Comparison and evaluation of empirical and machine learning models in estimating global solar radiation in Limpopo province

Thalukanyo Witney Murida

Student No: 14000953

Supervisor: Dr T.S Mulaudzi
Co-supervisors: Dr N.E Maluta and Mr N Mphephu

This research project is presented as partial fullfiment of the requirements of the degree:

Master of Science in Physics

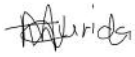
at the University of Venda
Faculty of Science, Engineering and Agriculture

15 August 2023

Declaration

*I, **Thalukanyo Witney Murida**, declare that this research tittle: **Comparison and evaluation of empirical and machine learning models in estimating global solar radiation in Limpopo province**, submitted in partial fulfilment of the requirements for the conferral of the degree **Master of Science in Physics**, from the University of Venda, is my work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.*

I further declare that I have not previously submitted this work, or part of it, for examination at University of Venda for another qualification or at any other higher education institution.



14/08/2023

Ms T.W Murida *Date*

(Student)

Abstract

This study investigated the performance of machine learning techniques as compared to the empirical models to forecast the global solar radiation in Limpopo regions. The machine learning techniques used in this study are Support Vector Machines, Random Forest, and Artificial Neural Network, and the empirical models used are the Clemence and Hargreaves-Samani models. To assess the efficiences of the machine learning models against the empirical models, the researchers calculated and compared the models performance evaluation using statistical equations such as Coefficient of determination, Mean Square Error, Mean Absolute Error, and Root Mean Square Error. Calibaration was done to improve performance of the empirical models. The present study found that machine learning techniques perform better than the empirical models when estimating the global solar radiation in the selected Limpopo regions.

Keywords: *Machine Learning, Empirical models, Random Forest, Support Vector Machines, Artificial Neural Networks*

Acknowledgments

On everything, I really like to thank God as the capability to further my studies, not to mention the motivation and encouragement I got from my supervisor Dr TS Mulaudzi and my co-supervisors Mr N Mphephu and Dr NE Maluta. I would further like to extend my thanks to the University of Venda for all the resourses I used to complete my work and Agriculture Research Council (ARC) for providing me with their data. Last but not least, I thank my family and friends for the support along the journey of their studies.

Contents

Chapter 1

Introduction

1.1 Background

There has been consensus that several countries globally are confronted with significant energy crises. As energy consumption rises, the world will face a substantial shortage of fossil fuels in the future because such sources provide most of the world's energy (Dorian et al., 2006). The existing consumption models across the globe are based entirely on non-renewable energy sources such as coal, gas and oil, however in recent times, there has been a call for the world to move away from such energy generation fuels as they create environmental problems. The fundamental issue is the energy crisis, particularly acute in emerging countries where there is a need to power families and industrial sectors (Kessides, 2013). It is estimated that in the next 30 years, the world will be overpopulated and energy demand will also increase. The majority of the world's population is unaware of the importance of preserving energy for future generations.

The economies of developing countries that are highly industrialized, for example, South Africa, have a high energy demand. As the primary power provider in South Africa, Eskom relies on environmentally hazardous fossil fuels (Pegels, 2010). Carbon dioxide emissions from burning fossil fuels, contribute to global warming and harm biodiversity and the ecosystem (Herzog et al., 2000). The best solution to this challenge is for the world to move away from non-renewable resources and use renewable resources. Solar energy is a clean, plentiful, renewable, and sustainable energy source that originates from the sun and travels to the earth as light and heat. (Panwar et al., 2011). In any given location, the demand for renewable and sustainable energy has increased.

Solar radiation from the sun is quickly becoming a viable alternative to other traditional en-

ergy sources. Solar energy looks to be the most popular alternative among the various forms of clean energy sources because of their endless and non-polluting nature (Khambalkar et al., 2010). Solar radiation is the world's oldest energy source, and virtually all fossil and renewable energy sources depend on it. Free and easily accessible solar energy could be used to reduce reliance on energy derived from fossil fuels (Chow et al., 2012). In the architecture of the solar system, accurate instruments for estimating solar radiation are essential.

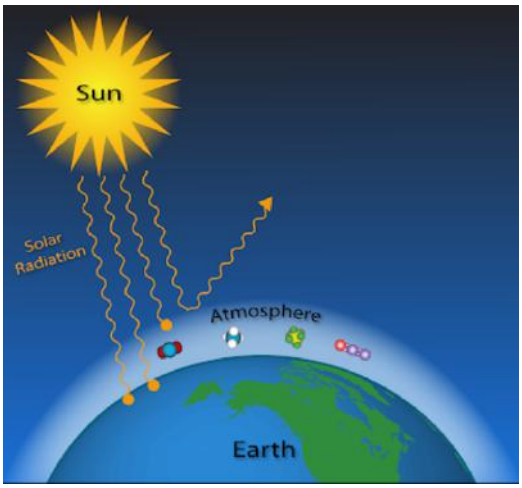


Figure 1.1: Solar radiation from the sun (Segalstad, 1998)

Solar radiation is the energy emitted by the sun as an electromagnetic wave (Wald, 2007). This energy energy has an impact on atmospheric and climatological processes, as well as phenomena such as photosynthesis. Furthermore, it is in charge of maintaining planet's temperaturestable for life and wind creation. It is a measurable energy source that may be utilized to heat buildings and objectsas well as collected by photovoltaic cells. It manifests itself in a variety of ways, including ultraviolet rays, visible light rays and infrared waves. Furthermore, this energy changes throughout time depending on weather and location. The amount of solar radiation that reaches the Earth's surface is affected by a variety of factors, including the time of the day, season, geographical location and local weather (Shukla et al., 2016). Solar radiation is very important for the following reasons: It emits light. It maintains the Earth's temperature by supplying heat. It emits energy that can be collected and transformed into power and also required for photosynthesis.

