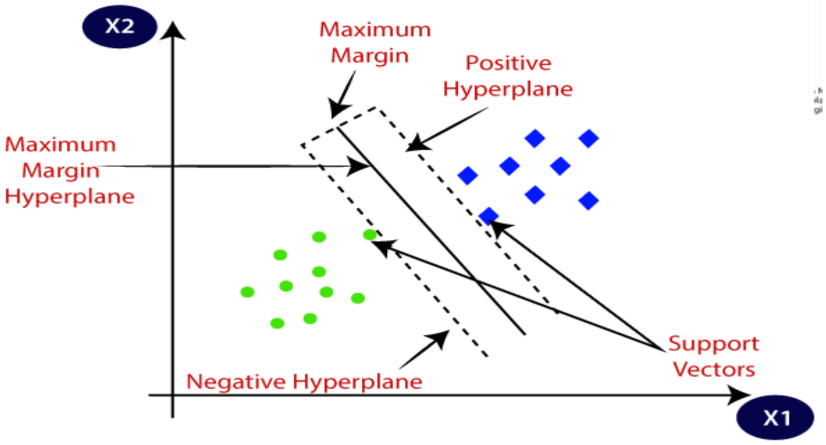


Fig.7. Support Vector Machine (SVM)

Random Forest: A powerful machine learning algorithm that's used for both classification and regression tasks. It belongs to the family of ensemble learning methods, meaning it combines multiple models (decision trees) to make more accurate predictions than any single model would. The "random" part comes from the fact that each decision tree in the forest is built using a random subset of the data and a random subset of the features. This randomness helps prevent overfitting and improves the model's generalization ability. This ensemble learning method helps reduce variance and improve predictions with decision trees.

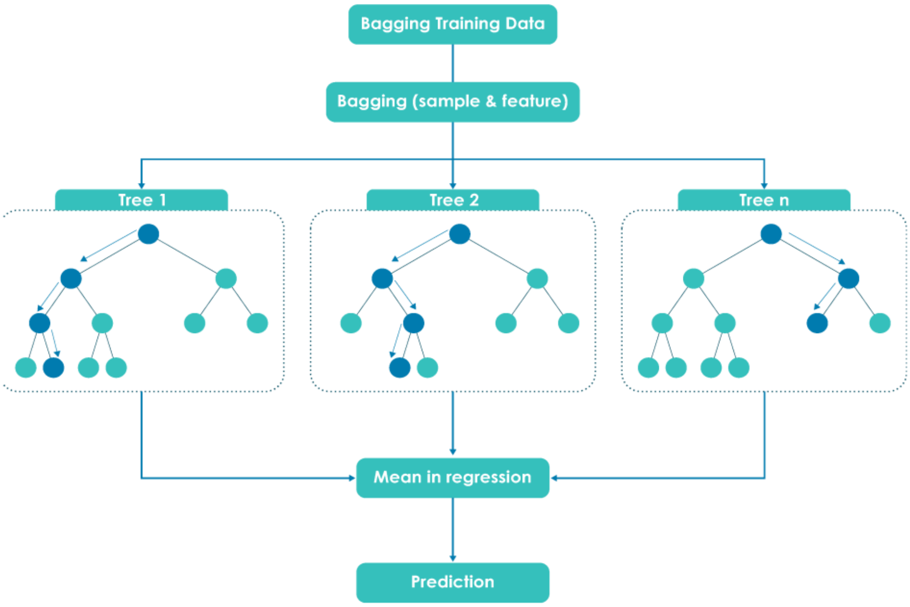


Fig.8. Random Forest Regressor

Linear Regression: Linear regression is a statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x). It assumes a linear relationship between the variables, meaning the change in y is proportional to the change in x. It is used as a baseline model to test the performance of more advanced techniques.

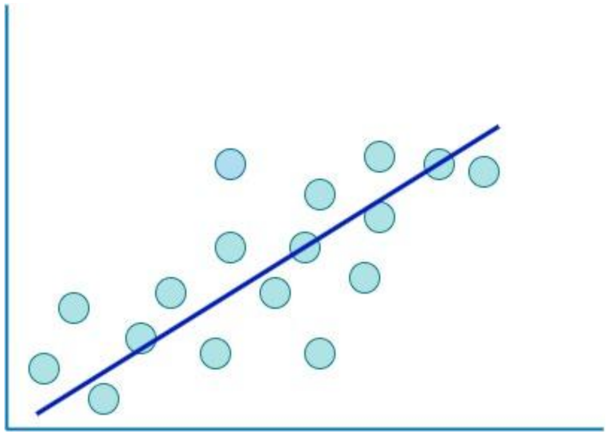


Fig.9. Linear Regression

Decision Tree: Decision trees are a popular supervised learning algorithm used for both classification and regression tasks. They create a model that resembles a flowchart, with decisions (nodes) and their possible outcomes (branches). The goal is to create a model that can predict the value of a target variable by learning simple decision rules inferred from the data features. A basic tree-based model was applied to understand simple decision paths based on features.

