GRADUATE PROGRAMS INSTITUTE

A SYSTEM TO PREDICT SOLAR RADIATION BY USING METEOROLOGICAL AND GEOGRAPHICAL INPUTS

Safi CENGİZ

MASTER'S TERM PROJECT

İstanbul, May 2022

A SYSTEM TO PREDICT SOLAR RADIATION BY USING METEOROLOGICAL AND GEOGRAPHICAL INPUTS

Safi CENGİZ

MASTER'S TERM PROJECT
COMPUTER ENGINEERING DEPARTMENT
COMPUTER ENGINEERING MASTER'S PROGRAM WITHOUT THESIS

MASTER'S PROJECT ADVISOR
Asst. Prof. Burçin KÜLAHÇIOĞLU

TABLE OF CONTENTS

ABSTRACT

A SYSTEM TO PREDICT SOLAR RADIATION BY USING METEOROLOGICAL AND GEOGRAPHICAL INPUTS

In these days of global energy crisis, solar energy is emerging as a viable alternative to fossil fuels. Even if they are closer to the Equator or receive more sunlight, developing or underdeveloped countries cannot benefit as much from solar energy as developed countries. Solar energy panels are installed and adjusted according to the angle of direct normal irradiance to achieve efficiency. However, direct normal irradiance is measured using a pyrheliometer. The pyrheliometer is a sensor that requires ongoing maintenance (calibration) and is more expensive than other sensors. This study attempted to predict direct normal radiation with machine learning models on the dataset created with meteorological data obtained from other sensors without using a pyrheliometer.

Keywords: Solar Irradiance Forecast, Meteorological Data, Machine Learning

ÖZET

METEOROLOJİK VE COĞRAFİK GİRDİLER KULLANARAK GÜNEŞ RADYASYONUNU TAHMİN EDEN SİSTEM

Küresel enerji krizinin yaşandığı bu günlerde, güneş enerjisi fosil yakıtlara bir alternatif olarak ortaya çıkmaktadır. Ekvatora daha yakın olsalar veya daha fazla güneş ışığı alsalar bile, gelişmekte olan veya az gelişmiş ülkeler güneş enerjisinden gelişmiş ülkeler kadar yararlanamamaktadır. Verimliliği sağlamak için güneş enerjisi panelleri doğrudan normal ışınım açısına göre kurulur ve ayarlanır. Bununla birlikte, doğrudan normal ışınım bir pirhelyometre kullanılarak ölçülür. Pirhelyometre, sürekli bakım (kalibrasyon) gerektiren ve diğer sensörlerden daha pahalı olan bir sensördür. Bu çalışmada, Pirhelyometre kullanılmadan diğer sensörlerden elde edilen meteorolojik verilerle oluşturulmuş veri seti üzerinde, makine öğrenimi modelleri ile doğrusal normal ışınım tahmin edilmeye çalışılmıştır.

Anahtar Sözcükler: Güneş Işınımı Tahmini, Meteorolojik Veri, Makine Öğrenmesi

ABBREVIATIONS

GHI : Global Horizontal Irradiance

DHI : Diffused Horizontal Irradiance

DNI : Direct Normal Irradiance

RMSE : Root Mean Squared Error

R2 : R-Squared

LIST OF FIGURES

1 INTRODUCTION

The sun is one of the primary sources of energy and the source of life. Also, it is a direct or indirect source of renewable energy systems, such as solar thermal power plants, which directly transform sun irradiation into useable electric energy. The cube of the distance from the sun lowers solar irradiation. Solar irradiance outside the earth's atmosphere varies between 1325 W/m2 and 1420 W/m2 as a result of the varying distance between the earth and the sun during the year (Quaschning, 2022). Numerous academic disciplines examine the sun, such as gravitational physics, agriculture, and solar energy. Predicting solar radiation might be used in agriculture, photovoltaic plant development, and photovoltaic electricity in an effort to lower the cost of solar power plants.

Numerous research have been conducted about solar irradiance. Numerous opinions on the subject have been expressed. (Al-Taani, H., & Arabasi, S., 2018) prioritised predictability above accuracy and created predictions without the usage of sensors using a smartphone-assisted setup. Yan, Shen, and others for accurate forecasting, deep learning approaches like ResNet and recurrent neural network (RNN) were used. Using an artificial neural network, (Yan, K. Shen &H., Wang & L., Zhou & H., Xu, M., & Mo, Y., 2020) (Pazikadin, A. R., Rifai, D., Ali, K., Malik, M. Z., Abdalla, A. N., & Faraj, M. A., 2020) also predicted DNI. Some research concentrated on forecasting both Global Horizontal Irradiation and Direct Normal Irradiation (Trapero, J. R., Kourentzes, N., & Martin, A., 2015) and suggested a novel time-series model known as Dynamic Harmonic Regression. (Yang, D., Jirutitijaroen, P., & Walsh, W. M., 2012) used an Auto-Regressive Integrated Moving Average (ARIMA) model that incorporates components of cloud cover effects. Li et al. employed Support vector machine regression for short-term solar irradiance forecasting. (Li, J., Ward, J. K., Tong, J., Collins, L., & Platt, G., 2016). Lastly Cao and Xchun Lin mapped live functions and projected hourly global sun radiation using Wavelet neural networks (WNN) (Cao, J., & Lin, X., 2008). This paper focuses on predicting Direct Normal Irradiance because measuring it with sensors is costly and labouring. Predictive DNI aims to lower costs and make it more accessible to everyone. Pyrheliometer sensors are extremely sensitive, and dust can cause problems with data collection. Furthermore, the sensors must be cleaned and maintained on a daily basis, and they must be replaced every three months in extreme climates such as deserts or extremely hot conditions (Perez-Astudillo, D., & Bachour, D. D. N. I., 2014).

2 SUN IRRADIANCE

2.1 SUN

The Sun is a 4.5 billion-year-old star that is a hot, incandescent ball of hydrogen and helium at the centre of our solar system. The Sun is approximately 150 million kilometres away from Earth, and life on our planet would not exist without its energy. Our solar system's greatest object is the sun. The Sun's volume is equivalent to 1.3 million Earths. The solar system is held together by gravity, which keeps everything in orbit around it, from the largest planets to the tiniest meteorite bits. The Sun's core is the hottest place on the planet, with temperatures exceeding 15 million degrees Celsius. The photosphere, which we call the Sun's surface, is 5,500 degrees C lower than the core. The corona, the Sun's outer atmosphere, gets hotter as it reaches further away from the surface. The corona is significantly hotter than the photosphere, reaching temperatures of up to 2 million degrees Celsius. (Amanda Barnett, 2022)