Around 71% of the sunlight that reaches the earth is absorbed by its surface and atmosphere (Ångström, 1929). The temperature of an object or surface rises as a result of sunlight's absorption, which causes the molecules it hits to vibrate more quickly. The Earth then emits this energy as heat or long-wave infrared radiation. A surface warms up and radiates heat

more as the amount of sunshine it absorbs increases. The Earth is a sphere, not every portion of it receives equal amounts of solar energy (Lindsey, 2009). More solar radiation is received and absorbed near the equator than at the poles. Light is primarily reflected if the surface does not absorb it. When incoming solar radiation meets an item or character in the atmosphere or on the ground, it bounces back and is not converted into heat.

Solar radiation predictions for specific regions are critical for guiding solar power conversion systems, focusing on design, modelling and operation (Moore-O'Leary et al., 2017). Furthermore, for decision-makers responsible for future green energy investment regulations, selecting appropriate places with sufficient solar radiation is critical. The solar system's power production is determined by solar radiation, it is crucial to have a reliable solar radiation prediction to assure the power grid's proper and sufficient operation.

However, the solar irradiance forecast depends on the available data and the forecasting methodologies used. Although empirical models have been developed and employed in the past, high-frequency data availability has lately shown that intelligent algorithms offer more prediction capability (Kearns and Nevmyvaka, 2013). In light of this, the Clemence and Hargreaves-Samani models will be used in this work to empirically predict the global solar irradiance surrounding the Vhembe district of the South African province of Limpopo.

In addition, machine learning and deep learning techniques, namely: support vector machines (SVM), random forest (RF), and artificial neural networks (ANN) will also be employed to predict global solar irradiance in the same areas. The predicted values for global solar radiation will be compared to the recorded data from the weather stations in these regions in order to evaluate the efficacy of these empirical and machine learning models.

1.2 Problem statement

Solar radiation data are extremely useful for designing and implemening various renewable energy techniqies in any given area. Contrastingly, due to the lack of weather stations in developing countries and rural regions, in many situations where there are no observed data, empirical methods have been employed to approximate the solar radiation. The use of the various empirical models has its limitations due to multiple parameters needed. Due to the limited measurement coverage, there is a need to develop better methods with low data requirements and low computational costs for estimating solar radiation.

1.3 Research aims and objectives of the study

This study compares and evaluates global solar radiation in specific regions around Limpopo using empirical models and various machine learning algorithms.

The objectives of the study are to:

- Explore the deficiencies of the empirical models in estimating global solar radiation.
- To calibrate the parameters of the empirical models for the specific study area.
- Estimate the solar radiation using empirical models and machine learning models.
- Obtain better machine learning models through hyper-parameter tuning.
- Compare the performance of empirical models and machine learning models using the experimental data

1.4 Significance of the study

Since the amount of solar radiation in the world is a crucial parameter for the majority of ecological models and is used as input for various photovoltaic conversion systems, renewable energy sources are commercially significant. The amount of solar radiation that reaches Earth's surface depends on the local climate. For a solar energy system to be accurately predicted and designed, this is crucial. Global solar radiation data for a given area must be known for better and more accurate estimations. This can be attained through a perfect prediction models that is to be assessed in this study.

1.5 Motivation for the study

One of the most popular research areas has been solar radiation, which has emerged as a benchmark for improving energy generation methods that employ renewable energy sources. In order to determine the monthly average daily global solar radiation, empirical models have been developed. Still, due to sudden changes in meteorological conditions, such as cloud cover and wet days, these models are unable to accurately estimate short-term solar radiation. These models produced partially satisfactory forecast results for global solar radiation. However, due to the availability of high-frequency data, intelligent algorithms have recently been demonstrated to have more predictive capacity.

This technology makes it possible to integrate solar energy into the grid, which has good

effects by raising the standard of power delivered to the grid and decreasing the price of solar-related accessories. The combination of these factors has fueled the development and design of models in a complex field of research aimed at improving solar resource predictions and, as a result, being able to predict the amount of output power that can be generated based on the technology used and the period use in order to improve solar resource predictions and, as a result, be able to forecast the amount of output power that can be created based on the technology and time taken. The complicated field of research has been driven by the combination of these aspects. In this work, global solar radiation forecasting models will be estimated using empirical and machine learning techniques, and their performances will be compared to the observed data from meteorological stations.

1.6 Research question

As highlighted in the problem description, the solar energy generation system is reliant on output power predictions. From this it can already be observed that to have a good solar power generation system, one needs to predict solar radiation. Also pinpointed that the prediction models are categorized into empirical and machine learning. Against this background, this study was conducted under the following research question:

- Can machine learning models predict global solar radiation accurately?

