National College of Ireland

Solar Irradiation Components Prediction Using Meteorological Measures in Odeillo

MSc Research Project
Data Analytics

Yogesh Maruti Patil
Student ID: x17169828

School of Computing
National College of Ireland

Supervisor: Dr. Catherine Mulwa

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Yogesh Maruti Patil
Student ID:	x17169828
Programme:	Data Analytics
Year:	2019
Module:	MSc Research Project
Supervisor:	Dr. Catherine Mulwa
Submission Due Date:	16/09/2019
Project Title:	Solar Irradiation Components Prediction Using Meteorological Measures in Odeillo
Word Count:	8609
Page Count:	25

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Solar Irradiation Components Prediction Using Meteorological Measures in Odeillo

Yogesh Maruti Patil
x17169828

Abstract

The rise in the usage of solar energy for electricity generation has led to growing interest among the researchers to predict the solar radiations at a more granular level. A good characterization of solar irradiation components is also important to produce the best performance from photovoltaic plant(PV) and concentrated power plant(CSP). The three most prominent solar irradiation components include Global Horizontal Irradiation(GHI), Diffuse Horizontal Irradiation (DHI) and Diffuse Normal Irradiation (DNI). The objective of this project is to predict the individual components of solar irradiations utilizing added meteorological measures. And, also to evaluate the ability of different types of time series models in the prediction of individual components of solar irradiation. Solar and meteorological data at the site of Odeillo is used for training and testing the models. Six machine learning models namely Classification And Regression Trees, Linear Regression, Stochastic Gradient Boosting, KNN, Support Vector Machine and Random Forest model are implemented to predict GHI, DHI and DNI on an hourly basis. Four evaluation metrics were used namely RMSE,nRMSE, MAE, nMAE. The results concluded that the addition of meteorological measures significantly helps in improving the prediction of individual solar irradiation components. This study can help firms dealing with solar installations and energy generations to decide as to which plant to set up where.

1 Introduction

Solar irradiation is defined as the total amount of radiation per unit area that is received from the sun at a particular point on the surface of the earth at some given moment[1]. The three most prominent components of solar irradiance include GHI, DHI, and DNI. GHI is defined as an aggregate of the amount of short-wave radiation that is obtained from the above via a surface that is horizontal to the earth. Whereas DNI is defined as the amount of irradiances which are received directly on a surface which is normal to suns rays over the entire solar spectrum. DNI is the total amount of suns rays that arrive on the surface with a straight path from the sky, but in a distributed fashion to some extent because of the particles in the environment (Blanc et al.; 2014). Rays from the sun can be converted into power with the help of advanced systems amongst which the two most common and prominent being concentrating photovoltaic (CPV) and photovoltaic (PV) systems. GHI is the main driver for photovoltaic plants so they require GHI forecast and eventually make use of both solar irradiation components i.e. direct and diffuse. Flat plate solar systems like PV require a semiconductor device which converts the GHI into electricity. Whereas for concentrating photovoltaic (CPV) technology, DNI is the main driver (Schroedter-Homscheidt et al.; 2013). DNI is converted to heat energy by the use of focusing receivers. Both systems i.e. CSP and PV need highly accurate forecasting of GHI and DNI to yield solar energy in an effective manner (Ramírez and Vindel; 2016).

[1] https://www.nasa.gov/mission$_p ages/sdo/science/solar - irradiance.html$

1.1 Project Background

A tremendous increase is observed in the business of these plants with 100 MW of CSP plant and 100 GWp of Photovoltaic power plants set up in 2017. The total capacity touched 405 GW for PV and for CSP it reached at 4845 MW throughout the world.[2] Europe is an emerging market for solar energy and was named the third-largest place accounting for new installations (9.7 GW) and managed to hold the 2nd rank for total operating capacity. 8.3 GW of solar PV were installed in 2018, which is an increase of 36 percent from the previous year's installations which brought the total capacity to 115 GW. In comparison to 2017, 22 out of 28 EU countries boosted the solar plant's installations. France was also ranked third for the highest number of new installations(0.9 GW)[3]. With the increasing demand for solar energy its important to develop models that can predict and forecast the solar energy in order to use it efficiently.

1.2 Motivation

Energy derived from solar radiation is believed to the cleanest and the most inexhaustible source of renewable energy. This energy is associated with two types of energy i.e. light and heat. With appropriate design, deployment and build-up of solar power plant, solar energy can be utilized remarkably. In order to reduce the cost incurred by the power grid regulators, it is very necessary to manage the resources efficiently for which an accurate solar irradiance and the solar power forecasting is necessary. Over the years there has been a huge advancement in the technology and with it a boost in the kind of methodologies and equipment employed for the solar irradiance measurement as well as for weather forecasting (Kalnay; 2002). An important point to focus is that as compared to GHI component, DHI and DNI components are quite difficult to predict as they are very sensitive to the variable meteorological conditions (Benali et al.; 2019). So to effectively develop, control and administer the power grid, forecasting of solar irradiation components with meteorological measures is necessary (Notton et al.; 2019).

This project is concentrated at presenting different methodologies to predict solar irradiations individual components mainly GHI, DHI, and DNI on an hourly basis. An important objective of this project is to assess and evaluate the effect of additional meteorological measures in the improvement of the individual GHI, DHI and DNI forecast. This project has taken into consideration hourly data solar components measured at a site at Odeillo, France along with meteorological data. A performance comparison of the models is done with the help of four error metrics mainly RMSE, MAE, nRMSE, and nMAE.

1.3 Project Requirement Specification

1.3.1 Research Question

RQ:*"Can the prediction of the individual components of the solar irradiation namely GHI, DHI, and DNI be significantly improved by using additional meteorological measures compared with any model that uses only solar irradiation components inorder to help the firms dealing with solar installations and energy generations to decide as to which type of plant to set up?"*

Sub-RQ :*"Can the inclusion of meteorological measures like cloud opacity index, azimuth, humidity, wind speed, pressure, EBH, temperature and zenith improve the prediction accuracy of individual solar irradiation components in Odeillo."*

It is beneficial to look for the best locations while determining the cost that will be needed to harness the best photovoltaic potential from that area. There has been a tremendous in-

[2]https://www.eurobserv-er.org/

[3]https://wedocs.unep.org/bitstream/handle/20.500.11822/28496/REN2019.pdf?sequence=1isAllowed=y

crease in the amount of investment in solar plants across the world. Harvesting of solar energy can be achieved in the best productive manner with the accurate prediction of solar irradiation components namely GHI, DHI, and DNI. Though a lot of research is being done previously, the predictive accuracy can be improved by the use of the ideal combination of predictors. The solar components are sensitive to different meteorological conditions which should be taken as into consideration while prediction. The best selection of predictors to predict solar irradiation can help in improving the accuracy of the model and thereafter help the firms dealing with solar plants installation to decide which kind of solar plant to setup.

To solve the above mentioned research and the sub research question, a certain objectives are listed in section 1.3.2 and these are later implemented, evaluated and the results corresponding to them are exhibited in section 4.

1.3.2 Research Objectives and Contributions

Objectives to be achieved in this project work are mentioned below:

Obj.1 A review of the global literature on photovoltaic potential and solar irradiance from 2005 to 2019. The results will help in understanding the different predictors and models used until now

Obj.2 Perform exploratory data analysis of the solar irradiation and meteorological data.

Obj.3 Feature selection using Boruta algorithm and Random Forest

Obj.4 Implementation, evaluation, and results of solar irradiation regression models.

- Obj 4.1 CART model.
- Obj 4.2 Linear Regression model.
- Obj 4.3 Stochastic Gradient Boosting model.
- Obj 4.4 k-nearest neighbors Algorithm.
- Obj 4.5 Support Vector Machine Model.
- Obj 4.6 Random Forest.

Obj.5 Annual performances comparison of the developed models for time horizons(h+1 to h+6)

- Obj 5.1 For Hourly GHI prediction.
- Obj 5.2 For Hourly DHI prediction.
- Obj 5.3 For Hourly DNI prediction.
- Obj 5.3 For GHI, DHI and DNI in terms of nRMSE.

Obj.6 Seasonal performance comparison of developed models for time horizons(h+1 to h+6)

- Obj 6.1 For Hourly GHI prediction (Autumn, Winters, Summers, and Spring)
- Obj 6.2 For Hourly DHI prediction (Autumn, Winters, Summers, and Spring)
- Obj 6.3 For Hourly DNI prediction (Autumn, Winters, Summers, and Spring)

Obj.7 Compare the accuracy of the developed model with state of the art models.

The rest of the technical report is divided into separate sections. Section 2 presents a comparison and an overview of the predictors and models used in the previous researches. It also discusses the statistics of solar energy production in Europe. A brief overview of research methodologies employed in this project is presented in Section 3. Section 4 presents a thorough explanation of the six machine learning models used in this project. It also talks about the outcomes and comparisons between different models. And in section 5 the project report is concluded.

2 Literature Review

Several past researchers have tried solar irradiation prediction using different models and also with the different predictors. This section is intended to provide an overview of some prominent researches in this field. This section is further divided into five more subsections. Sections 2.1 presents a brief review of solar irradiation research in Europe. Sections 2.2 identifies the different predictors used in earlier researches. Sections 2.3 presents a critique of the models employed by other researchers in solar irradiation prediction at various locations. A comparison of existing models and predictors used in previous researches are presented in Section 2.4. This literature review is concluded in section 2.5.

2.1 A Review of Solar Irradiation in Europe

With the advancement in the technology a lot of research is being conducted around the globe to harness the solar energy in the best possible way. But an emerging trend is observed in the European solar market, as Europe was ranked as the third in terms of highest number of new installation in 2018. This trend can be seen in the literature solely drawn in Europe in the researches conducted by Black et al. (2006), and Li et al. (2011). Also one of the highest installation was observed in France and in 2018 it was ranked third for highest number of new installations. This project also focuses on a location named Odeillo based in France. Odeillo is also famous for its solar furnace which is the world's largest solar furnace. Some previous researches have been conducted at this location and the discussion is being provided as follows.