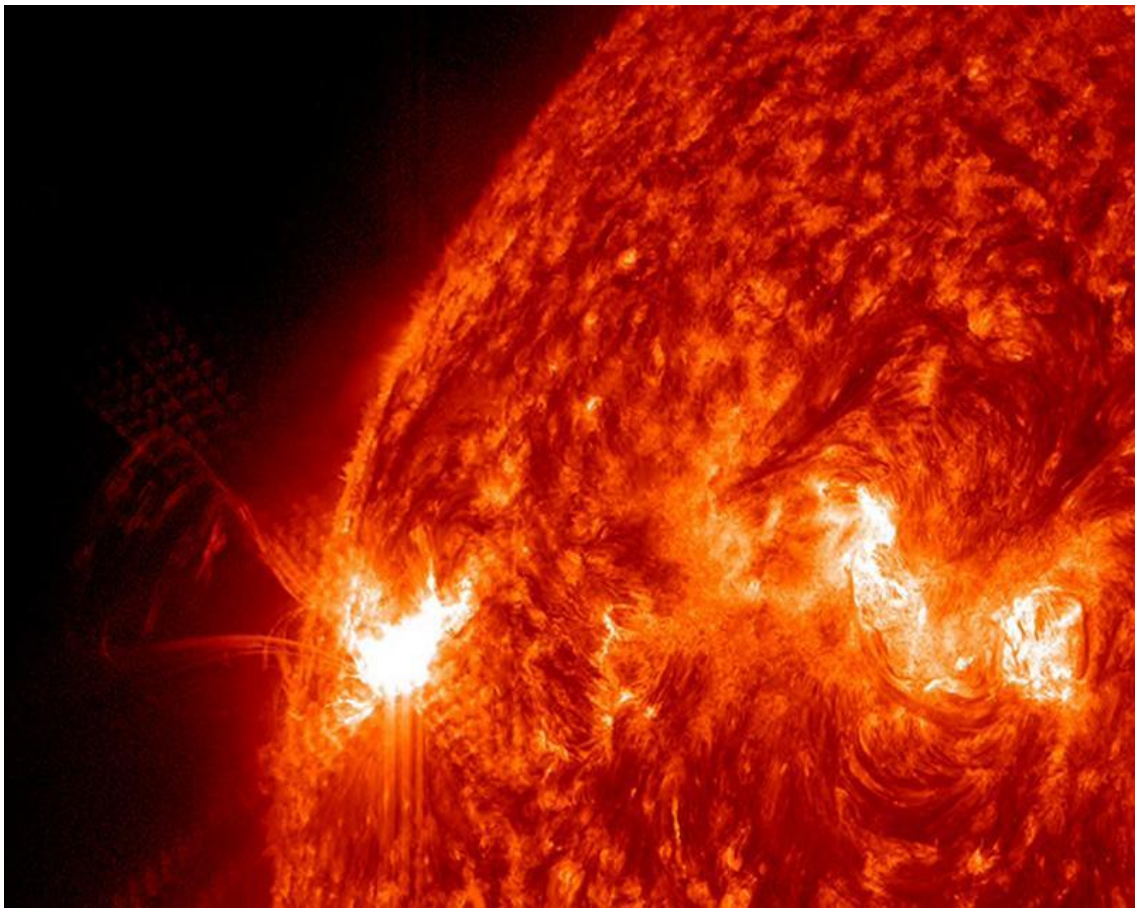

Figure 2.1: Sun surface , Taken from NASA Photo Gallery (Amanda Barnett, 2022)

The temperature in the Sun's core is estimated to be 15 million degrees Celsius. This is the temperature of a hydrogen bomb exploding, which is hot enough to drive thermonuclear reactions that convert hydrogen atoms into helium, thereby powering the Sun. In this manner, the Sun consumes approximately 5 million tons of nuclear hydrogen fuel every second. (A Meeting with the Universe, 2022)

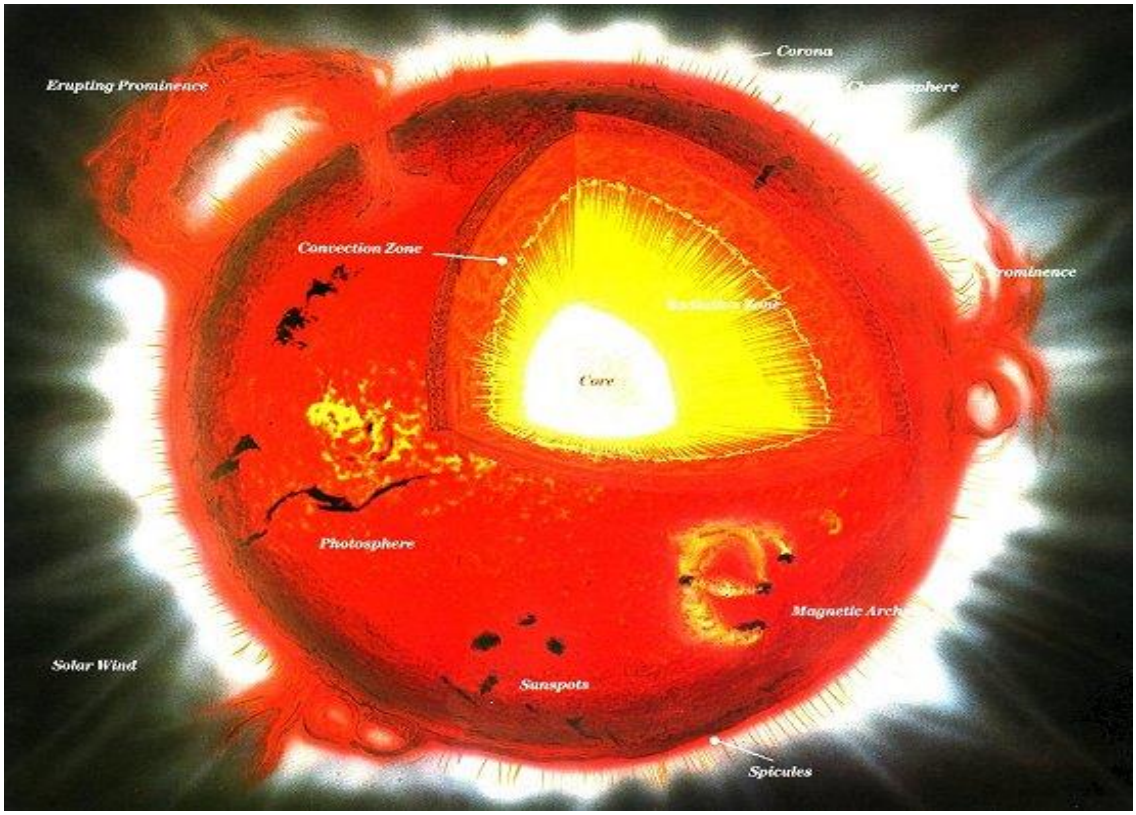

Figure 2.2: Sun Drawing from NASA Gallery (A Meeting with the Universe, 2022)

2.2 IRRADIANCE

Heat functions as radiation in space, sending an infrared energy wave from warmer to cooler objects. Radiation waves cause molecules to heat up when they come into touch with them. This is how heat travels from the sun to Earth, although radiation only heats molecules and stuff in its straight route. Everything else remains cold. Compare this to Earth, where the air surrounding you stays warm even when you're in the shade, and it's

even dark at night in some seasons. This is due to the fact that heat passes through our planet in three routes from space: conduction, convection, and radiation. When the sun's radiation strikes and heats molecules in our atmosphere, the excess energy is passed on to the molecules around them. These molecules smash, heating their surroundings. Conduction is a chain reaction that heats locations outside the sun's path by transferring heat from molecule to molecule. However, because space is a vacuum, it is essentially empty. Because the gas molecules in space are so far apart, they cannot collide and heat up. That is, even if the sun heats them with infrared rays, the heat is not transferred through conduction. Convection, a type of heat transfer that occurs in the presence of gravity, is also crucial in transporting heat to Earth, although it does not occur in zero g-space. (Coffey, 2022)

2.3 SOLAR IRRADIANCE EFFECTS ON LIFE

Solar radiation is the energy released by the Sun and transmitted as electromagnetic waves across space. This energy, emitted by the sun's surface as a result of chemical reactions, influences atmospheric and climatological processes. It is the cause of plant photosynthesis, which also causes the planet to maintain a temperature suitable for life and wind creation, which is required for wind power generation. Wind, as known, is another sustainable energy source. The Sun emits energy in the form of short-wave radiation, which is diminished by clouds in the atmosphere and absorbed by gas molecules or suspended particles. Solar radiation reaches the marine and continental land surfaces after travelling through the atmosphere, where it is reflected or absorbed. Lastly, the surface reflects it back into space as long-wave radiation. (Iberdrola, 2022)

These descriptions of these concerns are intended to clarify the sorts of solar radiation that will be detailed shortly. There is just one sort of radiation in space, unlike on Earth.

2.4 SOLAR IRRADIANCE TYPES

There are 6 different solar irradiance types, which are Total Solar Irradiance, Direct Normal Irradiance, Diffuse Horizontal Irradiance, Global Horizontal Irradiance, Global Tilted Irradiance, Global Normal Irradiance. This study is related only 3 of them.

Direct Normal Irradiance (DNI) is the quantity of light that strikes the surface perpendicularly. The surface in this case indicates the ground or anything parallel to the earth. This sort of irradiance is caused by rays that travel in a straight line from the sun's current position in the sky. Solar collectors and panels maximise this DNI by tilting or rotating with the angle of the sun.

Diffused Horizontal Irradiance (DHI) is solar radiation that has been diffused by clouds and particles in the atmosphere and comes equally from all directions.

Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received from above by a horizontal (parallel) surface to the ground.