Chapter 2

Literature Review

2.1 Background

Numerous studies on the estimation and prediction of global solar radiation have been carried out recently using various machine learning techniques, including artificial neural networks (ANN), random forests (RF), support vector machines (SVM), kernel nearest neighbors (K-NN), and an adaptive network-based fuzzy inference system (ANFIS), as well as empirical models that are sunshine hour based models and temperature-based models globally in various areas (Kaba et al., 2018). With each passing day, solar energy, which contributes significantly to the production of electricity, has emerged as one of the most promising renewable energy sources gaining experts' interest since, in contrast to fossil fuels, it is clean, limitless, and sustainable. As a result, it is the most precise way of obtaining long-term data in a specific geographic area (Dincer, 2011).

However, measuring the solar radiation (SR) everywhere is generally impossible due to the high cost, time, and precision required. Furthermore, most countries cannot accurately assess radiation levels because measurements can only be taken in specific areas. Consequently, methodologies based on experimental, statistical, and Artificial Intelligence (AI) have been developed to estimate the value of SR globally (Vakili et al., 2017). Machine learning (ML), a branch of AI, is one of the most systematic approaches to measuring global solar radiation.

To design any solar energy conversion equipment and determine the viability of any solar energy application, solar radiation measurement is always necessary. Solar energy is the most plentiful source of renewable energy and is also secure and clean (Mohtasham, 2015). Total solar radiation statistics must be understood to carry out the study. The models of daily

solar radiation consider the effect of time, as well as variations in weather conditions and the distributions of hourly total solar radiation through time, resulting in a normal distribution. To estimate global solar radiation, the total solar radiation from several local sites must be available (Korachagaon and Bapat, 2012).

2.2 Empirical Models

Empirical models are based on correlations obtained from the analysis of experimental data (Shi et al., 2016). The experimental solar radiation data is not available for all locations. Therefore, numerous estimating procedures have been developed to determine the total solar radiation. Modelling is a more accurate approach for estimating global solar radiation in specific locations (Hassan et al., 2016). The researchers used empirical models such as Angstrom(sunshine hour-based model), Clemence(temperature-based model) and Hargreaves-Samani(temperature-based model) to estimate the radiation.

The amount of solar energy present there must be known before using global solar radiation to generate electricity at any one site (Montes et al., 2009). The instrumentation required to measure global solar radiation is expensive and dangerous. However, the majority of meteorological stations lack access to information on solar radiation, which has an impact on many stations in developing countries. There must be additional techniques available for estimating the needed data. These techniques rely on the accuracy and volume of the measurable data used. It is also regarded as a useful tool for generating global solar radiation for areas where measured data are unavailable or absent altogether.

To estimate global solar radiation, a number of models have been developed. (Amprat-wum and Dorvlo, 1999) has used original(linear) and modified(linear logarithm) Angstron Black type regression models with quadratic, a power relationship and a power-trigonometric model to estimate the global solar radiation. In this study they found that all the models perform well as estimators of global solar radiation from sinshine hours, but the non-linear models were better than the basic linear models. The global solar radiation data was correlated with the relative sunlight length using a modified Angstrom type model called a linear regression model. In this study, they found that the two relations shows that the effect of the lattitude and altitude can be ignored. They also found that the bright sunshine hours are correclated linearly with the global solar radiation. Aad et al. (2008) studied an existing relationship between monthly mean, clearness index and the number of bright sunshine hours using the data obtained from Romania. Trabea and Shaltout (2000)examined the relationship between measured global solar radiation and meteorological parameters derived from

solar radiation in various parts of Nigeria, including mean daily maximum teprature, relative humidity, sea level pressure, vapour, and bright sunshine hours. Their resuts show that the model were satistically significant and could be used to estimate solar radiation at that region. Sfetsos and Coonick (2000) use artificial intelligence techniques to forecast hourly global solar radiation in the same area. In the current work, global solar radiation was estimated using the temperature-based models developed by Hargreaves-Samani and Clemence. In this study, they found that the developed artifical intelligence models predict the solar radiation time series more effectively compared to the conventional procedures based on the clearness index. In Pretoria, Maluta and Mulaudzi (2018) researched the effectiveness of temperature-based models for estimating solar radiation from the sun. In this study they found that the two temperature based models have potential to estimate the global solar radiation in the area under study.

Multiple empirical models have been developed to calculate global solar radiation using a variety of parameters. Angstrom's first model for projecting global radiation used data on the amount of sunshine and the radiation from clear skies. The other method to estimate solar radiation was obtained by using atmospheric transmittance model (Podestá et al., 2004) while other authors have used diffuse fraction (Reindl et al., 1990) and clearness index models. Information on specific atmospheric characteristics is needed for the parametric or atmospheric transmittance model (Wong and Chow, 2001). Since this model performs well in clear skies and without clouds, several authors have used it to assess how well an empirical model performs best in warm weather.

2.3 Machine Learning Models

As new application fields develop, ML, a dominant AI form, gains appeal and acceptance (Antonopoulos et al., 2020). ML gives computers the ability to interpret and estimate unknown outputs on their own. The ML algorithm's performance has been dramatically influenced by the attributes chosen and the algorithm's training success. AI has become increasingly popular in practically all technical sectors in recent decades as a result of technological advancements (PK, 1984).

