Solar Radiation Prediction

***Abstract*—For an efficient conversion and utilization of solar power, solar radiation data should be measured continuously and accurately over the long-term. However, the measurement of solar radiation is not available for all countries in the world due to some technical limitations. Hence, several studies were proposed in the literature to find mathematical and physical models to estimate the amount of solar radiations depending on various atmospheric features. In this project, we review our work for forecasting hourly solar radiation based on the combination of unsupervised and supervised machine learning algorithms and artificial.**

This dataset has even more great value for predicting the radiation levels as the features capture most of the useful information. The Radiation column is the target feature which must be predicted (dependent variable) and rest others are the trainable features (independent variables). Dimensionality reduction techniques like PCA and other feature engineering techniques are also applied on the data to reduce the higher dimension in the data.

We are going to do an analysis of what factors like temperature, pressure, humidity etc, the solar radiation depends on. Such a model has several applications and can be used in real life applications like planting of solar panels at place depending on the above factors. We are plannning to use Linear Regression, Decision Tree Classifier, K-Means, PCA, Logistic Classification.

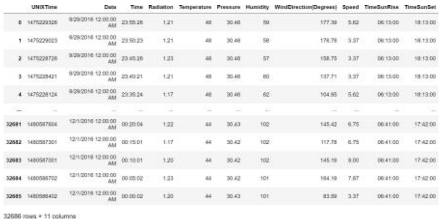
I. INTRODUCTION

The sun is the planet's main energy source, and solar radiation affects the hydrological cycle, photosynthesis of plants, weather extremes, and temperature on a large scale. Therefore, accurate solar radiation forecasting is crucial for both the solar business and climate research. A stacking model employing the best of these algorithms was produced to predict solar radiation. We built 12 machine learning models to predict and compare daily and monthly values of solar radiation. The findings demonstrate the significance of meteorological variables for machine learning models, including sunshine duration, land surface temperature, and visibility.

The role of solar radiation in combining extreme climate events was demonstrated by trend analysis between extreme land surface temperatures and the amount of solar radiation. The ability to forecast daily and monthly solar radiation was better using regression models. New features were also created to classify whether it is a high radiation or low radiation. The segmentation of the radiation categories is also done using clustering algorithms like K-Means clustering and Gaussian Mixture Model. To reduce the dimensionality, Principal Component Analysis is also applied on the data.

II. OVERVIEW OF DATASET

The dataset contains such columns as: "wind direction", "wind speed", "humidity" and temperature. The response parameter that is to be predicted is: "SolarRadiation". It contains measurements for the past 4 months and we have to predict the level of solar radiation using Machine Learning techniques. We worked with 3 supervised methods and 2 unsupervised methods to analyse and predict the data. We also applied some feature engineering techniques to transform the data for better prediction and analysis and also found correlation between the different categorical features of the dataset.



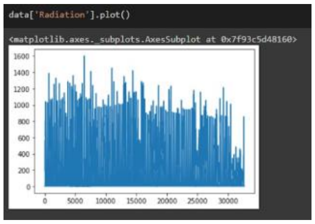
	UNIXTime	Data	Time	Radiation	Temperature	Pressure	Humidity	WindDirection(Degrees)	Speed	TimeSunRise	TimeSunSet
0	1475229326	9/29/2016 12:00:00 AM	23:55:26	1.21	48	30.46	59	177.39	5.62	06:13:00	18:13:00
1	1475229023	9/29/2016 12:00:00 AM	23:50:23	1.21	48	30.46	58	176.78	3.37	06:13:00	18:13:00
2	1475228726	9/29/2016 12:00:00 AM	23:45:26	1.23	48	30.46	57	158.75	3.37	06:13:00	18:13:00
3	1475228421	9/29/2016 12:00:00 AM	23:40:21	1.21	48	30.46	60	137.71	3.37	06:13:00	18:13:00
4	1475228124	9/29/2016 12:00:00 AM	23:35:24	1.17	48	30.46	62	104.95	5.62	06:13:00	18:13:00
...	...	...	...	...	...	...	...	...	...	...	...
32681	1480587604	12/1/2016 12:00:00 AM	00:20:04	1.22	44	30.43	102	145.42	6.75	06:41:00	17:42:00
32682	1480587301	12/1/2016 12:00:00 AM	00:15:01	1.17	44	30.42	102	117.78	6.75	06:41:00	17:42:00
32683	1480587001	12/1/2016 12:00:00 AM	00:10:01	1.20	44	30.42	102	145.19	9.00	06:41:00	17:42:00
32684	1480586702	12/1/2016 12:00:00 AM	00:05:02	1.23	44	30.42	101	164.19	7.87	06:41:00	17:42:00
32685	1480586402	12/1/2016 12:00:00 AM	00:00:02	1.20	44	30.43	101	83.59	3.37	06:41:00	17:42:00