Global solar irradiation was forecasted by Fouilloy et al. (2018) utilizing eleven statistical and machine learning models and a comparative analysis was done at three locations in France mainly Tilos, Odeillo and Ajaccio.Here the author concluded that the regions with higher meteorological variations seemed to be practicing models with greater complexity to obtain the highest accuracy.In the regions with less meteorological variability, ARMA and MLP demonstrated good results and were termed the most efficient methodologies in these regions. In the study conducted at Odeillo, the meteorological variability was observed to be high and limited scope for reliable solar irradiation forecasting was seen. Another study conducted by Benali et al. (2019) at Odeillo also dealt with the prediction of solar irradiation components. This prediction was done on an hourly basis with different time horizons. Here three predictive models were evaluated namely ANN, smart persistence and random forest. GHI, DHI, and DNI were used this research and the results obtained revealed that random forest was the best amongst all the models. The author also stated that it is more difficult to predict DNI and DHI as compared to GHI. As DNI and DHI are more susceptible to variability in meteorological conditions.From the conclusions drawn by Fouilloy et al. (2018) and Benali et al. (2019) the best accuracy was obtained in case of random forest model. Odeillo being a highly variable meteorological region there is a need to see the effect of other additional meteorological measures on the accuracy of solar irradiation components. Various researches with different predictors are being done previously which is being discussed in the next section.

2.2 An Investigation of Input Variables Used in Solar Irradiation Forecasting and Identified Gaps.

An input variable is essential for any model so as to forecast in the best possible manner. The prediction accuracy varies with the kind of input variable used as well as the number of variables used. So in order to achieve high accuracy, the best choice of input variables amongst those present in the solar irradiance data is vital. This selection eventually helps in improving the performance of the model (Pedro and Coimbra; 2015). In this section, a brief investigation of the different kinds of predictors used by some previous researchers is being done. Two predictors were used by Jamil and Siddiqui (2018) in order to forecast the diffuse solar

radiation. First being the clearness index and the second being the sunshine ratio. For the forecasting of solar irradiation for a shorter time period, Francis M. Lopes (2018) made use of two predictors namely GHI and DNI. Another study by Qing and Niu (2018) made use of data that consisted of measures like humidity, temperature, wind speed, and dew point. In order to forecast the global solar irradiation, Bouchouicha et al. (2019) proposed a new correlation in which sunshine term and air temperature were used. Solar irradiation was predicted on the basis of some air pollutants by Fan et al. (2018) which consisted of pollutants like SO2, O3, PM2.5, NO2, PM10, and CO along with Air Quality Index(AQI). This study by Fan et al. (2018) was mainly focused on predicting diffuse solar radiation (Rd) and Global solar radiation (Rs) that too for a period of two years. In order to predict hourly GHI values, its historic values data were taken into consideration, also in the same research the effect of cloud covers was checked for predicting GHI. To enhance the ability to predict the solar irradiation Chu and Coimbra (2017) thought of incorporating some new external predictors mainly the sky images data and measurement of diffuse irradiance. Though sunshine duration was used in many studies, the combination of sunshine duration with air pollution index was done by Sun et al. (2016) for solar irradiation forecasting. This research also concluded that the highest contributing factor for the estimation of solar irradiation was found to that of sunshine hours and then the API index. Images of the sky with cloud cover information and the data of local irradiance were used in the researches conducted by Pedro and Coimbra (2015) and Pedro et al. (2018). For the prediction of DHI and DNI, Alobaidi et al. (2014) used their historic values were taken along with the effects of clouds and clouds free conditions. In the study conducted by Wu et al. (2017) the importance of different input variables was highlighted that proved to provide good results when used with models that are temperature based. These variables included relative humidity and precipitation on a daily basis.

Though a number of predictors are being used for the prediction of solar irradiation in the previous researches a combination of predictors for predicting individual component of solar irradiation mainly GHI, DHI and DNI are not been done before. Also, the work done at this granularity i.e. on an hourly basis is also limited due to the dearth in the availability of substantial time series on an hourly basis that too for all three components. The limited and poor data availability of DHI and DNI is also being discussed in Benali et al. (2019). The accuracy with which the prediction is done is not dependent only on the type of predictor used, but a lot depends on the kind of model used for it. In the previous researches, a lot of models are being used and a critique of those researches is presented in the next section.

2.3 A Critique of Existing Models, Methods and Techniques Used for Solar Irradiation Forecasting with Identified Gaps

In the past, various models have been used for forecasting of solar irradiation, this section aims to discuss those models. In order to perform a prediction of solar irradiation components mainly DNI and GHI on an hourly basis, Francis M. Lopes (2018) made use of a new forecasting model derived from the European Centre for Medium-Range Weather Forecasts (ECMWF). This model was capable of predicting solar irradiation on an hourly basis for a span of 24 hours. This research was performed for a site in south Portugal on a data of one-year data. To attain a forecast of solar irradiation components(GHI and DNI) on a short term a model named Numerical Weather Prediction (NWP) model was used. The results showed a high correlation in the case of GHI whereas, in the case of DNI, simulations were largely influenced by the cloud as well as aerosols representation. The results and the predicted value obtained proved the model to be efficient. A good selection and analysis of solar irradiation components are necessary so as to obtain the best output from the CSV plants as well as PV plants. Another study by Yang (2012) conducted at two weather stations namely Orlando and Miami in the USA were used to forecast the solar irradiation. In this study DHI, DNI, GHI and cloud covers were used as the predictors. Here the forecasting was performed on an hourly basis. In the first scenario, the

model was fed with some GHI historic values by use of additional seasonal decomposition which was accompanied by a model named Auto-Regressive Integrated Moving Average (ARIMA) model. The second scenario also involved the same procedure but in this case, firstly DHI and DNI were forecasted and thereafter the GHI values were obtained by combining the outputs obtained earlier. The third scenario made use of the relationship of GHI with a polynomial expression of elevation angle measured from vertical under numerous cloud cover situations. It was found out that incorporating the cloud cover in the model increased the accuracy of the model. It was also stated that with the increase in the time scale there were more possibilities of associated errors in context to forecast due to the loss of the meteorological data associated with it. So if the data is available at a granular level that is on an hourly basis, the forecasting can be improved by using some additional meteorological data as the predictors.

Qing and Niu (2018) made use of a new algorithm based on artificial recurrent neural network architecture named Long short-term memory(LSTM) for the hourly prediction of day ahead solar irradiance. Readings from the weather forecast as well as the data at the predicted time were altogether taken as the predictors. As the outputs a day ahead values of the solar irradiance were presented accordingly. The only difficulty observed was to predict the various output concurrently. This methodology proved to be much more efficient and better with accuracy (18 to 34 % higher) than that of Back-Propagation Neural Networks (BPNN) learning algorithm with regards to the root mean square error accounting for a data of two years. To how much extent the prediction is correct is totally dependent on the accuracy of day ahead weather forecasting. Further in research submitted by Fan et al. (2018) Global and Diffuse solar radiation (Rs and Rd) were predicted based upon the data consisting of air pollution as well as the meteorological data. Furthermore, this was conducted for a time period of two years commencing from January 2014 to December 2016 in Beijing, China. To understand the effectiveness of the models, Support Vector Machine (SVM) models were applied which also helped in understanding the influence of the input parameters be it single or a combination. This prediction was performed utilizing various predictors combination. The most important input parameter was observed to be AQI succeeded by O3 and then PM2.5. The most effective combination was seen to be that of PM2.5 and O3 as well as PM10 and O3. It was also seen from the results that as the number of predictors increased the accuracy of prediction decreased gradually.

For the prediction of GHI and DHI individually Pedro et al. (2018) made use of the local solar irradiance data and the data from the images of the sky. To this two machine learning algorithms were applied i.e. k-nearest-neighbors (kNN) and the other being gradient boosting. In order to evaluate the performance of both the algorithm, a comparative analysis was also performed. The analysis showed that the gradient boosting model to which sky images data was fed proved to be showing higher accuracy during each deterministic forecast. The results were totally opposite in case of probabilistic forecast where KNN outperform GB. A similar study with detailed analysis was presented by Pedro and Coimbra (2015) with the same input parameters. From the conclusions, it was evident that the in order to enhance the forecast using KNN the best possible selection of input parameter is a must. Minimal improvement was seen by the addition of data from the sky images. They proved to be helpful and seemed to be relevant only in cases where there was a possibility of cloud cover.

Sun et al. (2016) performed the forecasting of solar radiation by the use of Random forest model. This model was chosen so as to get a better knowledge of how significant is the factor to be considered which can help in the prediction with minimal error. This research was performed at 3 different areas in China with the data consisting of solar irradiation data, some meteorological data along with Air pollution Index. The models with Air pollution index as the predictor

proved to be better than the one without it. The model with API as the predictor proved to be lower at RMSE and also satisfied the NSE values. In some areas like Algeria, there is a dearth of available solar data, so sunshine models were used in these cases. Bouchouicha et al. (2019) suggested the use of a new correlation to forecast the global solar irradiance. The time scales on which this was done varied i.e. on both daily as well as the monthly basis with the use of different predictors especially the meteorological ones. An excellent performance was observed in the case of daily sunshine models with all RMSE values under 10 % and all coefficient of determination values to be greater than 70 %. The empirical models were observed to be much better performing and much more accurate at medium longitudes. The error estimation can be enhanced by the use of K-fold cross-validation method by reducing the standard deviation of RMSEs by more than 99% for all kinds of models.