SVM, deep learning (DL), K-NN, ANN and other AI are examples of ML models for predicting solar radiation data that have become increasingly popular. The literature has revealed that when it comes to estimating solar radiation, AI systems are more accurate than other methods (Ağbulut et al., 2021). Quej et al. (2017) used three ML algorithms, namely, SVM, ANN, and adaptive network-based fuzzy inference system (ANFIS) to forecast daily global

solar radiation data from six sites in Mexico. For the training of the algorithms, the authors employed extraterrestrial solar radiation, rainfall, minimum temperature, and temperature data, the best results were obtained in SVM.

Machine learning has lately become quite popular because it aims to make machines and electronic applications learn in the same manner that human brains and biological neural systems do (Aggarwal et al., 2018). Subsequently, this heuristic approach can be utilised to get the optimal solar radiation estimation in the field of solar radiation prediction. Global solar radiation is extensively dispersed throughout the world, but because it is challenging to construct meteorological stations everywhere, alternate approaches must be developed to replace expensive monitoring tools. Consequently, ML procedures are used because they produce more accurate results than traditional methods (Voyant et al., 2017).

In several disciplines, ANNs are the most prevalent and recently utilized ML approach (Kourou et al., 2015). They are increasingly being utilised to estimate solar radiation in a variety of investigations, providing reliable results when compared to traditional methods with few errors. It has been demonstrated that 79% of ML approaches used in weather forecasting problems are based on ANN, especially with the absence of meteorological stations in many parts of the world. This heuristic approach may be used to estimate optimal solar radiation. Elminir et al. (2006) also employed multilayer perceptron neural networks(MLPNN) to estimate global solar radiation at control sites in Helwan, Egypt. In addition, Angela et al. (2011) employed ANN to predict daily global solar radiation in Kampala, Uganda, based on a single variable, the sunshine duration, in a matter of monthly average values on a horizontal surface. Applying 65 neurons with the tansig transfer function gave the best results in the study.

Another popular machine learning technique is regression tree (RT), which is well known for its high effectiveness and is used in a number of technical, medical, and environmental difficulties (Rani et al., 2006). Unfortunately, it is rarely used in global solar radiation prediction. Junior et al. (2016) on the other hand, conducted a study of the feature selection problem for global solar radiation estimation in Tokyo, Japan. The study demonstrated the significance of global solar radiation, sunshine duration, and sun altitude in the authors' decision to anticipate solar radiation. In various domains of renewable energy, RT has been shown to be effective. By measuring it in a number of wind towers, Troncoso et al. (2015) demonstrated that hourly short-time wind speed forecast in Spain using RT worked effectively. RT fared better than support vector machines, multilayer perceptrons, multilinear regression, and other ML techniques (Troncoso et al., 2015).

Furthermore, in several studies, SVR has been shown to be more accurate than ANN and

regression statistical models in a variety of prediction (Balabin and Lomakina, 2011). Even though the fact that the SVR approach can be used to predict variables in a variety of engineering applications, there are few studies that have used it to estimate solar radiation. Ghasemi Mobtaker et al. (2016) introduced a research study on the estimation of global solar radiation using SVR. He compared the fuzzy linear regression method and SVR to forecast daily global solar radiation in Tehran, Iran, by using inputs such as daylight hours, Julian days, the minimum and maximum amounts of sunshine, clear-sky solar radiation, and extraterrestrial radiation. The study also elucidated that SVR outperforms fuzzy linear regression in terms of performance.

In a different study, Ağbulut et al. (2021) attempted to predict the daily global solar radiation of 13 distinct locations. In that work, the authors exclusively employed ANN as a machine learning algorithm. The best results were obtained in the study when extraterrestrial solar radiation, minimum temperature, and maximum temperature were used to train the ANN system. Gene expression programming (GEP), ANN, and ANFIS were the three distinct models that Ağbulut et al. (2021) utilized to calculate the daily global solar radiation in Karmen, Iran. In the pertinent study, the ANN model delivered the best results.

2.4 Gaps Identified

For calculating the amount of global sunlight at certain areas, modeling is a more accurate method. Empirical models have frequently been used to calculate the monthly average daily global solar radiation (Chen et al., 2013). To calculate the monthly average daily global solar radiation, empirical models have been frequently utilised. However, because weather conditions can quickly change due to cloud cover, rainy days, and other factors, these models are unable to accurately estimate short-term solar radiation (Golestaneh et al., 2016).

Solar radiation measurement equipment are used to measure solar radiation. This equipment, on the other hand, has high initial and ongoing costs, as well as calibration requirements (Tassey, 2000). As a result, they are not sufficiently available at most stations around the world. Due to this, it is crucial to forecast solar radiation data using measurably accurate environmental factors like humidity, temperature, wind speed, and cloud cover. To forecast solar radiation data from this perspective, few models have been presented. Due to some of them are based on mathematical formulas, they are known as empirical models (Jiang, 2009).

According to Jiang et al. (2019), in humid places where solar radiation is significantly influenced by dense clouds on rainy days, empirical models are unable to depict the complex and

non-linear relationships between dependent and independent variables. In previous investigations done by Ağbulut et al. (2021), it was revealed that these empirical models produced partially satisfactory forecast results for daily global solar radiation. Although ML models are better than empirical models in predicting solar radiation, comparisons of the prediction accuracy of various ML models, particularly their computational efficiency on sizable datasets for solar radiation prediction, have not frequently been conducted globally.