Global Horizontal Irradiance (GHI) = Direct Normal Irradiance (DNI)* cos(solar zenith angle) + Diffused Horizontal Irradiance (DHI) (Singh, 2022)

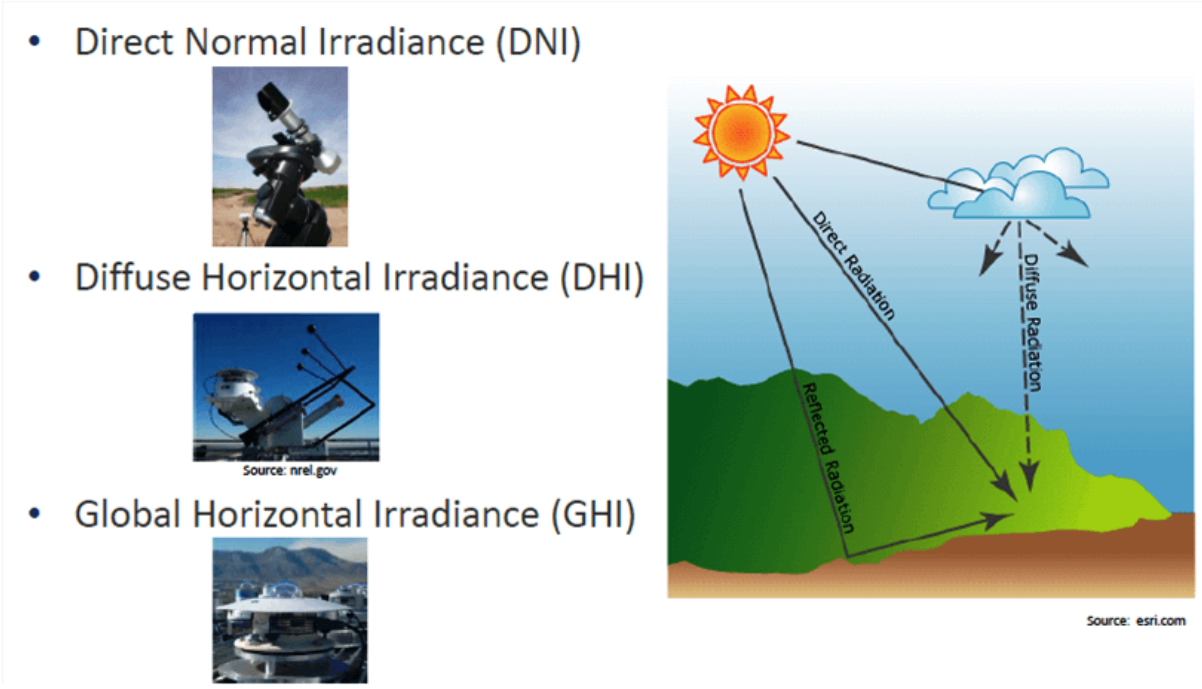

Figure 2.3: Different Sensors for per Irradiance Type (Singh, 2022)

2.5 SENSOR TYPES TO MEASURE IRRADIANCE

DNI can only be measured with a pyrheliometer. The difference between a pyranometer and a pyrheliometer is that the former measures global solar radiation while the latter measures direct sun irradiance. Pyranometers monitor radiation coming from all directions by looking up at the sky. Pyrheliometers observe the sun directly and only measure radiation from one direction. Because they are more complex and require more maintenance, pyrheliometers are more expensive than pyranometers. And this is the focus of our research: predicting DNI because measurement is costly and requires frequent maintenance. The sensor is calibrated during maintenance. Our research is being conducted to predict DNI globally without the use of an expensive Pyrheliometer that measures DNI. (Pyranometer vs Pyrheliometer: What's the Difference?, 2022)

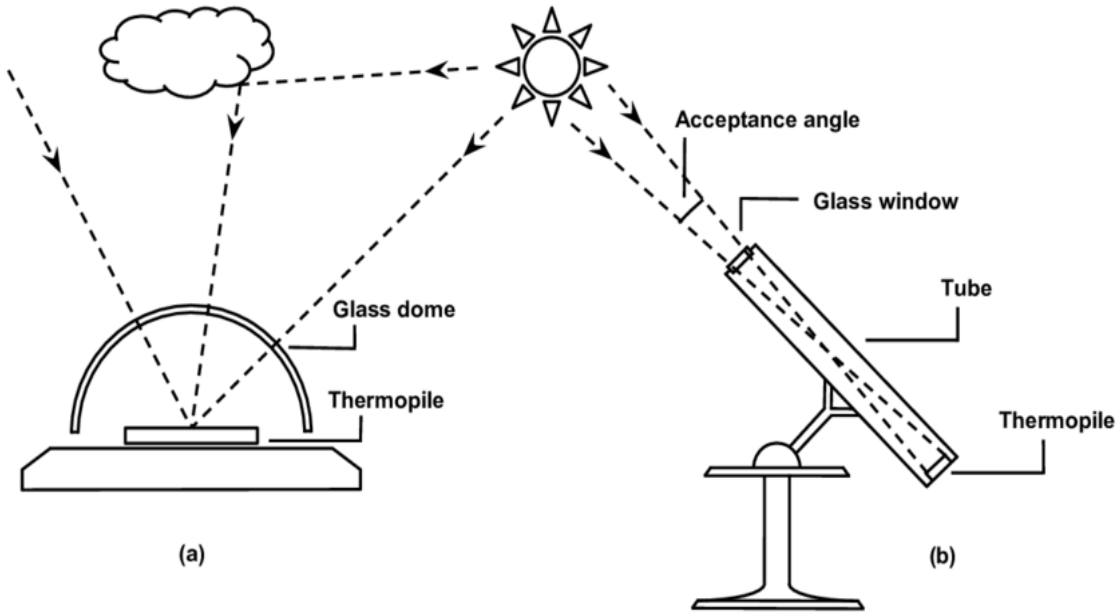

Figure 2.4: Difference between (a)Pyranometer and (b) Pyrheliometer (sevensensor, 2022)

3 METHODOLOGY

3.1 DATA COLLECTION

After discussing the principles of radiation and its path from the Sun to Earth, we concentrate on the specifics of this study. The dataset includes numerous types of solar radiation as well as numerous meteorological units. Let us examine the units of measurement for these characteristics and the creation of the dataset. The dataset was received from the National Renewable Energy Laboratory of the US Department of Energy (NREL). The NREL tool NRSB Data Viewer enables users to choose and query geographical inputs like longitude and latitude, as well as data series including years, months, and frequency of data. The dataset was queried using the coordinates 41.09, 29.1, and data from the closest meteorological weather station. The components of the dataset were measured by a barometer, thermometer, psychrometer, hygrograph, anemometer, pyranometer, and ultimately a pyrheliometer put on the meteorological station. (mgm, 2022)

Figure 3.1: Automatic Weather Observation Station (mgm, 2022)

The dataset consists of hourly data at starting 00:30, 01.01.2017 to 23:30, 31.12.2017, Latitude: 41.09, Longitude : 29.1, Beykoz/ İstanbul/ Türkiye respectively district, city and country name.

3.2 DATASET

There are 8760 rows and 23 columns in the raw data set. After passing it a data frame to ‘df’ variable, pandas data frame was used as a read csv function and the 'Unnamed: 23' variable created by pandas was discarded due to misinterpretation.