32686 rows × 11 columns

Fig. 1. DATASET

III. EXPLORATORY DATA ANALYSIS

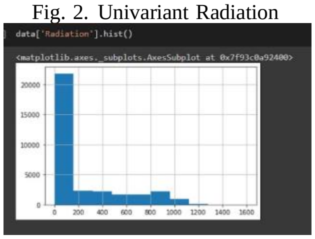
Exploratory data analysis is the crucial process of doing preliminary analyses on data in order to find patterns, identify anomalies, test hypotheses, and double-check assumptions with the aid of summary statistics and graphical representations. The distribution of radiation values shown in the figure below.



```
data['Radiation'].plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f93c5d48160>
```

Fig. 2. Univariant Radiation



```
data['Radiation'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f93c0a92400>
```

Fig. 3. Histogram for Radiation

The distribution between the mean radiation value and the hour of the day is displayed below. We can observe that the radiation value peaks from 10am in the morning and reaches maximum at 12. It gradually decreases from 1pm and goes back to normal.

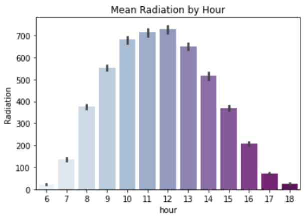


Fig. 4. Mean Radiation vs Hour of the day

The mean radiation value is also observed over the months. We can see that it is higher in the months of September, October and it gradually decreases in December and January. The figure is shown below

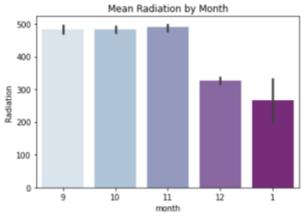


Fig. 5. Mean Radiation vs Month of the day

The distribution between temperature and radiation is shown below in Fig.5.

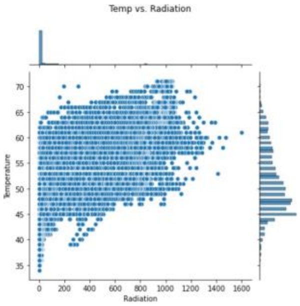


Fig. 6. Temperature vs Radiation

A correlation matrix is a table that shows the correlation coefficients for various variables is shown in Fig.6. The correlation between all potential pairs of values in a table is shown in the matrix. It is an effective tool for compiling a sizable dataset and for locating and displaying data patterns.

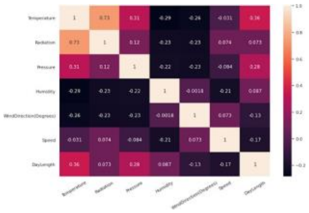
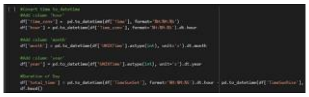


Fig. 7. Correlation Map

IV. FEATURE ENGINEERING

Feature engineering is the process of modifying your data set, including addition, deletion, combination, and mutation, in order to enhance the training of your machine learning model and achieve improved accuracy and performance. A solid understanding of the business issue and the data sources at hand is the foundation for effective feature engineering. Based on the attributes of the data, the following new features were created :