2.4 A Comparison of Existing Models, Methods, Techniques and Results

This section of the technical report discusses the comparison between a set of models. Wu et al. (2017) evaluated a set of 14 temperature-based models and amongst these 6 were recommended for the estimation of radiation in humid areas. The results from this study suggested that to increase the accuracy of the temperature based models, along with temperature other meteorological measures like relative humidity and precipitation can be considered and added to the model. However, these models proved to be good only for the places which had similar climatic conditions. In another study conducted by Jamil and Siddiqui (2018), 5 areas with different climatic zones were taken into consideration. A comparison of 16 empirical models was done that too under two category depending on the predictors namely the clearness index and the sunshine ratio. Also, the model which incorporated the use of different parameters proved to be the best. Apart from this there were other researchers who used different predictors and models to predict the solar irradiation which is being listed in Table 1

Table 1: Comparison of Existing Models, Methods and Results

Predictors	Model Used	Results	Authors
Tmax, Tmin, PM2.5, PM10 and O3	SVM	RMSE= 1.857 and MAE=1.320	Fan et al. (2018)
n, Tmax, Tmin, API	Random Forest	RMSE= 2.268	Sun et al. (2016)
Temp, Dewpoint, Humidity, Visibility, windspeed, weather type	Persistence, LR, BPNN, LSTM	RMSE= 209.25, 230.98, 133.53, 76.24(W/m2) respectively	Qing and Niu (2018)
Temperature and relative humidity	KNN	RMSE= 3.362 MJ/m2	Bouchouicha et al. (2019)
Cloud cover index, sunshine Duration	EANN (Optimized ANN)	RMSE= 9.09 rMBE=-1.03	Fan et al. (2018)

2.5 Conclusion

It is evident from the literature survey that a lot of research is been done previously to forecast solar radiation. These researches were being carried out using different predictors and also with a combination of predictors to enhance forecasting accuracy. In addition to this various models were used and a comparison of the models was performed by the help of some error

metrics. Also, several ensemble methods were used to enhance the accuracy of prediction. These ensemble methods used a combination of methodologies. To arrive at a result with higher accuracy different researchers made use of totally different methodologies. But these researches also have certain gaps and shortcoming which can be overcome by the use of enhanced hybrid models with a different combination of predictors and that too different meteorological sites. Not much research is done in the forecasting and prediction of all the three components of solar irradiation namely GHI, DHI and DNI individually on an hourly basis. This technical report is aimed at presenting the tests done at the site of Odeillo, France using different methodologies. With this section the research objective 1 set in section 1.3.2 is achieved.

3 Scientific Methodology Approach Used, Design and Data Preparation

3.1 Introduction

This section illustrates the methodology which is being used for this project. This section is subdivided into four more sections. Section 3.2 illustrates the methodology used. Section 3.3 shows the process flow diagram. Section 3.4 describes the data selection, pre-processing and Transformation. Section 3.5 concludes this section.

3.2 Solar Irradiation Methodology Approach

To reveal the patterns in the solar irradiance components and to understand the solar energy potential of Odeillo a new modified Knowledge Discovery in Databases(KDD) Fayyad et al. (1996) methodology is presented. In the figure 1 below the stages involved in this methodology are presented. The first stage involves the selection of solar irradiance and meteorological data on an hourly basis. The second stage involves the pre-processing and transformation of data into annual and seasonal data which is explained in detailed in section 3.4 . The third stage involves the feature selection followed by the data mining phase deals with the application of regression models to predict solar irradiance components. The fifth stage involves the evaluation of the models followed by the presentation of the results obtained to deliver knowledge.

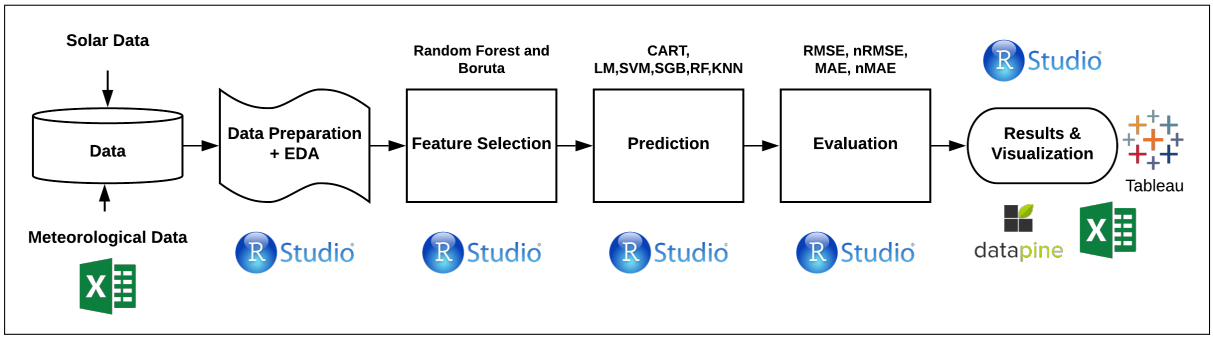

Figure 1: Solar Irradiation Methodology Approach

3.3 Design Architecture

The design process for this project is a two tier based architecture as shown in figure 2 with the Tier II accounting for the enhanced KDD approach as shown in 3.2. An overview on the selection process, pre-processing, and transformation altogether is explained in Sub-Section 3.4.The algorithms used and for data mining and their implementation is discussed in detail in section 4. The Tier I represents the final evaluation and knowledge realization phases of the enhanced KDD methodology approach.

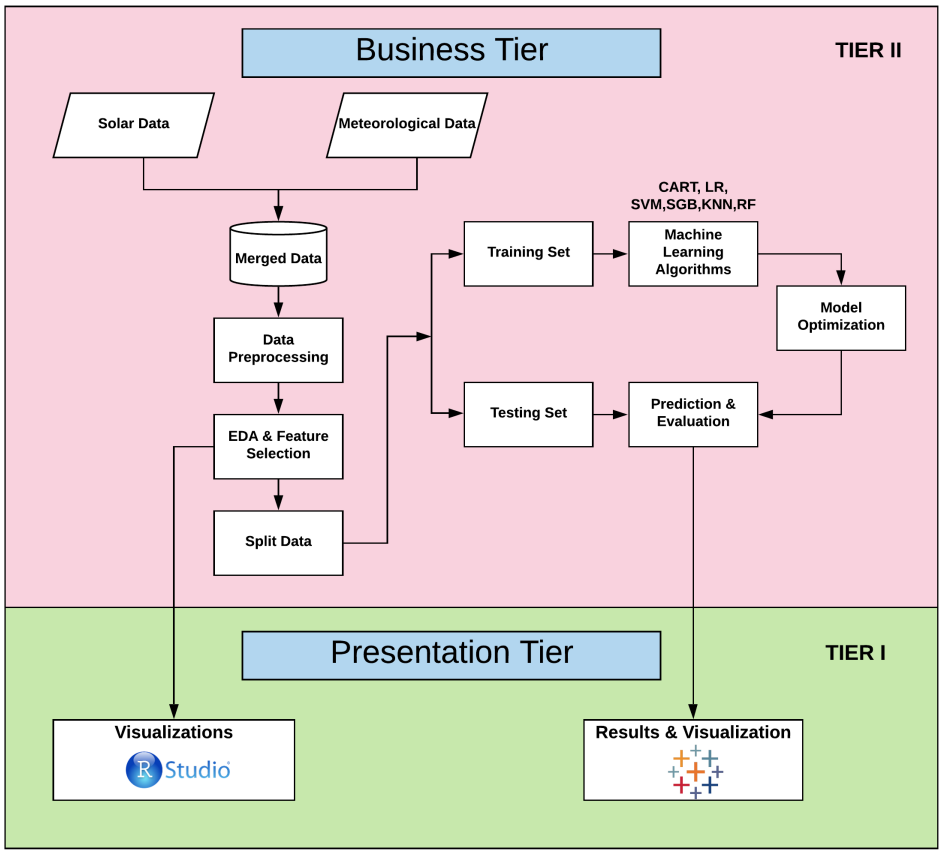

Figure 2: Project design process for solar irradiation prediction

3.4 Data Selection, Pre-processing and Transformation

3.4.1 Data Selection

The quality of data is an important factor in achieving higher predictive accuracy. Accuracy varies with the combination of predictors used. From the literature review done in section 2.2, it was observed that various combination of predictors were considered in the past researches. To carry out this project with new added meteorological measures, the following data sources were considered:

- Solar irradiance individual components data namely GHI, DHI and DNI at the site of Odeillo on an hourly basis was obtained from Solcast [4]. The data was downloaded by providing the coordinates for the location. This data was taken on an hourly basis from the year 2015 to 2018.
- The second source of data included meteorological measures like cloud opacity index, azimuth, humidity, wind speed, pressure, EBH, temperature, and zenith. This data was downloaded from a website name Openweathermap [5]. This data was also taken on an hourly basis from the year 2015 to 2018.

3.4.2 Data Pre-processing

Both the data sources were merged later by Date to form a final data on which the processing was done. Here the data was cleaned and transformed into a new data set. Over here the

[4]https://solcast.com.au/solar-radiation-data/
[5]https://openweathermap.org/

missing values were examined. Special characters and letters not required were removed. Date and time were separated into two different columns. As this project aimed at using regression models all the attributes were checked if they were numerical or not and if found not they were converted into numerical. The unnecessary columns were removed. The next stage involved data transformation according to the requirements set in the research.

3.4.3 Data Transformation

Here the original data after pre processing is transformed into few suitable datasets so as to use them to train the machine learning models. The steps involved in data transformation are as follows:
Step.1: In the first step data consisting of night hours were removed as they are of no use in solar irradiation prediction. Also it is not recommended to use sunset and sunrise hours data as they are consider non reliable for prediction. The solar irradiance component data with value '0' were removed as it was assumed that the data was not captured and won't be reliable to consider it for analysis.
Step.2: Once the data pre processing was done , feature selection was done using random forest and Boruta algorithm to find out the best predictors. Thereafter the important predictors were kept and rest were removed from the data.
Step.3: The second step involved dividing the data into six time horizons (h+1 to h+6).
Step.4: This data was further divided into four seasons (Winters, Spring, Summers and Autumn) each with six time horizons in order to compare the results with the results obtained by Benali et al. (2019).
Following the above steps a total of Five input models were developed i.e. Annual Solar Model, Summer Model, Spring Model, Autumn Model, and Winter Model.