Table 3.1: Dataset Features, Their Units, and Meanings

Dataset Features			
Numerical Values		**Categorical Values**	
Features	**Unit**	**Features**	**Meaning**
DNI	w/m²	Cloud Type 0	Clear
DHI	w/m²	Cloud Type 1	Probably Clear
GHI	w/m²	Cloud Type 2	Fog
Dew points	C(celsius)	Cloud Type 3	Water
Temperature	C(celsius)	Cloud Type 4	Super-Cooled Water
Pressure	mbar	Cloud Type 5	Mixed
Relative Humidity	(percentage)%	Cloud Type 6	Opaque Ice
Precipitable Water	cm	Cloud Type 7	Cirrus
Wind Direction	Degree°	Cloud Type 8	Overlapping
Wind Speed	m/s	Cloud Type 9	Overshooting
Surface Albedo	ratio	Cloud Type 10	Unknown
Clearsky DNI	w/m²	Cloud Type 11	Dust
Clearsky GHI	w/m²	Cloud Type 12	Smoke
Clearsky DHI	w/m²	Fill Flag 0	N/A
Solar Zenith Angle	Degree°	Fill Flag 1	Missing Image
		Fill Flag 2	Low irradiance
		Fill Flag 3	Exceeds Clearsky
		Fill Flag 4	Missing Cloud Properties
		Fill Flag 5	Rayleigh Violation

3.3 PRE-PROCESSING

All rows having a DNI values of zero are eliminated since the purpose of this research is to forecast DNI levels. If the 0 values were not removed, the night and day values would be confused. As is well-known, there is no daylight at night. Keeping these values also reduces the reliability of the model. Then, a new date column was created by merging five existing date columns: Year, Month, Day, Hour, and Minute. The index of the date-time column is then set in order to generate meaningful visualisations. Also available is Clearsky DNI column, which is measured by the same sensor as DNI. Therefore, the decision was made to eliminate it due to its nearly one-to-one correlation with DNI. It could threaten the quality of our model. The Solar Zenith Angle and DHI values also have been eliminated from the GHI equation, as previously demonstrated. Otherwise, every element of the equation would be known, and there would be no need to predict results.

The dataset now contains 3167 rows and 13 columns due to these modifications. In this dataset, there are no missing values.

3.4 VISUALIZATION

DNI is the target value, and the initial visual representation aims to convey a first impression to the reader.

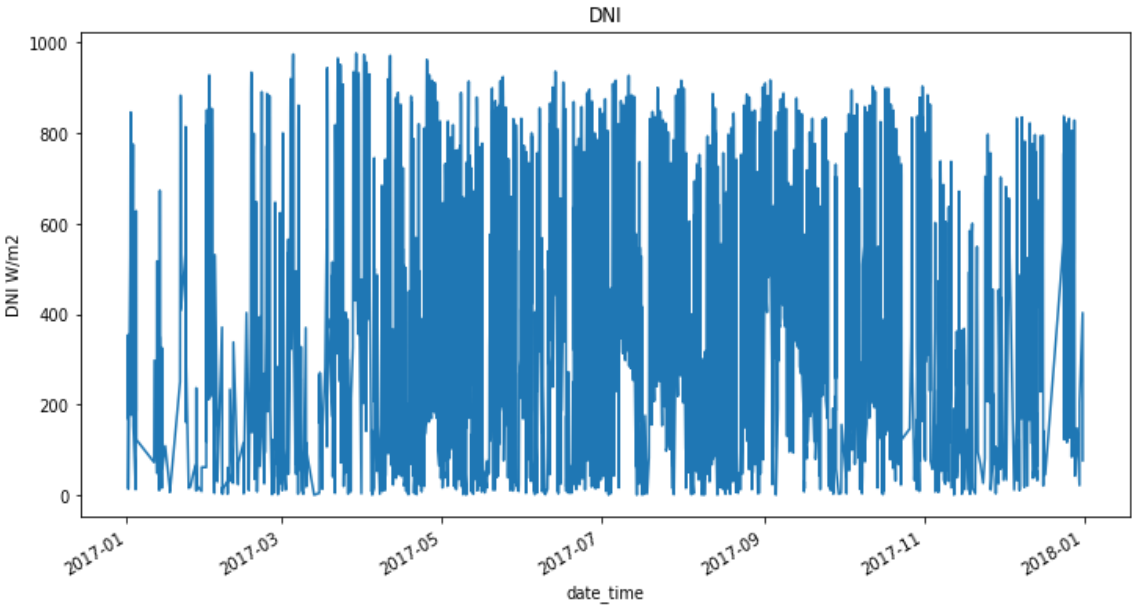

Figure 3.2: Target Value DNI Pattern

A second visualization is using a version of the dot pattern with a different colour for the purpose of making the visual more apparent. As can be seen in the graph below, the density of DNI is significantly higher in certain times of the year compared to other times of the year.

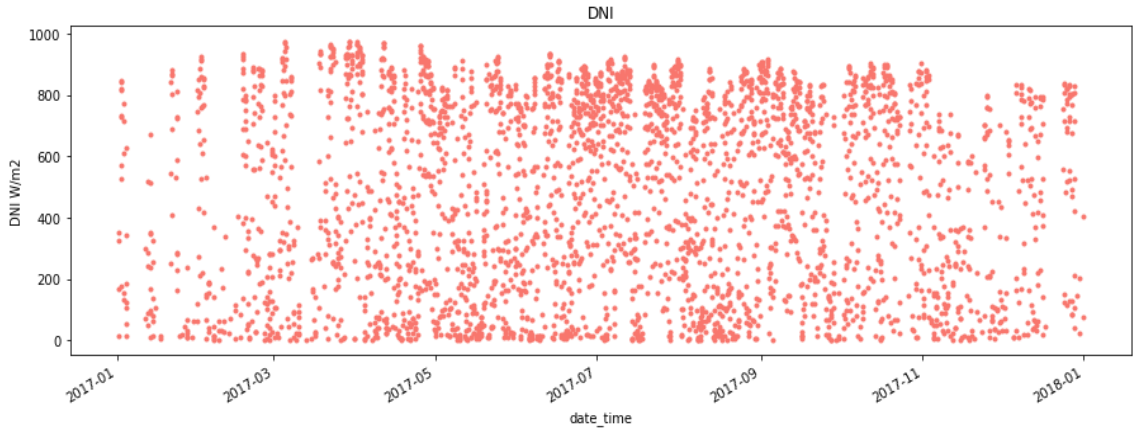

Figure 3.3: Target Value DNI with Dot Style and Red Colour

3.5 MISSING VALUES

If there are any missing values in the dataset, we will replace them with mean values. As a result, the overall distribution will be unaffected.

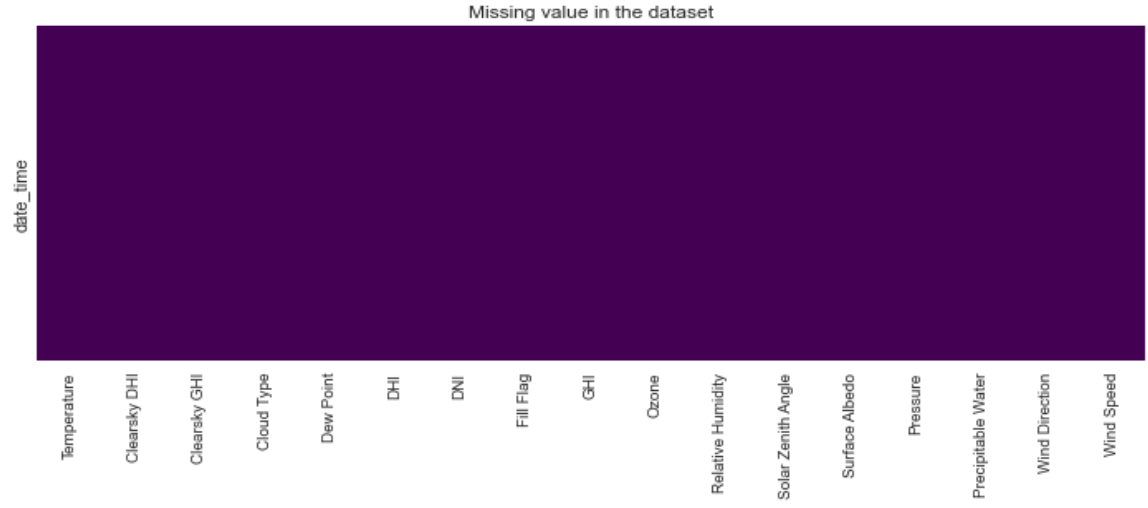

Figure 3.4: Missing Value Plot – Empty Graph Means There is No Null Value.

3.6 CORRELATION BETWEEN FEATURES

A linear relationship between variables is known as correlation. It provides an overview of feature selection. The correlation coefficients range from -1 to 1. If there is positive correlation, it will be between 0 and 1, whereas if it is negative correlation, it will be between -1 and 0. A strong linear relationship exists when a variable is close to 1 or -1. The heatmap below represents the correlation between all features in the entire dataset.