1.Hour - Derived from UNIXTime 2.Month - Derived from UNIXTime 3.Year - Derived from UNIXTime 4.Total Time - Derived from TimeSunset



```
[ ] #Convert time to_datetime
    #Add column 'hour'
    df['time_conv'] =  pd.to_datetime(df['Time'], format='%H:%M:%S')
    df['hour'] = pd.to_datetime(df['time_conv'], format='%H:%M:%S').dt.hour

    #Add column 'month'
    df['month'] = pd.to_datetime(df['UNIXTime'].astype(int), unit='s').dt.month

    #Add column 'year'
    df['year'] = pd.to_datetime(df['UNIXTime'].astype(int), unit='s').dt.year

    #Duration of Day
    df['total_time'] = pd.to_datetime(df['TimeSunSet'], format='%H:%M:%S').dt.hour - pd.to_datetime(df['TimeSunRise'],
    df.head()
```

Fig. 8. Feature Creation

V. PRINCIPAL COMPONENT ANALYSIS

It is a technique for dimensionality reduction in an unsupervised model. which splits correlated features into fewer, principal component-level uncorrelated variables. PCA is frequently used when no target variables exist. In this case, PCA was employed to weed out the least significant based on the explained variance ratio attributes. As PCA is primarily used for, we first sought to remove the classified columns. After that, we dimensioned the variables by scaling them. Assumed 3 randomly chosen primary components, which accounted for about 95. Next, we plotted the proportion of variation by primary using a scree plot. components and was successful in removing the characteristics with a low explained ration which are the transit count.

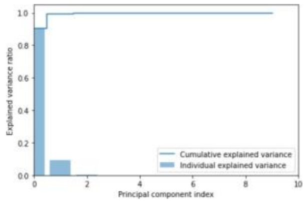


Fig. 9. PCA components with variance ratio

VI. RADIATION SEGMENTATION USING CLUSTERING

A machine learning approach called clustering or cluster analysis organizes the unlabeled dataset. A definition of it is ”A method of grouping the data points into various clusters, consisting of similar data points, and keeping the objects with potential similarity in a group that has fewer or no similarities with another group.” It accomplishes this by identifying comparable patterns in the unlabeled dataset, such as form, size, color, behavior, etc., then classifying the data according to the presence or absence of these patterns. Initially, the number of clusters are declared using the elbow method. It is calculated based on the inertia of the k means algorithm. The elbow graph is shown below. By understanding the elbow graph, we can conclude that 2 clusters are good enough for the data.

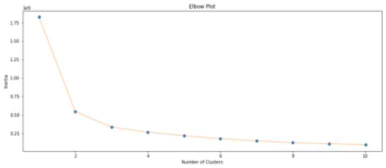


Fig. 10. Elbow Method for component1 and component2

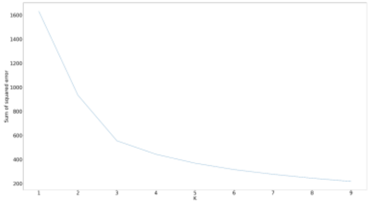
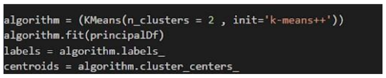


Fig. 11. Elbow Method for Pressure and Temperature

A. KMeans Algorithm

Unsupervised learning algorithm K-Means Clustering divides the unlabeled dataset into several clusters. Here, K specifies how many pre-defined clusters must be produced as part of the process; for example, if K=2, there will be two clusters, if K=3, there will be three clusters, and so on. It is an iterative approach that separates the unlabeled dataset into k distinct clusters, each of which contains just one dataset and shares a set of characteristics. It gives us the ability to divide the data into several groups and provides a practical method for automatically identifying the groups in the unlabeled dataset without the need for any training. Each cluster has a centroid assigned to it since the technique is centroid-based. In this clustering analysis, all the three principal components are given to the k means model with initiating parameter as kmeans++.



```
algorithm = (KMeans(n_clusters = 2 , init='k-means++'))
algorithm.fit(principalDf)
labels = algorithm.labels_
centroids = algorithm.cluster_centers_
```

Fig. 12. KMeans Algorithm

After segmentation, the plot between component1 and component2 is shown below.