Angstrom model and ANN approach were compared to predict global solar radiation data by et al., (Behrang et al., 2010). The authors discovered that when compared to the Angstrom model, the ANN technique produced higher prediction results. In a different study, Meenal and Selvakumar (2018) looked at the daily global solar radiation using SVM, empirical, and ANN models. SVM gave the greatest results out of all of these models, with a correlation of above 0.99. Premalatha and Valan Arasu (2016) forecasted monthly global solar radiation for four different places in Turkey using regression analysis and ANN. In the relevant study, the algorithm was trained using extraterrestrial solar radiation, longitude, maximum possible sunlight duration, sunshine duration, relative humidity, temperature, and day of the year. In the ANN model, the authors' results are the best.

Chen-Min Li (1978) used SVR to estimate the empirical Angstrom-Prescott models for daily GSR in several Chinese cities utilizing inputs from different sunlight duration combinations. The study unequivocally demonstrated that the SVR approach produced more reliable findings than the empirical models. Recently, improved and evolved SVR models were discovered to more accurately estimate hourly global solar radiation (HGSR) by utilizing (Zendehboudi et al., 2018) introduced penalized SVR and forward regression on a quadratic kernel SVM. As stated in the study's objective, the goal of this project is to compare machine learning and emprical models for predicting solar radiation. It is clear from the literature that machine learning learning models consistently produce better results than empirical models. To enhance empirical models' poor performance for future prediction, this work will calibrate them.

Chapter 3

Exploratory Data Analysis

3.1 Introduction

This study used the data that was collected by the Agricultural Research Council (ARC) for five different locations in Limpopo province. Exploratory data analysis is basically the first step that should be done in the data analysis process and it is used to understand and summarise the contents of a dataset to prepare for a good and accurate results when performance measures are made. In this chapter a full explanation is given for data analysis. Section 3.2 covers data and geographical description for this work, data cleaning and transformation of data is discussed in Section 3.3. Under Section 3.4 a description of data visualization is also given in more detail.

3.2 Data and Geographical Description

Limpopo Province, located in South Africa, has been considered an area of study and situated in the northern South African province. It has the name of the Limpopo River, which forms the western and northern borders of the province. Polokwane is the province's capital and largest city, while Lebowakgomo is where the provincial legislature is located. Limpopo is situated in the northern part of South Africa with geographical location of 23.4013° S (Latitude) and 29.4179° E (Longitude). January is the hottest month in Limpopo with an average temperature of 22.5° and the coldest month is June at 12.5°. The wettest month is November with an average of $100mm$ of rain. Figure 3.1 shows some of the locations that are considered for the study.

Figure 3.1: Limpopo province
(Cai et al., 2017)

In Table 3.1 a geographical location for each area is presented. Lower altitude denote that the area receives more solar radiation than other areas, as according to the statistics provided in Table 3.1 below. Musina and Mutale are the hottest area from the altitude at 429 and 550 respectively.

Stations	Latitude	Longitude	Altitude	Min temp	Max temp	Climate type
Mussina	-22.45404	30.502259	429	20.60	34.75	Hot Semi-arid
Burgersfort	-24.58072	30.32378	709	19.64	35.90	Temperate interior
Marble Hall	-25.02642	29.36634	846	17.71	30.44	Temperate interior
Ammondale	-23.72619	29.59517	1153	8.93	25.93	Temperate interior
Mutale	-22.73461	30.52188	550	13.02	26.23	Hot Semi-arid

Table 3.1: Station information

This study use the collected data for five different locations and Table 3.2 below summarises the period for each site:

Area	Period
Marble Hall	2006 - 2022
Burgersfort	2002 - 2022
Ammondale	2002 - 2020
Mussina	2006 - 2022
Mutale	2006 - 2020

Table 3.2: Study area and period

From the collected data the following variables was recorded for each site and Table 3.3

below summarises all the recorded symbols, their description for each element and also their measuring unit. The same variables were recorded for all different locations;

Symbol	Description	Unit
T_x	Daily Maximum Temperature	^oC
T_n	Daily minimum Temperature	^oC
RHx	Average Daily Maximum Relative Humidity	%
RHn	Average Daily Minimum Relative Humidity	%
Rs	Total Radiation	MJ/m2
$U2$	Average Wind Speed	ms
Rain	Rainfall	mm
ETo	Total Relative Evapotranspiration	mm

Table 3.3: variable description

3.3 Data Preparation and Transformation

Data must be put into the proper form before analysis can take place. Data manipulation, which is frequently chaotic and unstructured, is transformed into a more useful and structured form that is available for further analysis through the iterative process of ”data preparation.” Data cleansing and transformation are among the tasks included in the entire preparation process. To make the data more acceptable for modeling and analysis, normalization was performed. When utilising gradient descent optimization, techniques like support vector machines and neural networks, the data must be normalised.

Data cleaning is the process of finding and eliminating all unnecessary variables or components from the database or table, locating incomplete, erroneous, and inaccurate values of the data, and then replacing and updating them in a way that makes it ready to be utilised.

When combining various data sources, there are several possibilities for data to be duplicated or improperly categorised. Even if the algorithms and results seem to be accurate, faulty data makes them unreliable. The majority of the work is spent on cleaning datasets or producing an error-free data set when it comes to using machine learning data. Setting up a quality control system is one of the best practices for data cleaning in machine learning include setting up a quality strategy, filling in missing information, deleting rows, and minimizing data size. There is no one, unambiguous way to identify the specific stage in the data cleaning process because the procedures will vary depending on the dataset. Contrastingly, making a template for the data cleaning process is crucial.