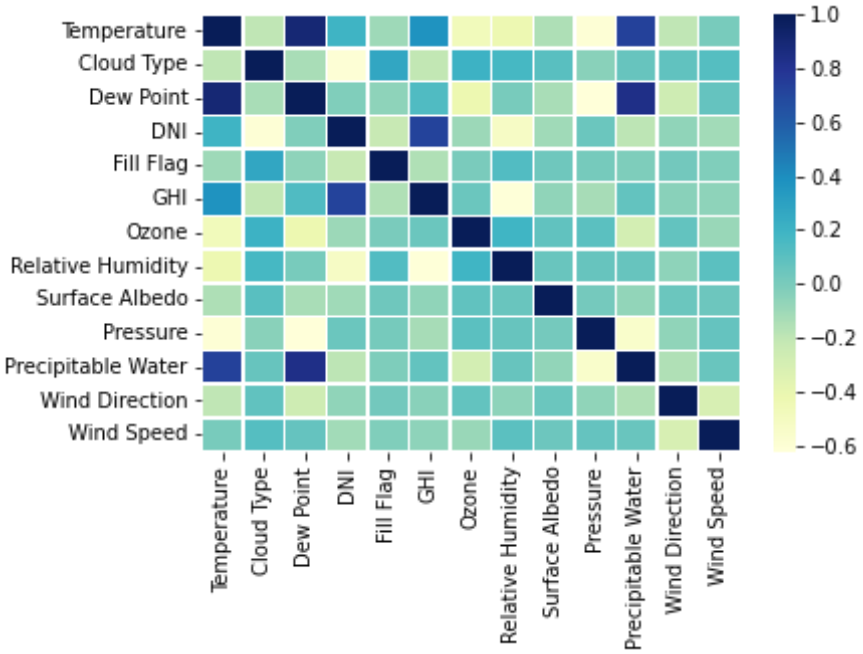

Figure 3.5: Correlation Heatmap

The correlation between all features is displayed in Figure 3.5 and Figure 3.6 depicts a correlation bar plot in which features are only compared to the target value DNI.

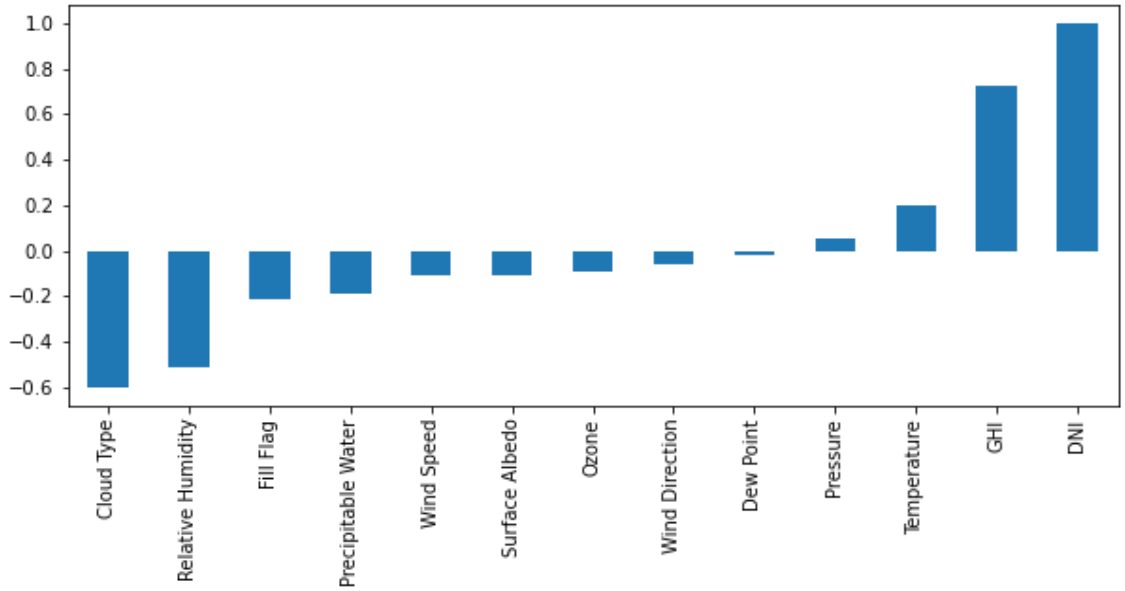

Figure 3.6: Correlation with Target Value DNI

Using the jointplot created in Figure 3.7, we analyze the relationship between the two variables. The study has previously described a mathematical equation that can be used to determine the value of the GHI, and this equation includes the target variable DNI.

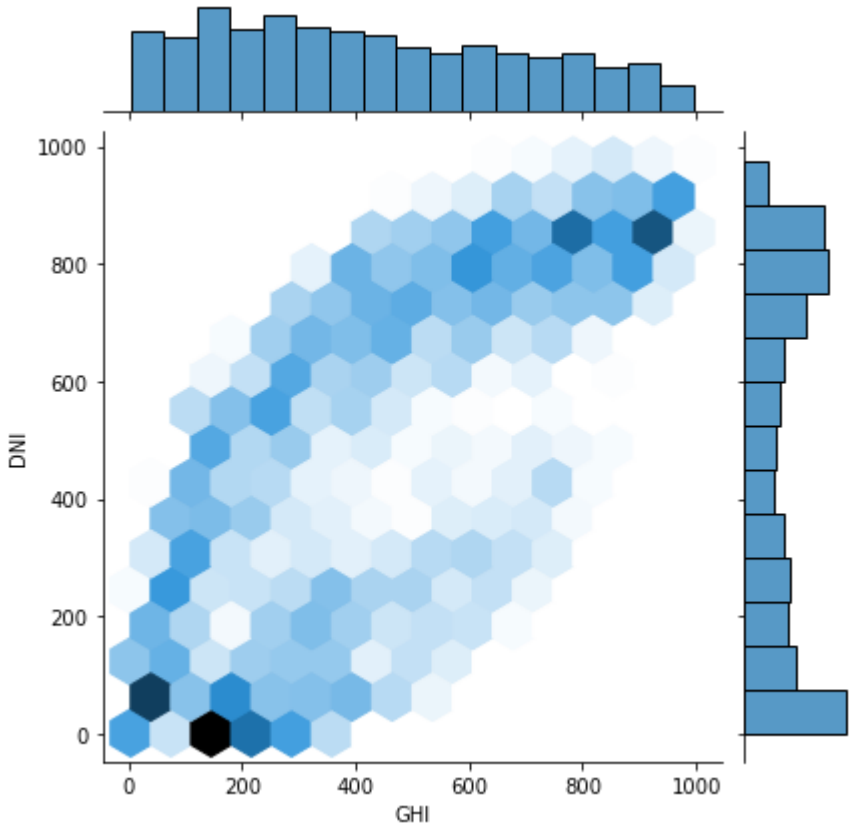

Figure 3.7: Hex Plot Show DNI and GHI

4 MODEL

After data sets have been produced, the modelling phase may begin. The dataset was divided into two parts: 5 percent validation data and 95 percent dataset, which would be divided into 30 percent test and 70 percent train data. The trained model is then evaluated using thirty percent of the data to confirm its success. Finally, we can use the finalised model to verify data and quantify model success using data that the trained model has never seen. X and Y variables are often separated into features and target values. Use X train, X test, y train, and y test for train and test data that has been divided. The DNI column, whose value is y, is our target. In summary, the model will be fed and created from input as train data, and its accuracy will be confirmed by applying it to test data. In actuality, the right phrase is regression since this is a regression accuracy. Root Mean Squared Error (RMSE), R-Squared (R2), Mean Absolute Error (MAE), and Mean Squared Error (MSE) are the performance measures for regression models.

The standard deviation of the estimation errors is the RMSE (residues). The RMSE is a measure of how widespread these residues are; residuals are a measure of how far the regression line is from the data points. As a result, it indicates how dense the data is around the best-fitting line. The model's success is also determined by the RMSE. Based on the RMSE score, the best model will be chosen. Because DNI values fluctuate greatly throughout the year and throughout the day (morning, evening, and noon).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y}i)^2} \quad (1)$$

$$RMSE = \sqrt{MSE} \quad (2)$$

In a regression model, R^2 is the proportion of variance explained by independent variables for a target variable. This metric will be used to show how well the selected model fits the data.