The cluster model is evaluated on silhouette score. The separation distance between the generated clusters may be investigated using silhouette analysis. The silhouette plot offers a visual approach to evaluate factors like the number of clusters by displaying a measure of how near each point in one cluster is to points in the surrounding clusters. The range of this metric is [-1, 1]. The sample is remote from the surrounding clusters if the silhouette coefficients (as these values are known) are close to +1. Indicated by a value of 0, a sample is on or very near the boundary between two nearby clusters, while negative values suggest that the sample may have been mistakenly allocated to another cluster. The Silhouette score for this model is 0.577.

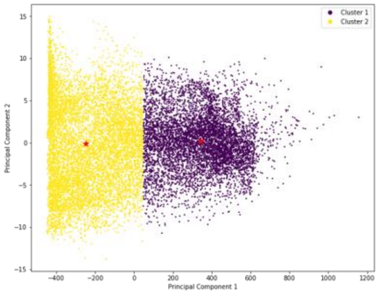


Fig. 13. Scatter Plot between Component1 and Component2

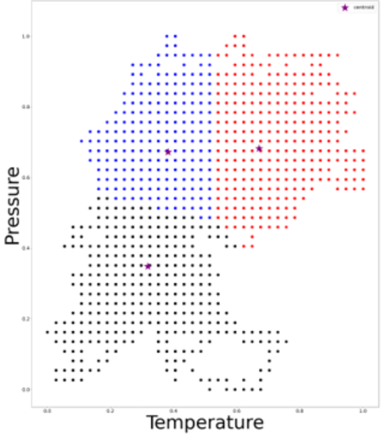
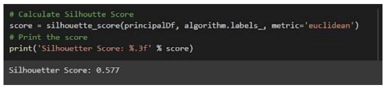


Fig. 14. Scatter Plot between Pressure and Temperature



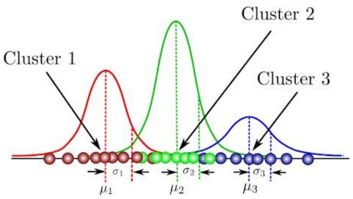
```
# Calculate Silhoutte Score
score = silhouette_score(principalDf, algorithm.labels_, metric='euclidean')
# Print the score
print('Silhouetter Score: %.3f' % score)
```

```
Silhouetter Score: 0.577
```

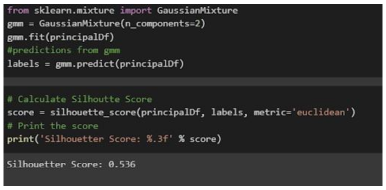
Fig. 15. Silhouette score

B. Gaussian Mixture Algorithm

A Gaussian Mixture is a function that is comprised of several Gaussians, each identified by k belongs to 1,. . . , K, where K is the number of clusters of our dataset. Each Gaussian k in the mixture is comprised of the following parameters: A mean that defines its centre. A covariance that defines its width. This would be equivalent to the dimensions of an ellipsoid in a multivariate scenario. A mixing probability that defines how big or small the Gaussian function will be. Let us now illustrate these parameters graphically:



The GMM model is trained on 2 clusters as shown below:



```
from sklearn.mixture import GaussianMixture
gmm = GaussianMixture(n_components=2)
gmm.fit(principalDf)
#predictions from gmm
labels = gmm.predict(principalDf)
```

```
# Calculate Silhoutte Score
score = silhouette_score(principalDf, labels, metric='euclidean')
# Print the score
print('Silhouetter Score: %.3f' % score)
```

```
Silhouetter Score: 0.536
```