3.5 Conclusion

This section exhibited the phases of the proposed methodology of this research project.In this project, the KDD methodology is customized to satisfy the requirements set in this research. A new solar irradiation methodology approach is explained first followed by the design architecture. An overview of the data selection at the site of Odeillo is discussed in section 3.4.1. The details of the data pre processing and its transformation was discussed in section 3.4.2 and 3.4.3 respectively. The next section 4 describes the data mining in detail, its implementation and evaluation of the models used for predicting solar iradiation components.

4 Implementation, Evaluation and Results of Solar Irradiation Models

4.1 Introduction

Several experiments are executed based on the objectives set in Section 1.3. These are conducted to obtain the answers to the research question. But before doing those experiments it is necessary to do exploratory data analysis and feature selection which is discussed in detail in the sections to follow.

4.2 Exploratory Data Analysis

An essential part of this research project is data exploration and transformation to aims to fulfill the requirements set in this research. Solar and meteorological data are merged by the use of R programming in R studio. Firstly a correlation matrix as shown in figure 3 was used in order to find the correlation between the dependent variable and the predictor.

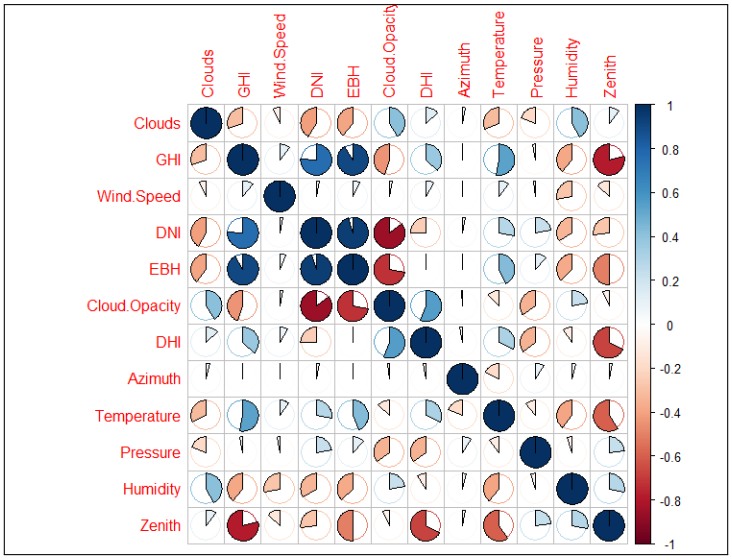

Figure 3: Correlation matrix

In Multiple regression, finding just the correlation between a dependent and predictor variable is not sufficient. So an Added Variable plot is used for the evaluation of the coefficients and the residuals of the independent variables while keeping the additional variables constant. If the coefficients are evaluated without keeping the additional variables constant, the results of the model can be misleading and in turn difficulty to understand the relation between variables. The influence of a selective independent variable on that of a dependent variable while keeping the additional variables constant, can be understood much in a much easier manner by the use of Added Variable plot.

The plot in figure 4 displays the values of an individual predictor after the additional predictors are taken into consideration on the X-axis and value of the GHI after the additional predictors are taken into consideration on the Y-axis.The influence of the variables on GHI prediction can be assessed by looking at the slope of all charts in the figure.The analysis shows that Zenith and Cloud Opacity have the highest influence on GHI prediction. A similar relationship was observed in case of DHI and DNI. With this section, the research objective 2 set in section 1.3.2 is achieved.

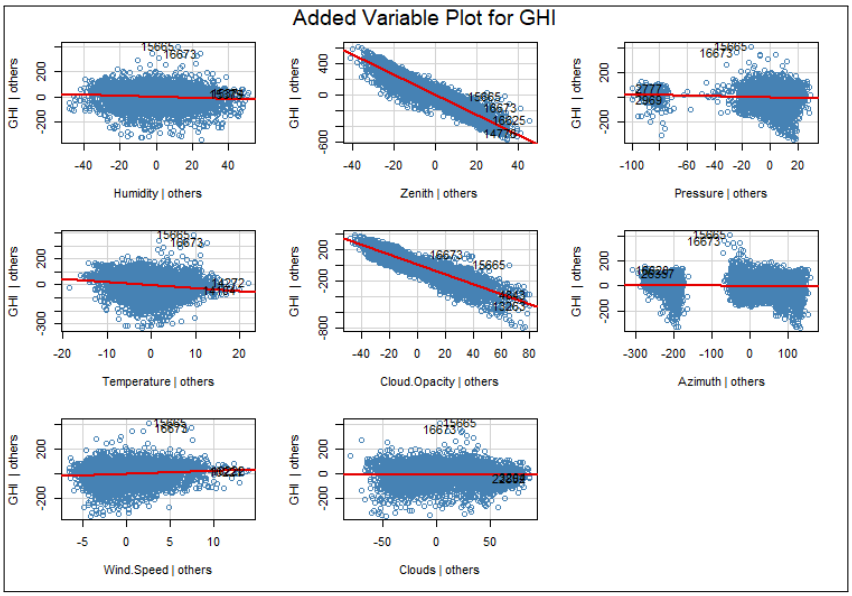

Figure 4: Added Variable Plot for GHI

4.3 Feature Selection

Once the solar and meteorological data is merged, the data frame consisted of three dependent variables namely GHI, DHI and DNI and other eight being the predictors. It is necessary to find out the importance of each of these predictors which is being discussed in this section. In order to choose the best predictors feature selection is mandatory. Some advantages of doing feature selection include faster training of the model, lessens the complexity, simpler interpretation, and improved accuracy.

Random Forest: Random Forest is used for the selection of variables based on the Mean Decrease Gini (MDG). An important benefit of using random forest over univariate methods for feature selection is that RF takes into consideration the influence of each predictor independently and in a multivariate interaction along with other independent variables. In regression the mean node impurity is calculated by residual sum of squares. Larger the MDG value better is the variable. This is achieved in R using a randomForest package. The results showing the important variable for GHI, DHI and DNI are being present in figures 8. In random forest variable importance is decided by the Z score which is calculated by dividing the mean of the accuracy loss by its Standard deviation, but one cannot rely on this Z score as it is not in any way linked with the statistical significance of the importance of a variable.

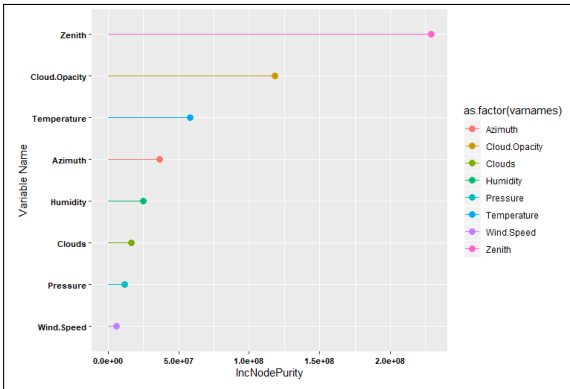

Figure 5: For GHI

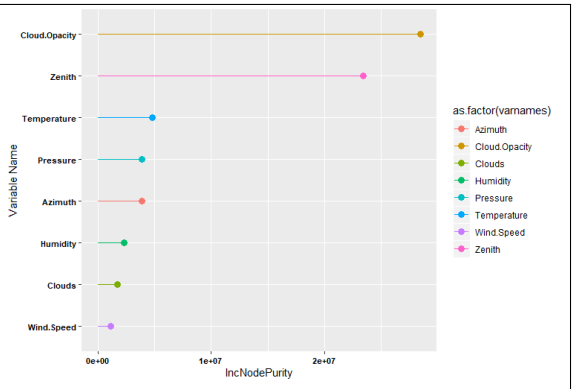

Figure 6: For DHI

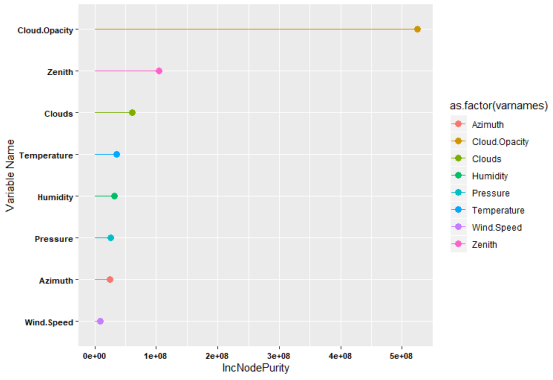

Figure 7: For DNI

Figure 8: Important meteorological Variables by Gini Index

Boruta algorithm: In order to overcome the drawbacks discussed above, Boruta algorithm was used in this project to find out the importance of each of the predictors so as to achieve the best results. Boruta operates RF on both the attributes i.e. on the original as well as the random ones and thereafter calculates the variable importance. This whole process is reliant on permuted copies, so random permutation procedure is repeated to get statistically significant results. A package named Boruta is available in R. An important aspect of this algorithm is that

it decides the importance of variable based on the maximum information about the dependent variable. Data get iterated multiple times in order to find the significance of every predictor. This algorithm divides the variables into a category of three i.e. confirmed important, confirmed unimportant, and tentative/rejected features. After applying Boruta algorithm on the merged solar data, all the eight attributes were confirmed as important. The results for GHI, DHI, and DNI are shown in the Figures 9.