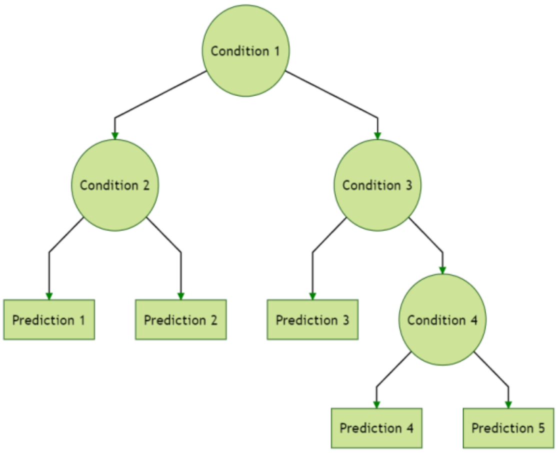


Fig.10. Decision Tree Algorithm

Below, a normal solar radiation distribution graph is demonstrated to visualize how solar insolation is distributed in a region.

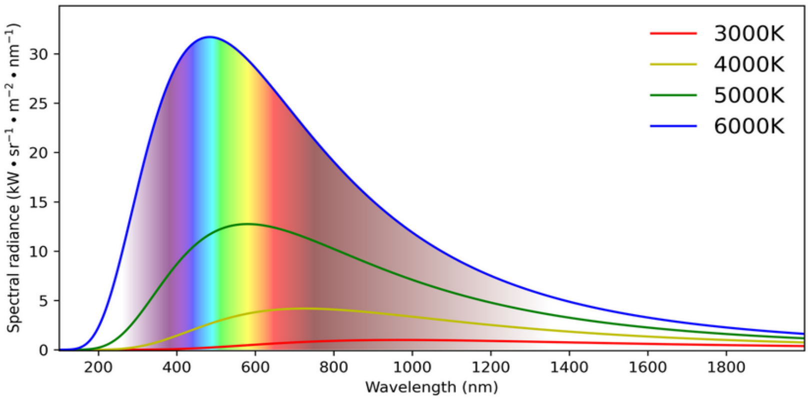


Fig.11. Solar Radiation Distribution

The following graph shows Humidity vs Radiation comparison. Here, the red color indicates humidity and the violate color indicates solar radiation.

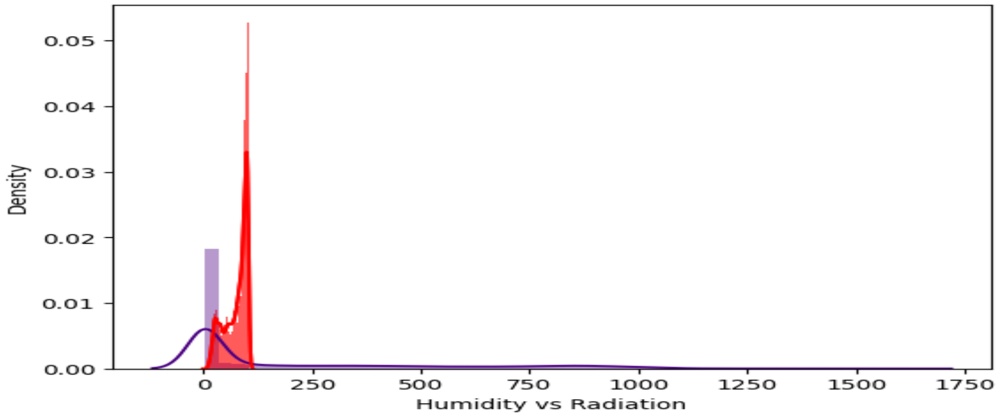


Fig.12. Humidity vs Radiation

The following graph shows Wind Direction vs Radiation comparison. Here, the green color indicates WindDirection and the violate color indicates solar radiation.