VII. RADIATION PREDICTION USING REGRESSION

Any forecasting or predictive model must include regression analysis, making it a frequent technique in machine learning-powered predictive analytics. Regression is a typical application for supervised machine learning models in addition to classification. The input and output training data for this method of training models have to be labeled. Accurately labeled training data is essential because machine learning regression models need to grasp the link between features and outcome variables. Since regression is a crucial component of predictive modeling, it may be found in a wide range of machine learning applications. Regression analysis may provide organizations with crucial information for decision-making, whether it's used to enable financial forecasts or forecast healthcare trends. It is already utilized in several industries to map pay changes and anticipate home values as well as stock and share prices. In this project, the regression task is performed on predicting the radiation value. The following independent features are removed and rest all are considered for model training : 'Radiation', 'Data', 'Time', 'TimeSunRise', 'TimeSunSet','RadorNot','UNIXTime' and 'Radiation' is taken as dependent / target variable.

A. Linear Regression

The base model for any regression problem is the linear regression. This model will give basic understanding of how regression works. It fails to regularize the data which leads to false convergence. Data transformation also plays an important role. The linear regression procedure, often known as linear regression, demonstrates a linear relationship between a dependent (y) and one or more independent (y) variables. Given that linear regression demonstrates a linear connection, it may be used to determine how the dependent variable's value changes as a function of the independent variable's value. For the linear regression model built, the scatter plot between true y and predicted y is shown below. 1.Overfitting can be reduced by regularization. 2.Sensitive to Outliers.

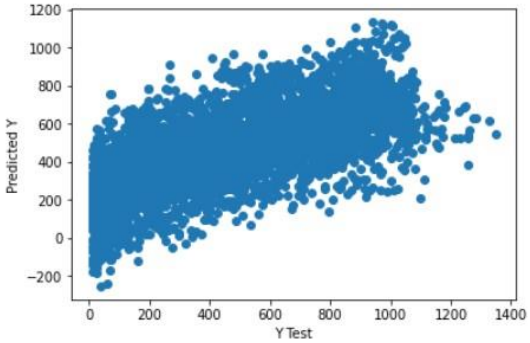


Fig. 16. Predicted values vs Actual values of Solar Radiation by Linear Regression

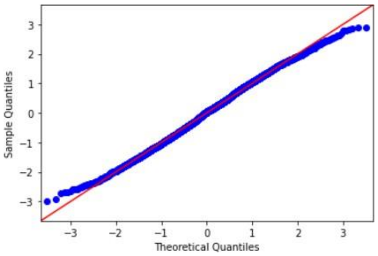


Fig. 17. Best Fit Line for Linear Regression

B. DecisionTree Regressor

Decision tree regression trains a model in the form of a tree to predict data in the future and provide useful continuous output by observing the properties of an item. Continuous output denotes the absence of discrete output, i.e., output that is not only represented by a discrete, well-known set of numbers or values. For the Decision Tree regression model built, the scatter plot between actual values of solar prediction and predicted values of solar radiation is shown below. The best fit line is also explained using the figure.

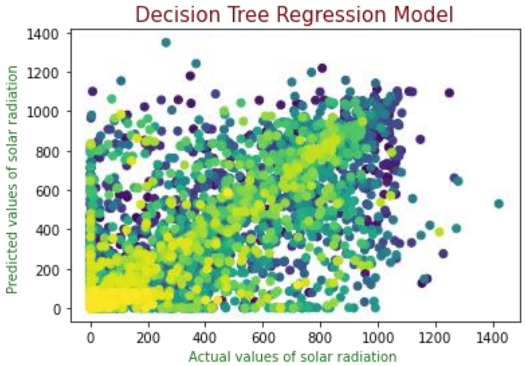


Fig. 18. Actual vs Predicted values of Solar Radiation By Decision Tree Regression