Variables	meanImp	medianImp	decision
Humidity	25.603049	25.4531609	Confirmed
Zenith	125.99264	125.948278	Confirmed
Pressure	21.269218	21.226349	Confirmed
Temperature	25.409888	25.5395614	Confirmed
Cloud.Opacity	175.38619	175.177973	Confirmed
Azimuth	32.193924	32.0704605	Confirmed
Wind.Speed	13.760854	13.6909637	Confirmed
Clouds	22.503465	22.2392382	Confirmed

Figure 9: Variables Confirmed Important by Boruta Algorithm

Variable inflation Factor: Multicollinearity is defined as a state in which two or more dependent or independent variables in a multiple regression model are very linearly correlated. The results achieved from the Boruta algorithm tells only about the importance of a particular variable and doesn't tell anything about the multicollinearity. In the case of regression, it's important to check for multicollinearity in order to avoid any bias in the estimation as well as in the coefficients. Multicollinearity is addressed amongst the variables by the use of variance inflation factor (VIF). VIF provides a score which tells the amount of variance of a regression coefficient that is being inflated because of the multicollinearity in the model. The test was performed on the data validated by Boruta. Companion to Applied Regression (car) package is used with a function vif() in R. The scores obtained are interpreted as values in the range of zero to five are considered acceptable while above five are unacceptable. In this project, all the eight variables had scored below five as seen in figure 10 and so all were considered to be the predictors. With this section, the research objective 3 set in section 1.3.2 is achieved.

```
> vif(xx_reg)
    Humidity       Zenith     Pressure  Temperature Cloud.Opacity      Azimuth
    1.485930     1.834583     1.141354     1.853288     1.260706     1.091405
  Wind.Speed       Clouds
    1.152571     1.399697
```

Figure 10: Variable inflation Factor

4.4 Implementation, Evaluation and Results of Classification And Regression Tree Model

4.4.1 Implementation:

This model was implemented in R using two useful packages named caret and caTools. After loading these two packages the data was divided into two sets i.e. Train data and Test data using a function named sample.split from caTools package. The data was divided in the ratio of 4:1 ratio i.e. 80% of data accounting for training and 20% of data for testing. To implement CART model "rpart" method was used (method="rpart"). The model was implemented for all three solar irradiation components namely GHI, DHI,and DNI individually. And the model was executed for all six time horizons(h+1 to h+6) data. So this model was executed 6 times for every solar irradiation component.

4.4.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only CART model in predicting the GHI, DHI and DNI. Four evaluation metrics were used in this project namely RMSE, nRMSE, MAE and nMAE. But nRMSE being a better indicator to access the performance, its been presented in the figure 11. The nRMSE values indicate lower values for h+1 as compared to h+6 time horizon for all three components. In terms of CART model's annual performance comparison (as shown in figure 23), it performed better in case of DNI with an nRMSE of 34% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), CART performed better in Summers with an nRMSE of 44.50%. For DHI (as shown in figure 27), the model performed better in Winters with an nRMSE of 41.80% for h+1. For DNI (as shown in figure 30), it worked better in Summers with an nRMSE of 26.80. With this section, the research objective 4.1 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	50.7	55.7	33.8
h+6	nRMSE	51.5	58.9	45.4

Figure 11: CART nRMSE for GHI, DHI and DNI

4.5 Implementation, Evaluation and Results of Linear Regression

4.5.1 Implementation:

The initial steps of loading the required packages and dividing the data in Training and Test data is the same as that of CART model. The only difference is the method used. To implement Linear Regression model "lm" method is used (method="lm").

4.5.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only LR model in predicting the GHI, DHI and DNI. nRMSE being a better indicator to access the performance, its been presented in the figure 12. The nRMSE values indicate higher values for h+1 as compared to h+6 time horizon for all three components. In terms of LR model's annual performance comparison (as shown in figure 23), it performed better in case of GHI with an nRMSE of 21% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), LR performed better in Summers with an nRMSE of 10.40%. For DHI (as shown figure 27), the model performed better in Winters with an nRMSE of 28.28% for h+6. For DNI (as shown in figure 30), it worked better in Autumn with an nRMSE of 28.90%. With this section, the research objective 4.2 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	21.8	74	35.5
h+6	nRMSE	20.6	61.2	31.7

Figure 12: LR nRMSE for GHI, DHI and DNI

4.6 Implementation, Evaluation and Results of Stochastic Gradient Boosting

4.6.1 Implementation:

The initial steps of loading the required packages and dividing the data in Training and Test data is the same as that of CART model. The only difference is the method used. To implement Stochastic Gradient Boosting model "gbm" method is used (method="gbm").

4.6.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only SGB model in predicting the GHI, DHI and DNI. nRMSE being a better indicator to access the performance, its been presented in the figure 13. The nRMSE values indicate lower values for h+1 as compared to h+6 time horizon for all three components. In terms of SGB model's annual performance comparison (as shown in figure 23), it performed better in case of GHI with an nRMSE of 7.9% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), SGB performed better in Springs with an nRMSE of 7.80%. For DHI (as shown in figure 27), the model performed better in Autumn with an nRMSE of 21.63% for h+1. For DNI (as shown in figure 30), it worked better in Winters with an nRMSE of 9%. With this section, the research objective 4.3 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	7.9	25.2	11.2
h+6	nRMSE	10.7	33.2	15.5

Figure 13: SGB nRMSE for GHI, DHI and DNI

4.7 Implementation, Evaluation and Results of k-nearest neighbors Algorithm

4.7.1 Implementation:

The initial steps of loading the required packages and dividing the data in Training and Test data is the same as that of CART model. The only difference is the method used. To implement k-nearest neighbors Algorithm model "knn" method is used (method="knn").

4.7.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only KNN model in predicting the GHI, DHI and DNI.nRMSE being a better indicator to access the performance, its been presented in the figure 14. The nRMSE values indicate lower values for h+1 as compared to h+6 time horizon for all three components except for DNI in which h+1 has higher nRMSE than h+6. In terms of KNN model's annual performance comparison (as shown in figure 23), it performed better in case of GHI with an nRMSE of 20% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), KNN performed better in Springs with an nRMSE of 19.80%. For DHI (as shown in figure 27), the model performed better in Summers with an nRMSE of 31.70% for h+1. For DNI (as shown in figure 30), it worked better in Summers with an nRMSE of 17.90%. With this section, the research objective 4.4 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	20	34.3	25.8
h+6	nRMSE	22.9	40.7	24.7

Figure 14: KNN nRMSE for GHI, DHI and DNI

4.8 Implementation, Evaluation and Results of Support Vector Machine Model

4.8.1 Implementation:

The initial steps of loading the required packages and dividing the data in Training and Test data is the same as that of CART model. The only difference is the method used. To implement

Support Vector Machine Model "svmRadial" method is used (method="svmRadial").

4.8.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only SVM model in predicting the GHI, DHI and DNI. nRMSE being a better indicator to access the performance, its been presented in the figure 15. The nRMSE values indicate lower values for h+1 as compared to h+6 time horizon for all three components. In terms of SVM model's annual performance comparison (as shown in figure 23), it performed better in case of GHI with an nRMSE of 12% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), SVM performed better in Springs with an nRMSE of 11.50%. For DHI (as shown in figure 27), the model performed better in Winters with an nRMSE of 21.87% for h+6. For DNI (as shown in figure 30), it worked better in Summers with an nRMSE of 19.10% for h+6. With this section, the research objective 4.5 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	11.9	26.9	16.3
h+6	nRMSE	12.4	31.8	17.4

Figure 15: SVM nRMSE for GHI, DHI and DNI

4.9 Implementation, Evaluation and Results of Random Forest

4.9.1 Implementation:

The initial steps of loading the required packages and dividing the data in Training and Test data is the same as that of CART model. The only difference is the method used. To implement Random Forest model "rf" method is used (method="rf").

4.9.2 Evaluation and Results:

The evaluation and results comparison in this section are in regards with only RF model in predicting the GHI, DHI and DNI. nRMSE being a better indicator to access the performance, its been presented in the figure 16. The nRMSE values indicate lower values for h+1 as compared to h+6 time horizon for all three components. In terms of RF model's annual performance comparison (as shown in figure 23), it performed better in case of GHI with an nRMSE of 7.3% for h+1. In seasonal performance comparison of GHI(as shown in figure 26), RF performed better in Summers with an nRMSE of 12.20%. For DHI (as shown in figure 27), the model performed better in Summers with an nRMSE of 24.50% for h+1. For DNI (as shown in figure 30), it worked better in Winters with an nRMSE of 8.90% for h+1. With this section, the research objective 4.6 set in section 1.3.2 is achieved.

Time Horizon	Metrics	GHI	DHI	DNI
h+1	nRMSE	7.3	25.5	11.7
h+6	nRMSE	10.6	32.5	16.3

Figure 16: RF nRMSE for GHI, DHI and DNI

4.10 Annual Performances Comparison of the Models for time horizons(h+1 to h+6) for Hourly GHI, DHI and DNI.

Comparison of six predictive models on an hourly basis for time horizon h+1 to h+6 is performed in this section. This comparison is done using four different evaluation metrics namely RMSE,

nRMSE, MAE and nMAE but over here RMSE is used to present the comparison. Figure 20 shows the values of the RMSE and respective component and time horizon.

One of the main objectives was to predict the three solar irradiation components for a time horizon of h+1 to h+6 using additional meteorological measures. The rank of a model is mostly identical from Root mean square errors point of view. From the figure 20 it appears that for GHI component RF proved to be best in prediction with an RMSE of 21 and 24 for h+1 and h+6 respectively. For DHI component RF proved to be best in prediction with an RMSE of 28 and 26 for h+1 and h+6 respectively. For DNI component SGB proved to be best in prediction with an RMSE of 42 and 47 for h+1 and h+6 respectively. With this section, the research objective 5 set in section 1.3.2 is achieved.