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

ŷ = predict value, y=test value, ȳ= mean of all test values

K-fold cross validation, piecewise resampling of the dataset, and evaluation of our model It will be tested on 10 different folds independently if k-fold is set to 10. Cross validation is used to avoid overfitting, which is the process of memorising the target (dependent) value for each independent value. Two categorical values are already label encoded in the dataset. Table -1 shows what numbers correspond to labels. There are 13 cloud types and 6 fill flags to choose from. However, there are only 9 different cloud type values and 4 different fill flag values after data cleaning. This dataset will be encoded with a one-hot encoder; otherwise, the model will believe there is a hierarchy between these two values. The dataset expanded to 27 columns after one-hot encoder used. The cloud type column was removed, and nine new columns were created, with names ranging from cloud type 0 to cloud type 9. Consider that the cloud type value of one row was 9. Following the operation, only the cloud type 9 column will have a value of 1, while the remaining eight columns will have a value of 0.

4.1 BUILDING THE MODEL

Pycaret was used to normalize the dataset, compare models, and tune hyperparameters for this project. Only the top three models will be shown to the improvement in the model score. In the first step, no feature selection algorithm or scaling, such as the standard scale, is applied to the dataset.

Table 4.1: Top 3 Successful Models

Model	MAE	MSE	RMSE	R2	TT(sec)
CatBoost Regressor	36.2402	2737.3051	51.9255	0.9708	0.4340
Light Gradient Boosting Machine	40.7388	3405.4235	58.0337	0.9637	0.1140
Extra Trees Regressor	40.7103	3445.4317	58.2483	0.9632	0.0680

After applying normalization to the dataset, the procedure was repeated. As seen in the table below, each variable has a distinct range of values, with distinct minimum and maximum values, mean, and standard deviation.

	Temperature	Cloud Type	Dew Point	DNI	Fill Flag	GHI	Ozone	Relative Humidity	Surface Albedo	Pressure	Precipitable Water
count	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000	3167.000000
mean	19.187528	1.431955	12.519230	473.062835	0.173350	428.693716	0.316864	66.489795	0.145122	1012.403221	2.108810
std	6.791943	2.271600	5.849655	308.579565	0.725371	268.163288	0.032649	11.604468	0.080748	5.258753	0.852973
min	0.200000	0.000000	-5.700000	1.000000	0.000000	4.000000	0.256000	24.960000	0.100000	998.000000	0.400000
25%	14.000000	0.000000	8.200000	171.000000	0.000000	201.000000	0.290000	58.340000	0.130000	1009.000000	1.500000
50%	19.700000	0.000000	13.000000	502.000000	0.000000	392.000000	0.311000	66.070000	0.140000	1012.000000	2.000000
75%	25.000000	2.000000	17.500000	768.000000	0.000000	644.000000	0.343000	74.860000	0.140000	1015.000000	2.700000
max	34.200000	12.000000	24.700000	976.000000	5.000000	996.000000	0.418000	98.930000	0.870000	1030.000000	4.400000

Figure 4.1: Sample of the Dataset' s Description

Normalization is the process of establishing a structural link between features in a dataset to reduce data redundancy and improve data integrity. After the normalization process, we examine the top three models again.

Table 4.2: Top 3 Successful Model After Normalization Process

Model	MAE	MSE	RMSE	R2	TT(sec)
CatBoost Regressor	36.2342	2736.9175	51.9218	0.9709	1.521
Light Gradient Boosting Machine	40.7843	3405.5338	58.0105	0.9637	0.096
Extra Trees Regressor	40.7103	3445.4317	58.2483	0.9632	0.567

There is no significant improvement on the CatBoost Regressor model, only the Light Gradient Boosting Machine model reduces errors, but these errors are not main indicators.

Then we observe how the top model performs.

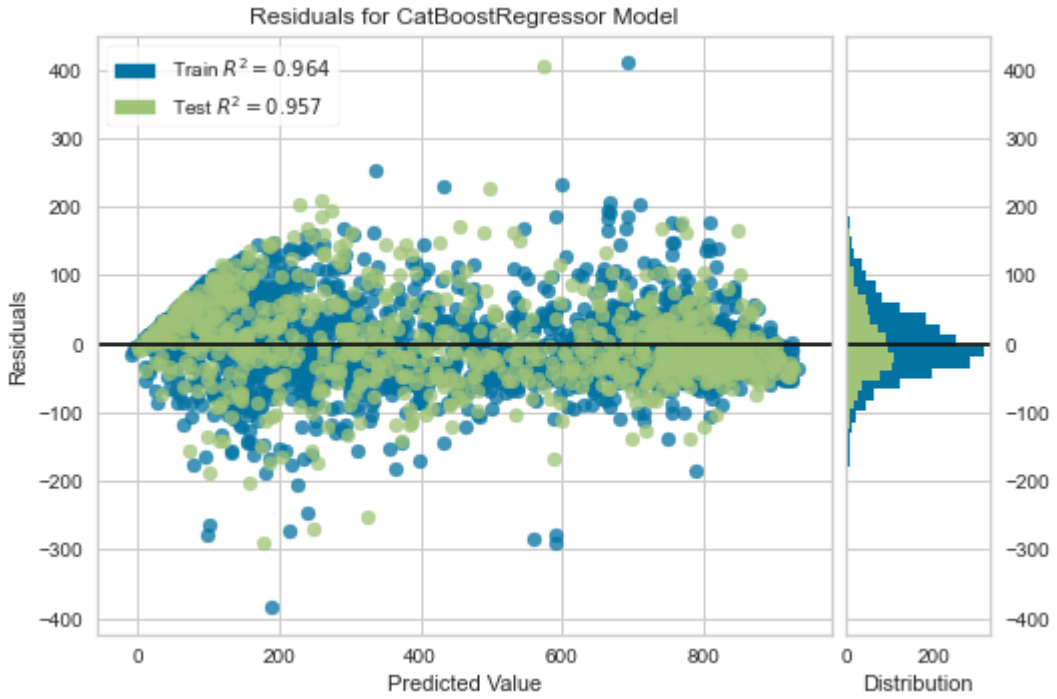

Figure 4.2: Residuals of CatBoostRegressor Model

As a result, we must modify the model's hypertuning, as it is undesirable for the model to search for outliers to extreme depths.

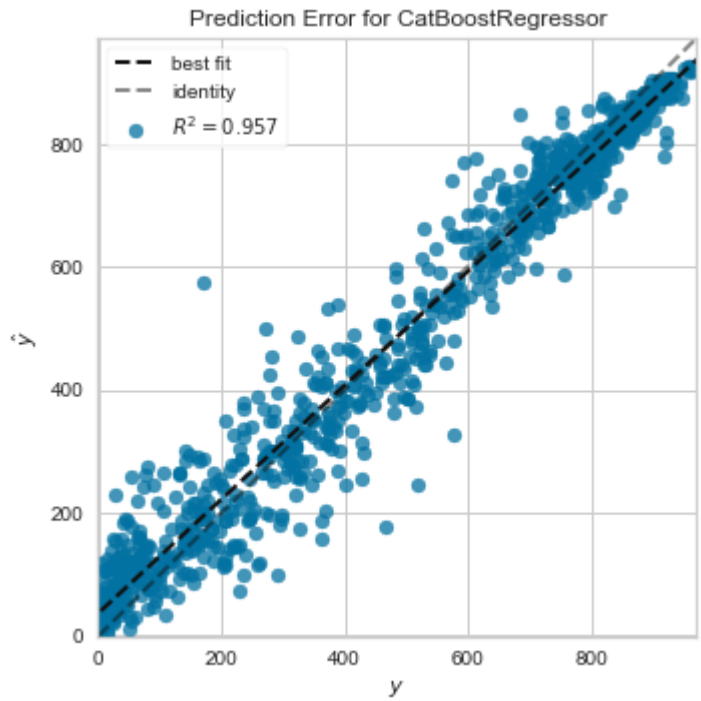

Figure 4.3: Errors of CatBoostRegressor Model

If the dots stay on a straight line, the predicted and observed values are the same. The greater the distance, the more errors appear.

The evolution of the model is represented in the graph below. These metrics are crucial for understanding model performance. Cross-validation is necessary for the model to prevent memorising the results. In order to ensure the reliability of the model it should also be tested on data that is new. (validation dataset).