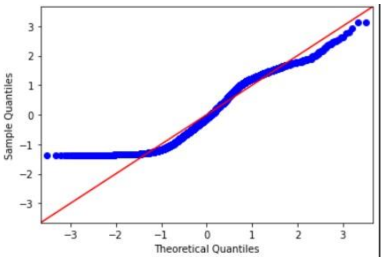
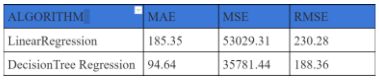


Fig. 19. Best Fit Line for Decision Tree Regression

VIII. RESULTS COMPARISION

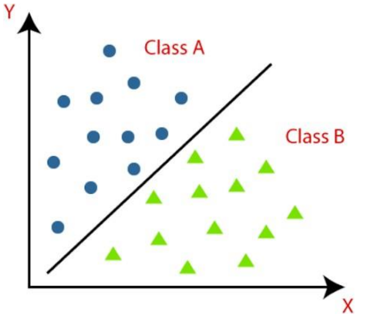
The above two algorithms are compared based on the evaluation metrics related to regression algorithms. The following table shows the results for both the algorithms.



ALGORITHM	MAE	MSE	RMSE
LinearRegression	185.35	53029.31	230.28
DecisionTree Regression	94.64	35781.44	188.36

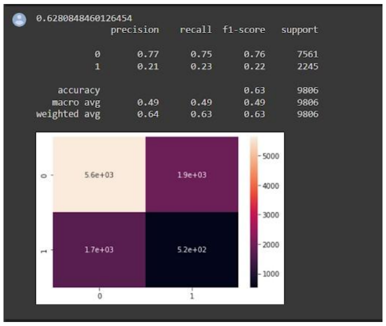
IX. RADIATION PREDICTION USING CLASSIFICATION

The primary objective of a classification algorithm is to determine the category of a given dataset, and these algorithms are mostly employed to forecast the results for categorical data. The graphic below can be used to better understand classification methods. Two classes, class A and class B, are shown in the diagram below. These classes have characteristics in common with one another and that set them apart from other classes.



A. DecisionTree Classification

A supervised machine learning algorithm called Decision Tree employs a set of principles to make judgments, much like how people do. A machine learning classification algorithm may be thought of as being created to make choices. The model is typically said to forecast the class of the novel, previously unseen input, but in reality, the algorithm must choose the class to be assigned. Below is the accuracy, precision, recall, f1-score and confusion matrix table. The accuracy is found to be 63%.



```
0.6280848460126454
              precision    recall  f1-score   support

           0       0.77      0.75      0.76      7561
           1       0.21      0.23      0.22      2245

    accuracy                           0.63      9806
   macro avg       0.49      0.49      0.49      9806
weighted avg       0.64      0.63      0.63      9806
```

X. FUTURE SCOPE AND CONCLUSION

In this project, we have taken a solar radiation prediction dataset which contains different data types of data belonging to numerical as well as categorical features. The data is then preprocessed for missing values and several new features were created based on the requirement. With the existing features and newly created features, data analysis is done to understand the nature of how solar radiation is distributed. After the data is ready, perform a clustering task on the data. Using K-Means the clustering of all the records based on the similarity is performed. Later, using different regression and classification techniques like linear regression, decision tree classification, decision trees, the radiation values are predicted. Further these algorithms are evaluated using Mean squared error, Root mean squared error for regression algorithms and Accuracy, precision, recall for classification algorithms. To further improvise this project, hyper parameter tuning can be done on the data to improve the metrics even more. Cross validation is also an improvement which can be applied on the models. Even WebUI can be created and hosted on cloud where we just have to give the values of temperature, pressure etc and it will predict the radiation value and also we can continously monitor the data. These techniques can also be used to predict solar power generation from a given solar station, depending on weather conditions on a particular day, we can coordinate different power stations accordingly for the optimum power generation.

XI. WORK CONTRIBUTIONS

Vinay Vaida