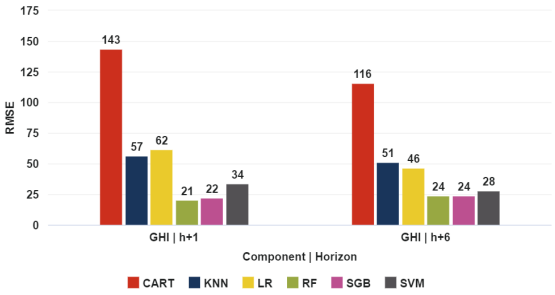

Figure 17: For GHI

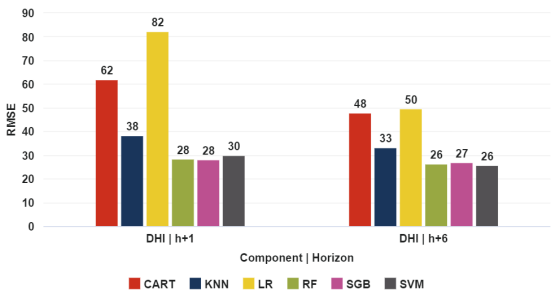

Figure 18: For DHI

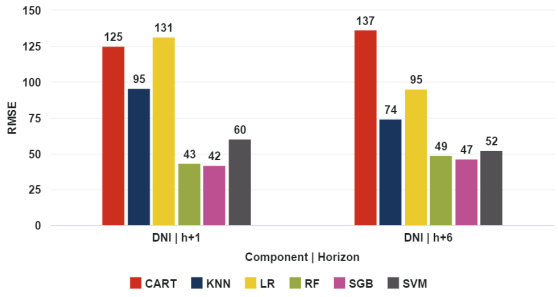

Figure 19: For DNI

Figure 20: Annual performance comparison for Solar Irradiation Components

4.11 Annual Performances Comparison of the Models for GHI, DHI and DNI in term of nRMSE.

Unlike in section 4.10 where RMSE was use to rank model's performance, in this section nRMSE was used as shown in figure 23 as the values of GHI, DHI and DNI are very different. It can be clearly seen that GHI is predicted with better accuracy compared to DHI and DNI. The accuracy obtained by RF and SGB is identical in all the cases. And LR and CART model achieved almost double nRMSE values as compared to RF.

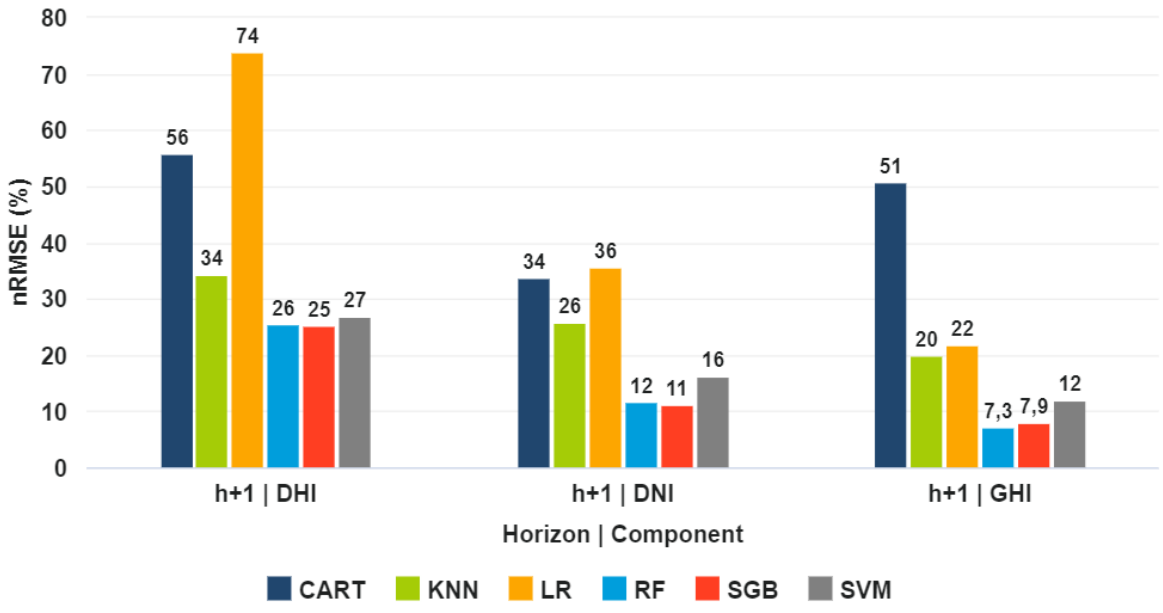

Figure 21: For h+1

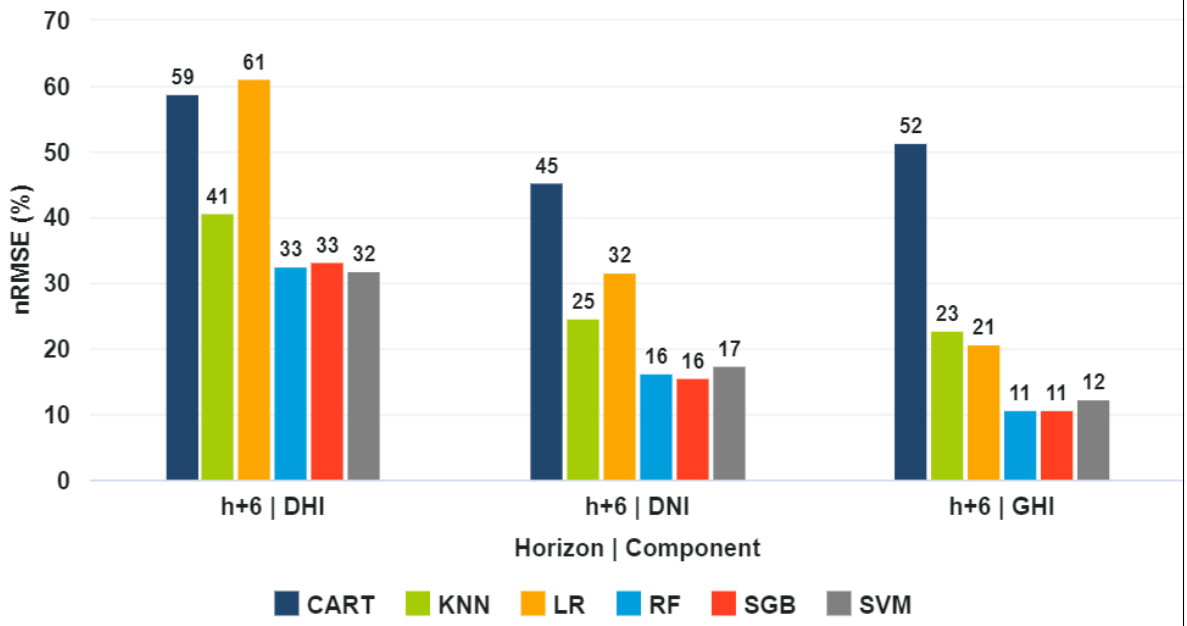

Figure 22: For h+6

Figure 23: Annual Performance Comparison of GHI, DHI and DNI

4.12 Seasonal Performance Comparison of Models for GHI prediction

To understand the importance of meteorological variability on the prediction of solar irradiation, a seasonal study is performed to see its effect on the models and check its accuracy. Here a seasonal performance comparison of models for GHI prediction are presented. Figure 26 shows the comparison of different model for time horizon h+1 and h+6 by Autumn, Spring, Summers, and Winters.