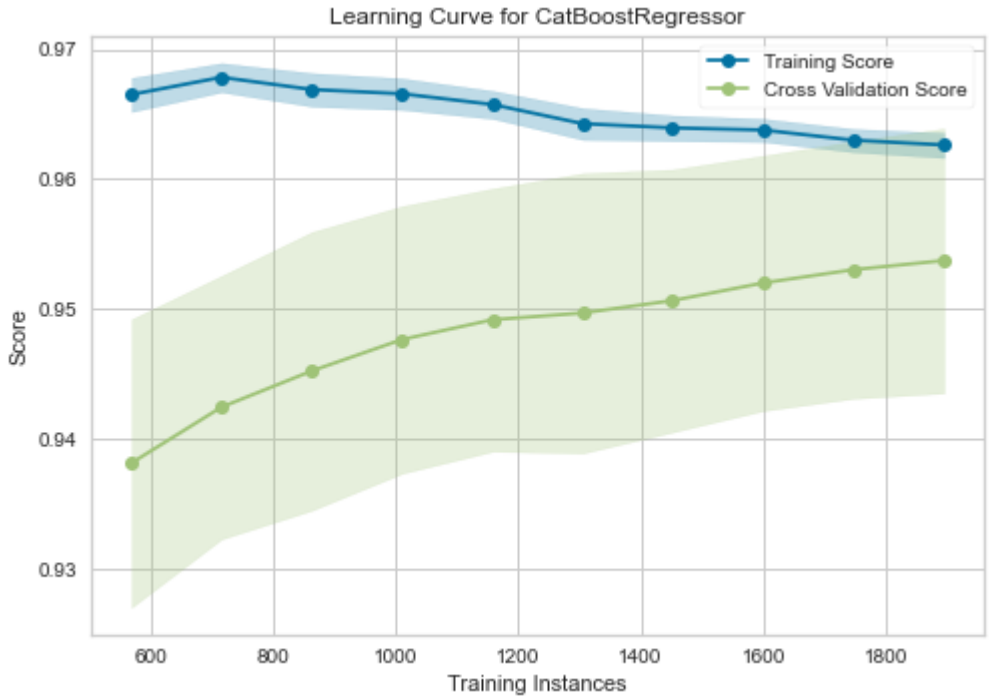

Figure 4.4: Learning Curve

Lastly, we consider the importance of features in the model. We compare the importance of these features to the correlation plots in Figure 3.5 and Figure 3.6.

Feature importance is useful for determining which features will be kept and which will be eliminated. The purpose of the project is to construct a model that can generate the most accurate prediction with the least amount of data in order to enhance the user experience. We attempted to reduce the number of features while maintaining the model's success. Numerous experiments led to the removal of the Fill Flag, Temperature, Wind Direction, Wind Speed, and Pressure columns. Their impact on the model is nearly nonexistent.

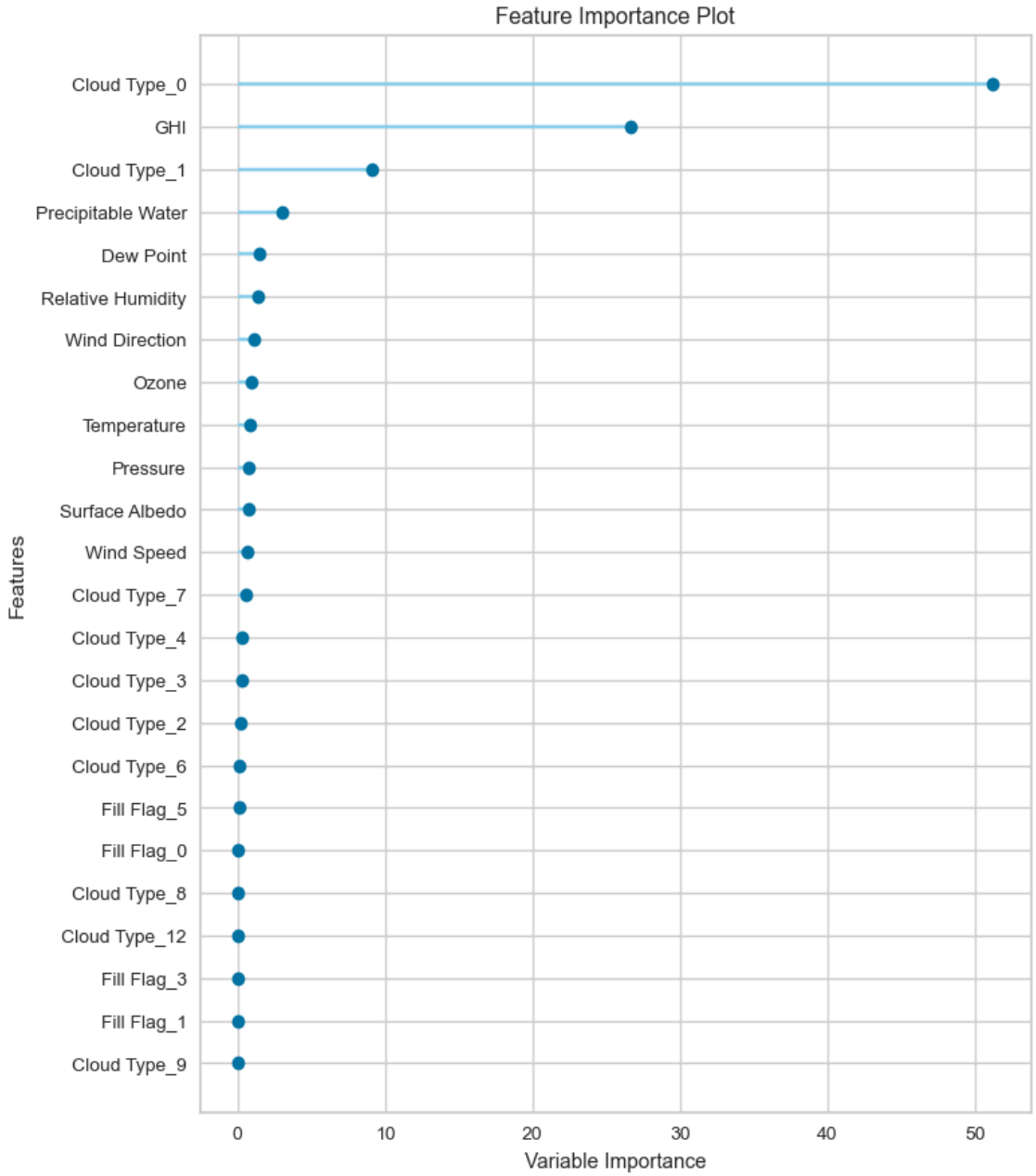

Figure 4.5: CatBoostRegressor Model's Feature Importance

The same steps were repeated after removing these columns from the dataset. Before the process, a fragment of validation data was divided. As a result, the model has never encountered these data before. Label is predicted variable (ŷ) while DNI is target variable (y).

Table 4.3: Fragment of Validation Dataset DNI Observed, Label Predicted

DNI	Label	DNI	Label
172	197.5754513	644	651.1688696
934	893.2224745	726	706.5690896
934	885.0462303	863	859.0745427
730	757.760343	377	369.0615352
878	861.5737746	766	747.705263
429	369.0592301	214	324.9054813
327	303.4730627	367	280.7366216

4.2 BEST MODEL

CatBoostRegressor was the most successful model by far. It was successful not only in RMSE, but also in 4 different error metrics which are RMSE, R2, MSE and, MAE. According to documentation of CatBoost; 'CatBoost uses gradient boosted decision trees as its foundation. A series of decision trees is built sequentially during training. Each successive tree is constructed with less loss than the previous trees.' (CatBoost, 2022)
The starting parameters determine the number of trees. Use the overfitting detector to avoid overfitting. Trees stop growing when it is activated. (CatBoost, 2022) Cross validation has been applied to avoid this issue.