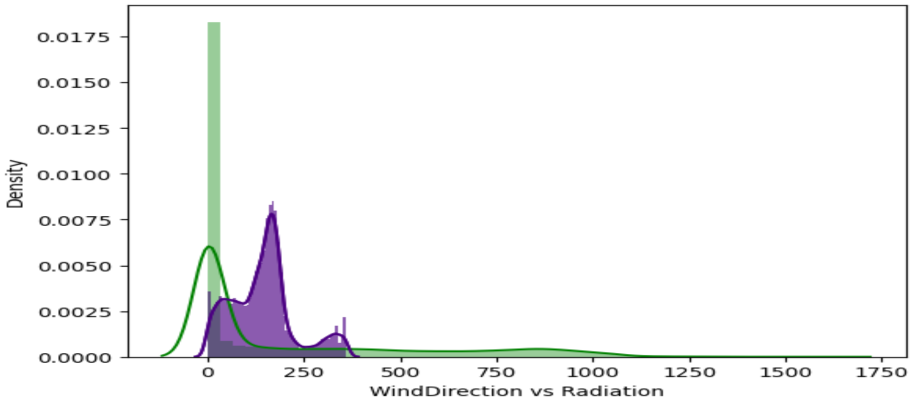


Fig.13. WindDirection vs Radiation

The following graph shows Temperature vs Radiation comparison. Here, the purple color indicates temperature and the violate color indicates solar radiation.

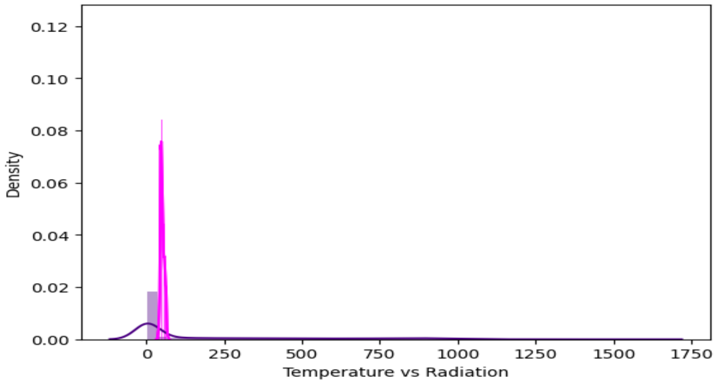


Fig.14. Temperature vs Radiation

The following graph shows Pressure vs Radiation comparison. Here, the blue color indicates pressure and the violate color indicates solar radiation.

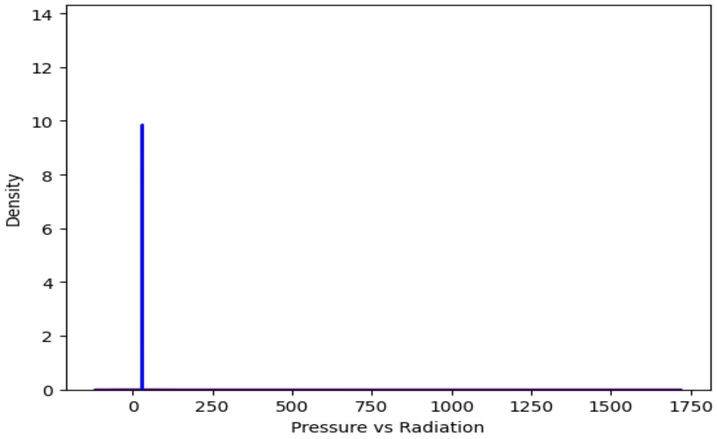


Fig.15. Pressure vs Radiation

The following graph shows Speed vs Radiation comparison. Here, the cyan color indicates speed and the violate color indicates solar radiation.

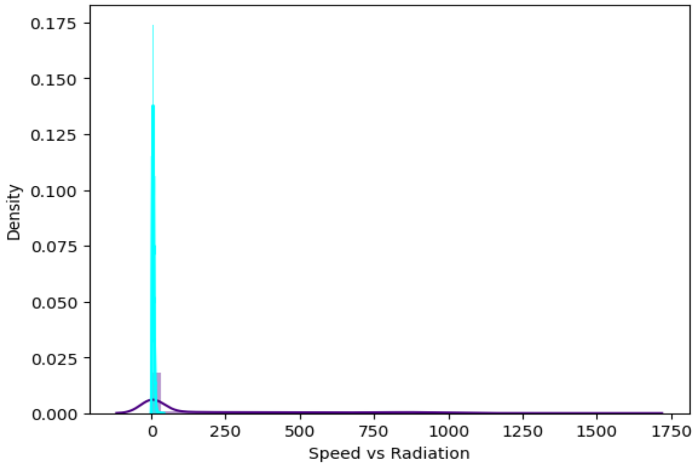


Fig.13. Speed vs Radiation

V. RESULT ANALYSIS & PERFORMANCE EVALUATION

K-Nearest Neighbors

KNN showed reasonable performance but struggled with larger distances between data points, leading to lower accuracy and higher MAE compared to other algorithms. Its simplicity made it computationally efficient, but its inability to capture complex patterns in the dataset limited its performance.