Data cleaning process done for each location is fully discussed below:

3.3.1 Treatment of duplicate and irrelevant observations

It is necessary to remove or fully delete all duplicate, unwanted, and undesirable observations from the dataset. Most duplicate observations happen when data are being collected. Data duplication might result from the merging of datasets from different sources. De-duplication is one of the crucial aspects of this process to consider. If there are observations, the are deemed unimportant for the study at hand if they do not fit in. To make the dataset more effective and manageable and to improve the effectiveness of the study, some observations must be discarded.

3.3.2 Handling missing values

The data value that is not stored for a variable in the relevant observation is known as missing values or missing data. Almost all research encounter the issue of missing data at some point, and it can significantly affect the inferences that can be made from the data. Given that many machine learning algorithms do not support missing values, the handling of missing data is crucial during the dataset's preprocessing. Rows or columns with null values can be removed to handle missing values. The missing values from the collected data will all be filled in using linear interpolation, and if there are more than 50% of missing values, it is recommended to drop all variables and any other considerations.

3.3.3 Remove unwanted outliers

The dataset's anomalous values are called outliers. Since they differ greatly from other data points, they may have an impact on how the results are analyzed. Whether to eliminate them depends on the data that are being analysed. Eliminating pointless outliers may also improve the performance of the data. The interquartile range(IQR) can be used to identify or detect any outliers in the dataset, and this can be done. The data set was totally purged of any and all outliers. Outliers were found and eliminated in the Marble Hall, Mussina, and Ammondale areas. To calculate IQR, the equation below can be used

$$\begin{aligned} IQR &= Upper\ Quartile - Lower\ Quartile \\ &= Q_3 - Q_1 \end{aligned} \tag{3.3.1}$$

where Q_3 refers to the third quartile and Q_1 refers to the first quartile.

3.4 Data Visualisation

The goal of data visualization is to identify trends, patterns, corrupt data, and outliers easily visible and understandable. By using visualization, we were able to identify any missing data. For instance, if a month or two is missing from the data, we can quickly see through visualize the absence of this data and look for outliers in the graphs. The tables and charts in the figures below provide an overview of all the input variables we utilized in the study and the goal variable for each study location. Table 3.4 and Figure 3.2 represent summary visualization for the Ammondale area. In Table 3.4, the average minimum temperature is $11,03^{o}$C, average maximum temperature is $26,24^{o}$C and the average rain is $1,102$mm.

$Stats$	T_x	T_n	RHx	RHn	$U2$	$Rain$	ETo	Rs
Mean	26.2441	11.0365	93.4338	37.0003	1.2336	1.1020	3.4620	17.6120
Std	4.5999	6.2207	8.0083	16.9437	0.8669	5.0229	1.2791	5.5581
Min	12.7000	-5.4000	39.2800	0.6600	0.0000	0.0000	0.0100	2.3800
25%	22.9800	5.7700	91.5000	23.8075	0.6300	0.0000	2.5200	13.8200
50%	26.5000	12.6000	96.7000	35.1000	1.2600	0.0000	3.3200	17.3500
75%	29.6900	16.3000	98.0000	48.4000	1.8400	0.0000	4.3200	21.4825
Max	38.6800	22.1000	100.0000	99.0000	4.5100	71.3800	7.3700	32.9600

Table 3.4: Statistical description for Ammondale

In Figure 3.2 below, minimum solar radiation for the area is 2.38, the maximum temperature of 32.960^{o}C and the average solar radiation of $5.558MJ/m^2$. The area receives low radiation during winter seasons as they increases when the season changes.

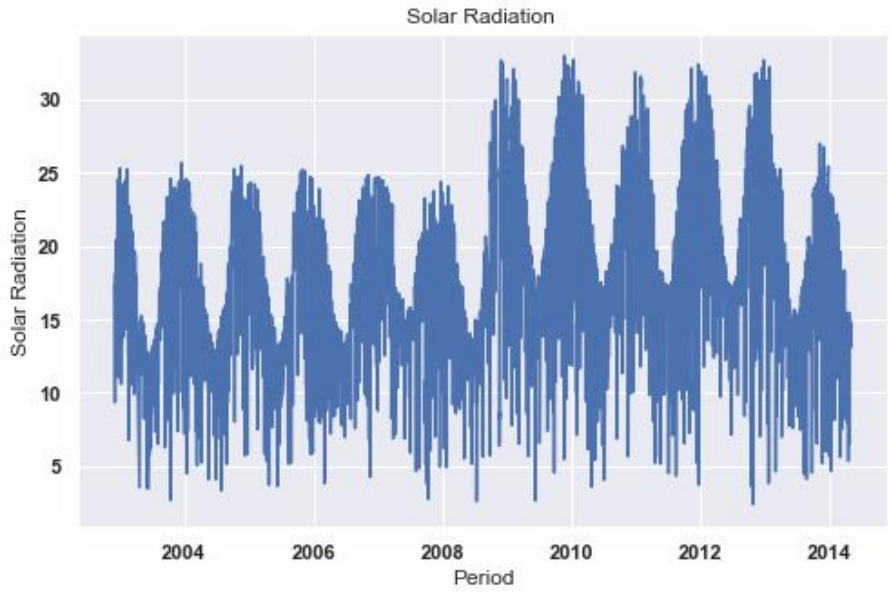


Figure 3.2: Global solar radiation observed for Ammondale

The table below shows the analysis for Burgersfort. This area receives a very minimal

amount of rainfall. The average rainfall for this area is 1.55mm. The maximum temperature is calculated 43.84^{o}C which is a clear indication that the area is very warm as the average of maximum temperature is shown to be 29.84^{o}C in Table 3.5.