After tuning; the model hyperparameters are; {evaluation_metric:RMSE, iterations: 250, estimator: SymmetricTree, depth:6, border_count:254, loss_function:RMSE, max_leaves:64}. These are the parameters obtained after performing a grid search on the model. The reason for doing so is that hypertuning parameters are used on trained models to generalise the model. Estimator is the type of estimator used by the model, SymmetricTree, depth is how far the model should seek, evaluation metric is the metric on which the model should focus, iterations is the number of iterations or samples to draw from the search space via the "n_iter" argument, leaves is the leaves of every tree, use all of the CPU cores by specifying the "n_jobs" argument as an integer with the number of cores in your system, in this study. (Brownlee, 2022)

CatBoost creates 1000 trees by default. To speed up the training, the number of iterations can be reduced. The learning rate must be increased as the number of iterations decreases. So iterations decreased to 250, to prevent overfitting.

4.3 DEPLOYMENT

After saving the model with the Python library Pickle, it is time to receive user input. Flask, a Python-based framework, was used to create this web application. jQuery, Bootstrap, HTML, and CSS were also used for structuring website. In brief, the user enters input, there is a one-second loading period, and then the model's predicted output value is displayed. When user clicks Predict DNI button, it executes getResult() function. The function getResult runs on flask interface, it gathers information from forms with POST method, variables created from forms that user entered. Then input list is created with these variables. Lastly saved model runs with .predict method and result returns as getResult() function's output. Values of predictions are presented as rounded numbers. Inputs are parsed as integers if the user enters words or symbols when a numeric value is expected.

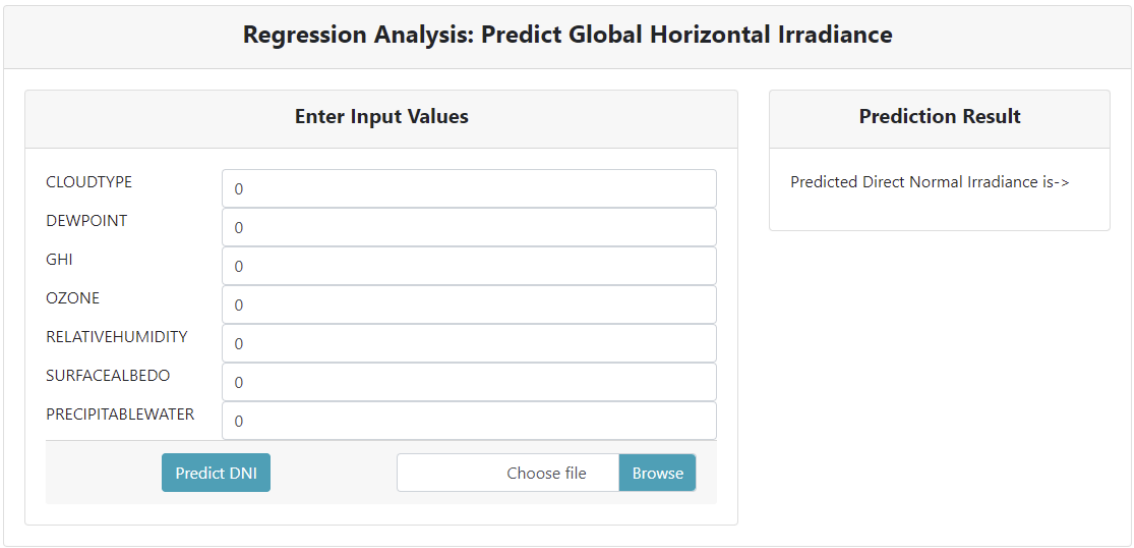

Figure 4.6: Interface of Flask Web-Application

5 CONCLUSION

The normalized dataset-trained models were only marginally successful than not normalized versions. In some ranges, the predicted values differed significantly from the existing values. The model had been trained insufficiently with some lower range values. The estimation values vary a lot, especially between 0 and 90. Since it is an hourly data, the DNI value suddenly increases. The trained model could make much better predictions if the data was collected every 15 minutes or more frequently. We compared 17 different regression models and found that CatBoostRegressor was the most effective one. Among regression models, tree-based models were the top three performers. Then we deployed our model and built a user input pipeline. We took the user's input and made DNI values predictable. The predicted values from the model also matched the observed values. In R^2, we had a very high success rate. The margins of error are quite large, especially at sunrise and sunset, and the DNI fluctuates quite a bit. Because the model could not be properly fed, its success rate remained low during these intervals. However, the application of machine learning in this field is promising.

6 FUTURE WORK

We can train the model by obtaining data from another source every 15 minutes or more frequently. Time-series forecasting can be done with at least 3,4 years of data. In addition to machine learning, regression deep learning methods can be used with the keras library or others.

REFERENCES

(2022, 05 25). Retrieved from IBERDROLA: https://www.iberdrola.com/social-commitment/solar-radiation

(2022, 05 25). Retrieved from sevensensor: https://www.sevensensor.com/what-is-pyranometer

(2022, 05 25). Retrieved from mgm: https://www.mgm.gov.tr/genel/meteorolojikaletler.aspx?s=9

(2022, 05 25). Retrieved from CatBoost: https://catboost.ai/en/docs/concepts/algorithm-main-stages

A Meeting with the Universe. (2022, 05 25). Retrieved from history.nasa.gov: https://history.nasa.gov/EP-177/ch3-2.html

Al-Taani, H., & Arabasi, S. (2018). Solar irradiance measurements using smart devices: A cost-effective technique for estimation of solar irradiance for sustainable energy systems. *Sustainability*.

Amanda Barnett. (2022, 05 25). *NASA Science*. Retrieved from solarsystem.nasa.gov: https://solarsystem.nasa.gov/solar-system/sun/in-depth/

Brownlee, J. (2022, 05 25). *Hyperparameter Optimization With Random Search and Grid Search*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/hyperparameter-optimization-with-random-search-and-grid-search/

Cao, J., & Lin, X. (2008). Application of the diagonal recurrent wavelet neural network to solar irradiation forecast assisted with fuzzy technique. Engineering Applications of Artificial Intelligence. *Engineering Applications of Artificial Intelligence, 21*(8), 1255-1263.

COFFEY, D. (2022, 05 25). *How cold is space, and how hot is the sun?* Retrieved from POPULAR SCIENCE : https://www.popsci.com/why-is-space-cold-sun-hot/

Li, J., Ward, J. K., Tong, J., Collins, L., & Platt, G. (2016). Machine learning for solar irradiance forecasting of photovoltaic system. *Renewable energy*, 542-553.

Pazikadin, A. R., Rifai, D., Ali, K., Malik, M. Z., Abdalla, A. N., & Faraj, M. A. (2020). Solar irradiance measurement instrumentation and power solar generation forecasting based on Artificial Neural Networks (ANN): A review of five years research trend. *Science of The Total Environment*, 715.

Perez-Astudillo, D., & Bachour, D. D. N. I. (2014). DNI, GHI and DHI ground measurements in Doha, Qatar. *Energy Procedia*, 2398-2404.

Pyranometer vs Pyrheliometer: What's the Difference? (2022, 05 25). Retrieved from Solartechadvisor: https://solartechadvisor.com/pyranometer-vs-pyrheliometer/

Quaschning, V. (2022, 05 25). *1*. Retrieved from volker-quaschning.de: https://www.volker-quaschning.de/articles/fundamentals1/index.php#:~:text=The%20sun%20is%20the%20source,irradiation%20directly%20into%20useable%20energy

Singh, S. (2022, 05 25). *Solar Irradiance Concepts: DNI, DHI, GHI & GTI*. Retrieved from yellowhaze: https://www.yellowhaze.in/solar-irradiance/

Trapero, J. R., Kourentzes, N., & Martin, A. (2015). Short-term solar irradiation forecasting based on dynamic harmonic regression. *Energy*, 289-295.

Yan, K. Shen &H., Wang & L., Zhou & H., Xu, M., & Mo, Y. (2020). Short-term solar irradiance forecasting based on a hybrid deep learning methodology. *Information, 11*(1), 32.

Yang, D., Jirutitijaroen, P., & Walsh, W. M. (2012). Hourly solar irradiance time series forecasting using cloud cover index. *Solar Energy, 86*(12), 3531-3543.

BIOGRAPHY

Personal Information

Name Surname: Safi CENGİZ

Place and Date of Birth: AYDIN, 1993

Education

Undergraduate Education: Bachelor in Economics, Marmara University

Graduate Education: Master in Computer Engineering, Beykoz University

Foreign Language Skills: English

Work Experience

T. Halk Bankasi A.S. 2017-2021

Contact

Email address: saficengiz1@gmail.com