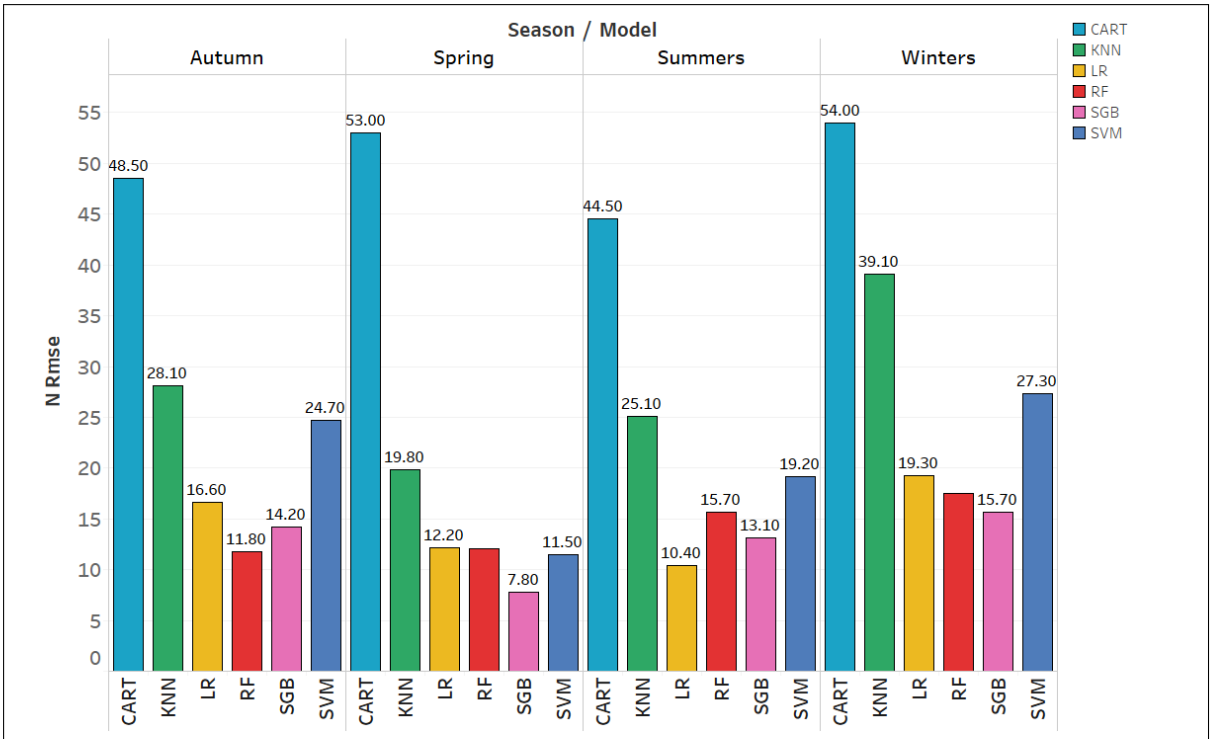

Figure 24: For h+1

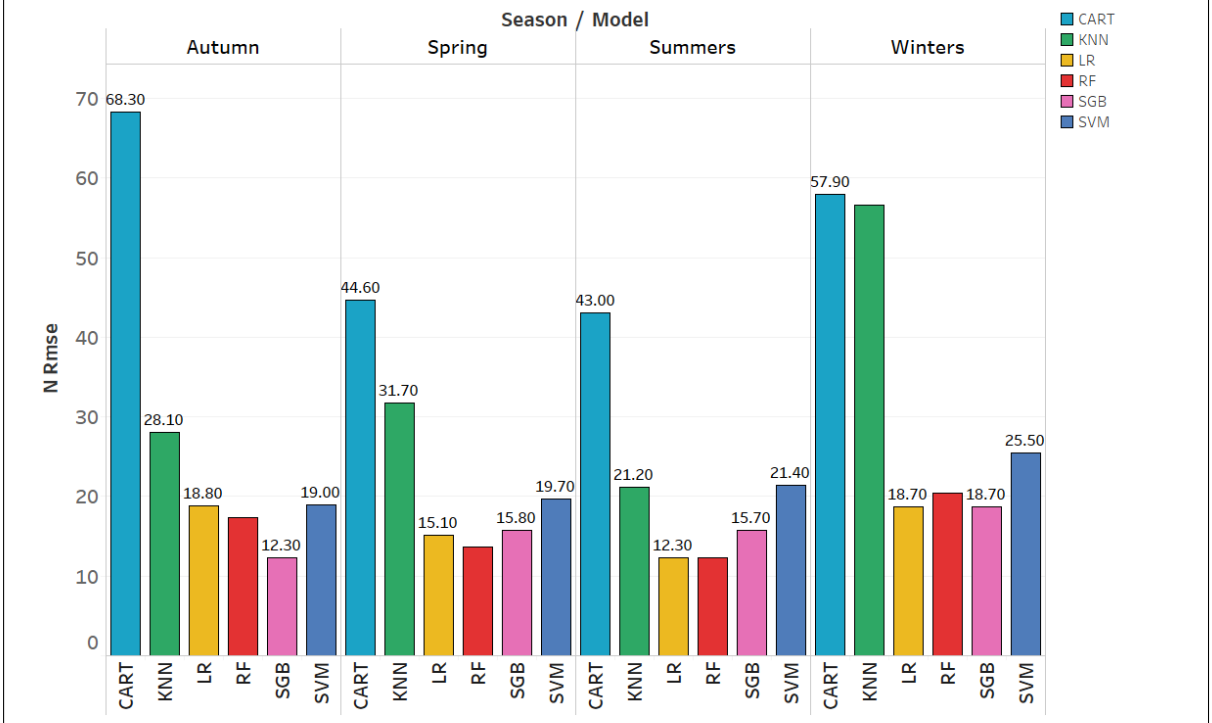

Figure 25: For h+6

Figure 26: Seasonal Performance Comparison for GHI

In order to check the seasonal variation in the accuracy of the prediction of three solar irradiation components for a time horizon of h+1 to h+6 using additional meteorological measures, it is necessary to decide as to how the model should be ranked. To compare and rank the models seasonally it is necessary to use nRMSE and not RMSE because as the seasons change the average hourly irradiation keep changing accordingly and they tend to be different. From

the figure 26, for time horizon h+1 it is observed that GHI component is predicted with a good accuracy in Spring, Summers and Winters using SGB model with an nRMSE of 7.80%, 13.10%, and 15.70% respectively. And RF proved to be better in Autumn with an nRMSE of 11.80%. For h+6, RF proved to be better at predicting in Springs and Summers with an nRMSE of 13.70% and 12.30% respectively. Overall in both the horizon RF and SGB proved to be the best for prediction of GHI seasonally.

4.13 Seasonal Performance Comparison of Models for DHI prediction

In this section a seasonal performance comparison for time horizon h+1 and h+6 of models used for DHI prediction are presented in the below figure 27.

Metric	Model	Autumn		Spring		Summers		Winters	
		h+1	h+6	h+1	h+6	h+1	h+6	h+1	h+6
nRMSE	CART	60.5	55.7	53.7	67.2	41.8	64.2	56.5	64.7
	LR	90	62	76	78.7	67.5	80	91.5	63.4
	SGB	27.1	39.2	27	47.1	26.9	49.7	31.1	33.9
	KNN	54.4	51.9	45.5	59.2	31.7	57.1	44.3	60.5
	SVM	39.7	39.6	39.9	52.8	36.7	51	45.2	47.6
	RF	27.1	41.2	27.7	45.1	24.5	49.2	32.1	32.4

Figure 27: Seasonal performance metrics for DHI

To compare and rank the models seasonally an evaluation metric nRMSE is used, as the seasons change the average hourly irradiation keep changing accordingly and they tend to be different. From the figure 27, it can be seen that in Autumns for time horizon h+1 and h+6 SGB performed the best with an nRMSE of 27.1% and 39.2% respectively. In Springs, for h+1 SGB performed the best with nRMSE of 27% and for h+6 RF performed the best with an nRMSE of 45.1%. In Summers for time horizon h+1 and h+6 RF performed the best with an nRMSE of 24.5% and 49.2% respectively. For Winters, for h+1 SGB performed the best with nRMSE of 31.1% and for h+6 RF performed the best with an nRMSE of 32.4%.

4.14 Seasonal Performance Comparison of Models for DNI prediction

In this section, a seasonal performance comparison of models for DNI prediction are presented. Figure 30 shows the comparison of different model for time horizon h+1 and h+6 by Autumn, Spring, Summers, and Winters.

Here in order to compare and rank the models seasonally it evaluation metric nRMSE is used as the seasons change the average hourly irradiation keep changing accordingly and they tend to be different. From the figure 30, for time horizon h+1 it is observed that GHI component is predicted with a good accuracy in Winters, and Autumns using RF model with an nRMSE of 10.10%, and 8.90% respectively. And RF proved to be better in Summers and Springs with an nRMSE of 9.20%. and 9.80% respectively. For h+6, SGB proved to be better at predicting in Autumn, Spring and Winters with an nRMSE of 16.80%, 13.30 and 18.0% respectively. And RF proved to be good in Summers with an nRMSE of 15.0%. Overall in both the horizon RF and SGB proved to be the best for prediction of DNI seasonally. With this section, the research objective 6 set in section 1.3.2 is achieved.

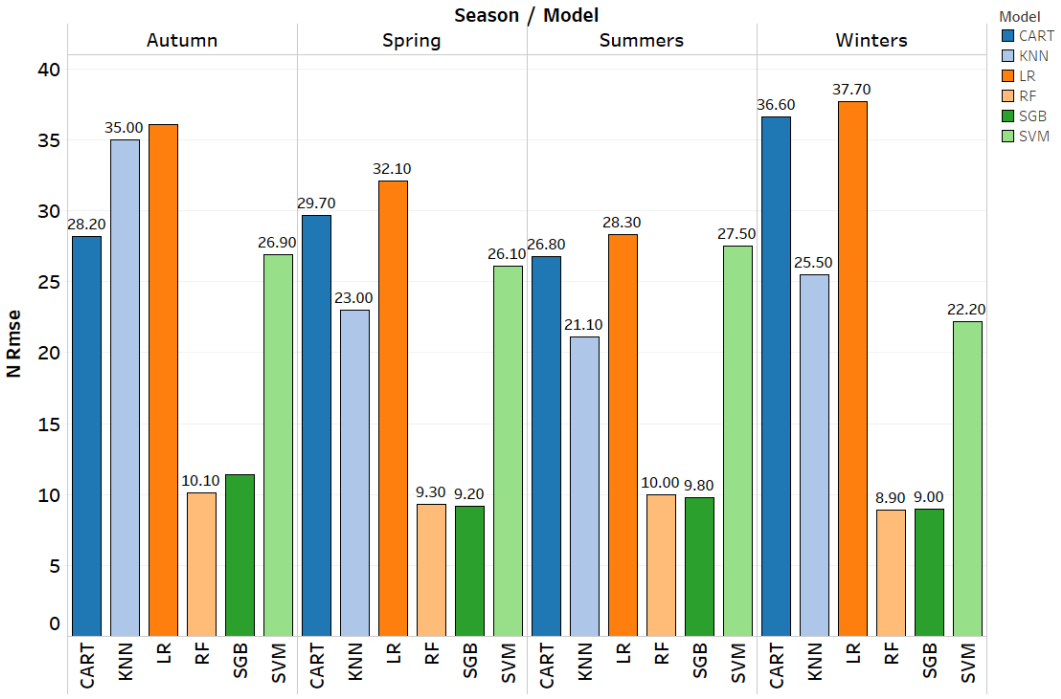

Figure 28: For h+1

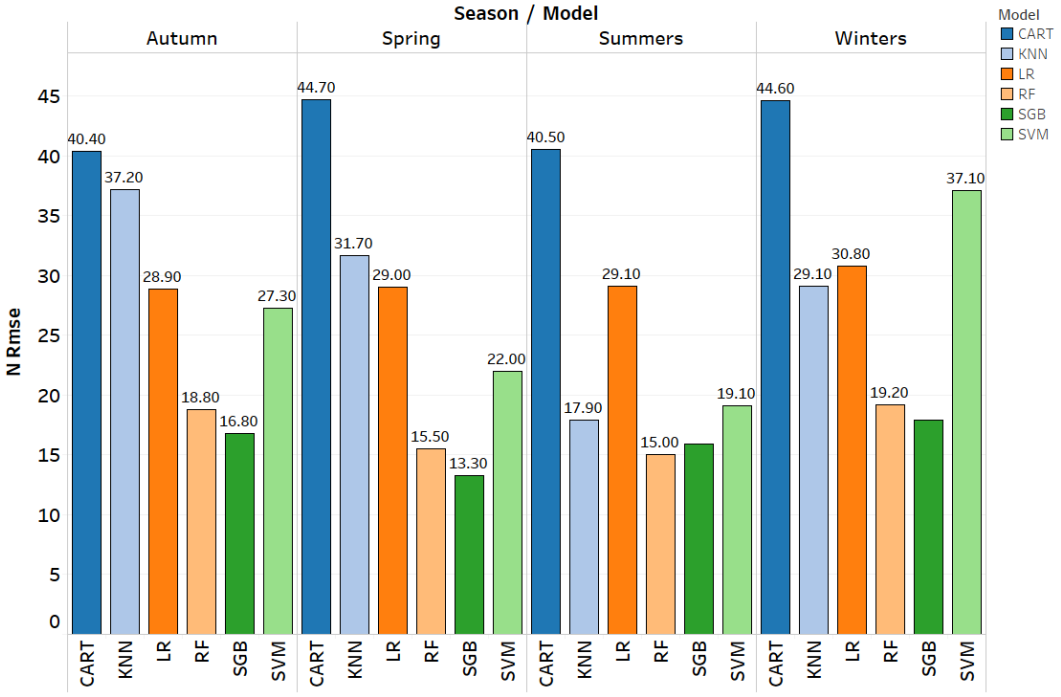

Figure 29: For h+6

Figure 30: Seasonal Performance Comparison for DNI

4.15 Comparison of Developed Models with Existing Models

The models developed in this project outperforms the models developed by Benali et al. (2019). The results are presented in the Table 2 as shown below. The results highlighted in red color are the improved results with the addition of meteorological measures. With this the objective 7 in section 1.3.2 has been achieved which partly helps in answering the research question as in

section 1.3.1.