$Stats$	T_x	T_n	RHx	RHn	$U2$	$Rain$	ETo	Rs
mean	29.8438	14.8441	78.5489	29.0171	0.7883	1.5516	3.8506	18.0984
std	4.6460	5.5120	11.2536	12.6450	0.4761	6.0820	1.6362	6.6180
min	14.6300	-0.6000	6.0000	3.0000	0.2000	0.0000	0.3600	0.0000
25%	26.3500	10.4500	72.6000	19.4775	0.4200	0.0000	2.5400	12.9900
50%	29.8250	16.1600	80.7000	27.8700	0.6800	0.0000	3.7000	17.5000
75%	33.3825	19.4200	87.2000	36.8275	1.0300	0.0000	5.0500	23.4900
max	43.8400	24.9200	96.8000	84.9800	2.9900	89.6600	30.8000	36.4800

Table 3.5: Statistical description for Burgersfort

The solar radiation is depicted in Figure 3.3 below, this area receives high average during summer seasons which are the end of the year and the beginning of the year. It can also be seen that the solar radiation drops during winter and changes from seasons to seasons. The average radiation in this area is $18.098MJ/m^2$ as reads from Table 3.5.

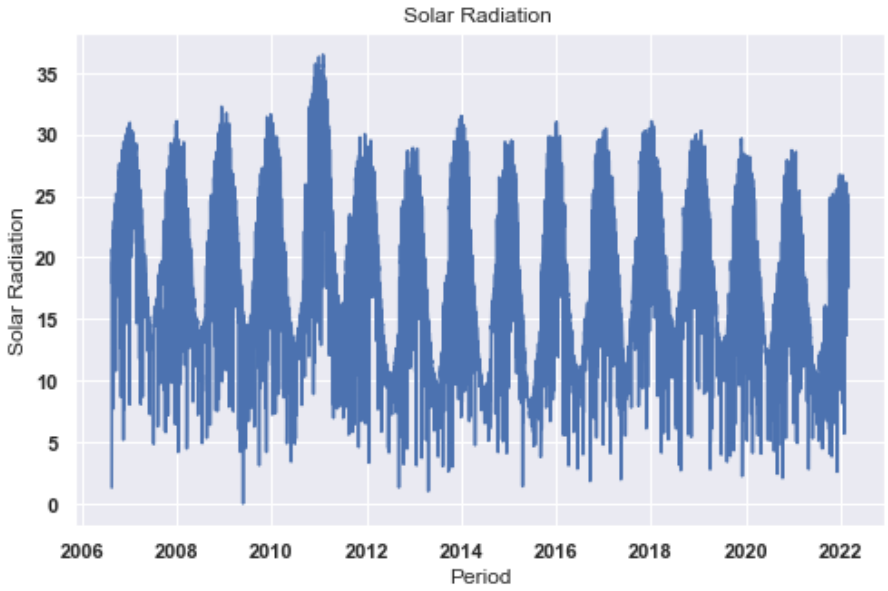


Figure 3.3: Global solar radiation observed for Burgersfort

The figures (Table 3.6 and Figure 3.4) show the summary visualization for all the input variables and the target variable plot for Marble Hall area. The average maximum temperature is 29.089^{o}C, and the minimum temperature is 12.433^{o}C. The average maximum relative humidity is 85.162% with a minimum average of 29.252%. The average rainfall for this area is $1.430MM$.

$Stats$	T_x	T_n	RHx	RHn	$U2$	**Rain**	ETo	Rs
Mean	29.0891	12.4337	85.1620	29.2526	1.3450	1.4308	4.0990	20.3960
Std	5.9914	6.5023	13.3263	12.6607	0.9755	5.3638	1.540504	6.2223
Min	6.8300	-15.7700	0.7000	5.4000	0.2000	0.0000	0.0000	1.820000
25%	25.5600	7.3300	82.1400	20.3100	0.8300	0.0000	2.8700	15.7100
50%	29.3400	14.0300	88.6000	27.9000	1.1600	0.0000	4.0100	19.8300
75%	32.5400	17.7400	92.1000	36.1333	1.5700	0.0000	5.3800	25.4700
Max	346.3000	24.6000	100.0000	90.5000	13.1300	117.8000	9.2700	36.6300

Table 3.6: Statistical description for Marble hall

Figure 3.4 shows the global solar radiation plot which learns the same pattern, during the end of each year and the beginning of each year which is summer season, the solar radiation is high during this time of the year and it also shows the times when the solar radiation drops.