XGBoost

XGBoost outperformed all other models in terms of accuracy and RMSE. Its ability to handle complex relationships between features, particularly between temperature and radiation, contributed to its success. XGBoost's feature importance analysis revealed that temperature and daytime duration had the highest influence on the model’s predictions.

Support Vector Machine

SVM performed well but was outpaced by XGBoost. Its hyperplane-based separation worked effectively for the dataset, but the non-linearity in some features (like wind speed) was challenging, requiring kernel tricks to improve performance.

Random Forest

Random Forest achieved solid results, especially in handling noisy data. However, its performance was slightly worse than XGBoost due to overfitting on certain subsets of the data.

Linear Regression

As expected, Linear Regression performed the worst due to the non-linear nature of the relationships between the features. The model failed to capture the underlying patterns in the data, resulting in high RMSE.

Decision Tree

Decision Tree provided interpretable results but suffered from overfitting, leading to higher errors in the testing set. Although it was computationally efficient, its predictions were not as accurate as those from ensemble methods like Random Forest or XGBoost.

Evaluation Metrics:

Models were evaluated based on several metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R² Score, and cross-validation accuracy.

Machine Learning Algorithms	Training Result	Testing Result	R^2	RMSE	MSE	MAE	Average Cross Validation
Linear Regression	77.75%	67.73%	73.17%	164.48	27325.16	93.91	72.23%
Decision Tree Regressor	77.77%	67.74%	75.72%	164.43	27285.81	93.96	73.89%
KNN Regressor	89.99%	71.51%	80.05%	85.54	14330.83	46.28	81.83%
Random Forest Regression	86.54%	74.53%	81.32%	110.05	14740.38	61.96	81.02%
SVM Regression	65.61%	65.79%	65.57%	185.56	34433.39	101.87	65.71%
XGBoost Regressor	79.14%	72.41%	77.57%	155.33	24253.91	92.26	76.17%
Gradient Boosted Model	79.88%	72.64%	76.29%	153.77	23798.38	91.51	76.92%

In the table mentioned above, we can see that Support Vector Machine Regression (SVM) has a training accuracy of 65.61% which is the lowest among all machine learning models. On the other hand, K-Nearest Neighbor Regressor holds 89.99% accuracy, being the highest among the training set. If we consider testing accuracy, we can clearly see SVM Regressor has the lowest testing accuracy of 65.79%. Random Forest Regressor holds the highest testing accuracy of 74.53%.

Performance Comparison:

- The XGBoost model outperformed other algorithms with the lowest MAE and RMSE, as well as the highest R² score. This can be attributed to XGBoost's ability to handle complex feature interactions and its efficiency in model tuning.
- Random Forest also provided strong results due to its ensemble nature, reducing overfitting and capturing non-linear relationships in the data.
- KNN and Decision Tree performed moderately well but struggled with high-dimensional relationships in the data.

- SVM showed strong performance but required longer training times and was sensitive to parameter tuning.
- Linear Regression had the lowest performance, mainly due to the complexity of the data which cannot be modeled accurately with a linear approach.

Feature Importance Analysis:

From the feature importance analysis, it was observed that Radiation had the strongest correlation with Temperature and Daytime Duration. Wind Speed and Direction had minimal impact on the models' predictive capabilities, while Humidity showed moderate significance.

VI. CONCLUSION & FUTURE WORK

This study demonstrates the effectiveness of various machine learning algorithms in predicting solar radiation using meteorological data. XGBoost and Random Forest were identified as the best-performing models, largely due to their ability to handle non-linearity and feature interactions. However, simpler models like Linear Regression and KNN were less effective, highlighting the importance of model complexity in solar radiation prediction tasks. In future work, the inclusion of more features, such as cloud cover and air quality, could improve model accuracy. Furthermore, deep learning techniques could be explored to enhance the prediction capabilities for large-scale renewable energy applications.

REFERENCES

[1]. IEA: World Energy Outlook 2009. International Energy Agency Publications (2008).

[2]. Benghanem, M., Maafi, A.: Data Acquisition System for Photovoltaic Systems Performance Monitoring. IEEE Transactions on Instrumentation and Measurement 47, 30–33 (1998).

[3]. Forero, N., Hernández, J., Gordillo, G.: Development of a monitoring system for a PV solar plant. Energy Conversion and Management 47(15-16), 2329–2336 (2006).

[4]. Jorge, A., Guerreiro, J., Pereira, P., Martins, J., Gomes, L.: Energy Consumption Monitoring System for Large Complexes. In: Camarinha-Matos, L.M., Pereira, P., Ribeiro, L. (eds.) DoCEIS 2010. IFIP AICT, vol. 314, pp. 22–24. Springer, Heidelberg (2010).