Table 2: Comparison of Developed Models with Existing Models

Models	nRMSE for h+1	nRMSE for h+6
Random Forest Model for GHI prediction Benali et al. (2019)	19.65%	27.78%
Developed Random Forest for GHI prediction	7.3%	10.60%
Random Forest Model for DHI prediction Benali et al. (2019)	35.08%	49.08%
Developed Random Forest for DHI prediction	25.50%	32.50%
Random Forest Model for DNI prediction Benali et al. (2019)	34.12%	49.14%
Developed Random Forest for DNI prediction	11.70%	16.30%

4.16 Discussion of the Results

Six predictive models namely CART, KNN, LR, RF, SGB, and SVM were used in order to predict the solar irradiation's components using additional meteorological measures at the site of Odeillo, France. The predictions were done at time horizon h+1 to h+6. From the results it was concluded that Random forest and Stochastic Gradient Boost predicted GHI, DHI and DNI with good accuracy.
- For GHI, and nRMSE of 7.3% for h+1 and 7.9% for h+6 was obtained with RF. And with SGB it was 7.9% and 11% respectively.
- For DHI, and nRMSE of 26% for h+1 and 33% for h+6 was obtained with RF. And with SGB it was 25% and 33% respectively.
- For DNI, and nRMSE of 12% for h+1 and 11% for h+6 was obtained with RF. And with SGB it was 16% and 16% respectively.
A study based on the seasonal performance comparison was also conducted. It was observed that DHI and DNI are more difficult to predict as compared to GHI as DHI and DNI are more prone to changes in the meteorological conditions. Seasonal performance comparison concluded that solar irradiation components prediction accuracy decreases in the winters and autumns as compared to summer and spring.

5 Conclusion and Future Work

The objectives for this research project was to predict the individual components of the solar irradiation namely GHI, DHI, and DNI using additional meteorological measures to increase the predictive accuracy. To find an answer to the research question mentioned in section 1.3.1 a series of steps were taken. Firstly the data was taken on an hourly basis and this data was explored by using correlation matrix and added variable plots. Important attributes of the data were taken into consideration by the use of feature selection using Random forest and Boruta algorithm. This data was then divided into six time horizon and later according to four seasons in order to understand the performances of models seasonally. Six machine learning models namely CART, SVM, RF, SGB, LR, and KNN were used to predict the solar irradiation components. The performances of all these models were compared using four evaluation metrics namely RMSE, nRMSE, MAE, and nMAE. Amongst all the models Random forest and Stochastic gradient boost model predicted the solar irradiation component much more accurately. A significant improvement in prediction was observed with the addition of meteorological

measures. And the results were also proved to be better than the results obtained by Benali et al. (2019) as shown in the table 2. Seasonal performance comparison was also performed and it was observed that solar irradiation components prediction accuracy decreases in the winters and autumns as compared to summer and spring and this can be due to a higher meteorological variation during those seasons. It was also observed that the GHI component is easy to predict as compared to the other two components namely DHI and DNI and the reason being its sensitivity to meteorological variations. From all these conclusions drawn above, it can be said that the research question as in section 1.3.1 is being answered. So to conclude, this project can help in improving the accuracy of the model and thereafter help the firms dealing with solar plants installation to decide where and which kind of solar plant to setup.
In future, these results can be validated at other sites with much different weather conditions. Also different optimization techniques could be employed to improve the results of the models presented in this research.

6 Acknowledgements

I am extremely grateful to Dr. Catherine Mulwa for her continuous support and guidance throughout my research project. I would also like to thank my parents and my friends for supporting me in every phase of my life. A big thanks to Solcast for providing solar irradiation data at a granular level for my academic research project.

References

Alobaidi, M. H., Marpu, P. R., Ouarda, T. B. and Ghedira, H. (2014). Mapping of the solar irradiance in the UAE using advanced artificial neural network ensemble, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(8): 3668–3680.

Benali, L., Notton, G., Fouilloy, A., Voyant, C. and Dizene, R. (2019). Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components, *Renewable Energy* **132**: 871–884.
URL: *https://doi.org/10.1016/j.renene.2018.08.044*

Black, K., Davis, P., Lynch, P., Jones, M., McGettigan, M. and Osborne, B. (2006). Long-term trends in solar irradiance in ireland and their potential effects on gross primary productivity, *Agricultural and Forest Meteorology* **141**: 118–132.

Blanc, P., Espinar, B., Geuder, N., Gueymard, C., Meyer, R., Pitz-Paal, R., Reinhardt, B., Renné, D., Sengupta, M., Wald, L. and Wilbert, S. (2014). Direct normal irradiance related definitions and applications: The circumsolar issue, *Solar Energy* **110**: 561–577.

Bouchouicha, K., Hassan, M. A., Bailek, N. and Aoun, N. (2019). Estimating the global solar irradiation and optimizing the error estimates under Algerian desert climate, *Renewable Energy* **139**: 844–858.
URL: *https://doi.org/10.1016/j.renene.2019.02.071*

Chu, Y. and Coimbra, C. F. (2017). Short-term probabilistic forecasts for Direct Normal Irradiance, *Renewable Energy* **101**: 526–536.
URL: *http://dx.doi.org/10.1016/j.renene.2016.09.012*

Fan, J., Wu, L., Zhang, F., Cai, H., Wang, X., Lu, X. and Xiang, Y. (2018). Evaluating the effect of air pollution on global and diffuse solar radiation prediction using support vector machine modeling based on sunshine duration and air temperature, *Renewable and Sustainable Energy Reviews* **94**(June): 732–747.
URL: *https://doi.org/10.1016/j.rser.2018.06.029*

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *AI Magazine* **17**(3): 37.
URL: *https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230*

Fouilloy, A., Voyant, C., Notton, G., Motte, F., Paoli, C., Nivet, M. L., Guillot, E. and Duchaud, J. L. (2018). Solar irradiation prediction with machine learning: Forecasting models selection method depending on weather variability, *Energy* **165**: 620–629.
URL: *https://doi.org/10.1016/j.energy.2018.09.116*

Francis M. Lopes, Hugo G. Silva, R. S. A. C. P. C. M. C.-P. (2018). Short-term forecasts of GHI and DNI for solar energy systems operation: assessment of the ECMWF integrated forecasting system in southern Portugal, *Solar Energy* **170**(January): 14–30.
URL: *https://doi.org/10.1016/j.solener.2018.05.039*

Jamil, B. and Siddiqui, A. T. (2018). Estimation of monthly mean diffuse solar radiation over India: Performance of two variable models under different climatic zones, *Sustainable Energy Technologies and Assessments* **25**(January): 161–180.

Kalnay, E. (2002). *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press.

Li, Z., Boyle, F. and Reynolds, A. (2011). Domestic application of solar PV systems in Ireland: The reality of their economic viability, *Energy* **36**(10): 5865–5876.
URL: *https://ideas.repec.org/a/eee/energy/v36y2011i10p5865-5876.html*

Notton, G., Voyant, C., Fouilloy, A., Duchaud, J.-L. and Nivet, M.-L. (2019). Some applications of ann to solar radiation estimation and forecasting for energy applications, *Applied Sciences* **9**.

Pedro, H. T. and Coimbra, C. F. (2015). Nearest-neighbor methodology for prediction of intra-hour global horizontal and direct normal irradiances, *Renewable Energy* **80**: 770–782.
URL: *http://dx.doi.org/10.1016/j.renene.2015.02.061*

Pedro, H. T., Coimbra, C. F., David, M. and Lauret, P. (2018). Assessment of machine learning techniques for deterministic and probabilistic intra-hour solar forecasts, *Renewable Energy* **123**: 191–203.
URL: *https://doi.org/10.1016/j.renene.2018.02.006*

Qing, X. and Niu, Y. (2018). Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM, *Energy* **148**: 461–468.
URL: *https://doi.org/10.1016/j.energy.2018.01.177*

Ramírez, L. and Vindel, J. M. (2016). *Forecasting and nowcasting of DNI for concentrating solar thermal systems*, Elsevier Ltd.
URL: *http://dx.doi.org/10.1016/B978-0-08-100516-3.00013-7*

Schroedter-Homscheidt, M., Oumbe, A., Benedetti, A. and Morcrette, J.-J. (2013). Aerosols for concentrating solar electricity production forecasts: Requirement quantification and ecmwf/macc aerosol forecast assessment, *Bulletin of the American Meteorological Society* **94**: 903–914.

Sun, H., Gui, D., Yan, B., Liu, Y., Liao, W., Zhu, Y., Lu, C. and Zhao, N. (2016). Assessing the potential of random forest method for estimating solar radiation using air pollution index, *Energy Conversion and Management* **119**: 121–129.
URL: *http://dx.doi.org/10.1016/j.enconman.2016.04.051*

Wu, L., Chen, B., Fan, J., Xiang, Y., Zhang, F. and Lu, X. (2017). Evaluation and development of temperature-based empirical models for estimating daily global solar radiation in humid regions, *Energy* **144**: 903–914.

Yang, X.-S. (2012). Hourly solar irradiance time series forecasting using cloud cover index, *Solar Energy* **86**(12): 3531–3543.
URL: *http://dx.doi.org/10.1016/j.solener.2012.07.029*