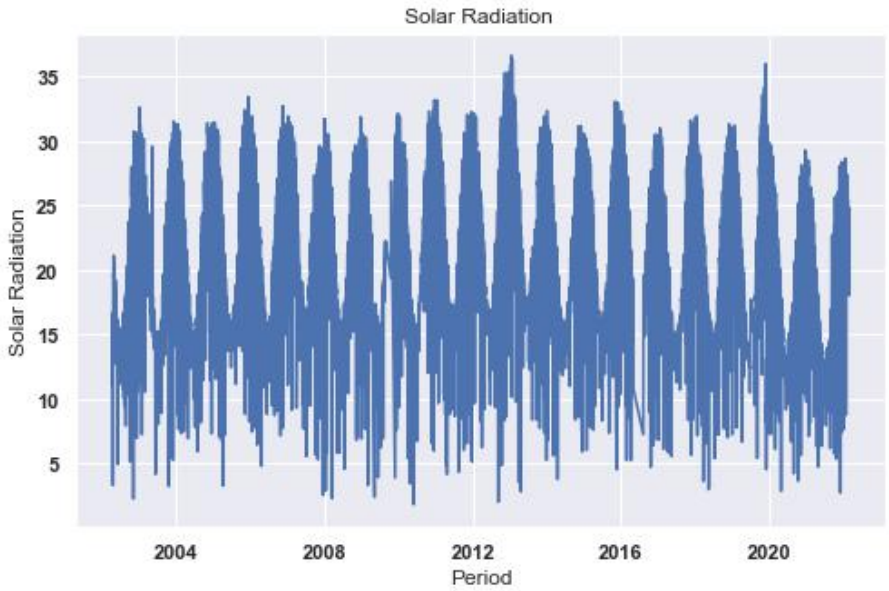


Figure 3.4: Global solar radiation observed Marble hall

The analysis for different variables in Mussina is shown in Table 3.7 and Figure 3.5 shows the solar radiation plot for the target variable which is solar radiation. The average solar radiation is $18.321 MJ/m^2$ as read from the figure below.

Stats	T_x	T_n	RHx	RHn	$U2$	$Rain$	ETo	Rs
Mean	31.2869	16.3274	82.0141	32.2822	0.9536	1.1887	3.9104	18.3217
Std	5.8988	5.7327	13.1067	14.3661	0.5862	7.0878	1.3737	5.6008
Min	12.1000	-13.2000	7.0000	0.0000	0.2000	0.0000	0.7000	3.4700
25%	27.6000	11.7750	73.9700	22.0150	0.5000	0.0000	2.9100	14.9300
50%	31.2100	17.4300	83.3700	30.3000	0.8000	0.0000	3.7700	18.2100
75%	34.7400	21.0900	92.7000	40.1000	1.2600	0.0000	4.9400	22.5000
Max	209.8000	29.4400	100.0000	94.8000	3.5700	246.8900	18.0700	31.0200

Table 3.7: Statistical description for Messina

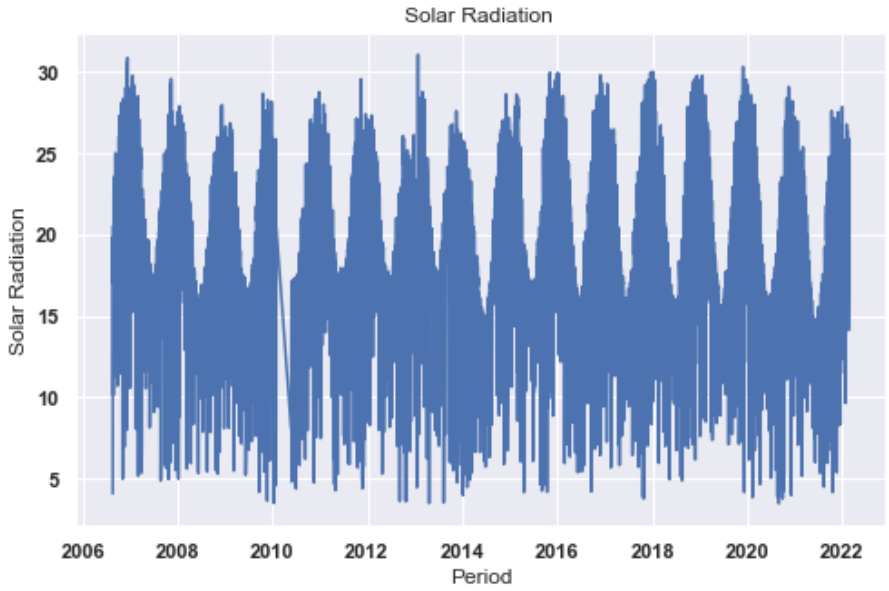


Figure 3.5: Global solar radiation observed for Messina

Table 3.8 describes the different input variables recorded for the study and output variable for Mutale area. Maximum and Minimum temperatures were shown in that table, maximum relative and minimum relative humidity, evapotranspiration and rainfall. The average for each variables is also calculated.

Stats	T_x	T_n	RHx	RHn	$U2$	$Rain$	ETo	Rs
Mean	27.5850	14.6113	88.3482	39.0408	0.6571	1.9695	3.0963	14.8225
Std	4.9678	4.3537	8.5684	18.0991	0.2788	8.9420	1.9857	6.9482
Min	13.1000	0.0000	33.3000	5.6400	0.2000	0.0000	0.1400	0.3400
25%	23.9575	11.1200	85.6075	25.8725	0.4500	0.0000	1.9600	9.9575
50%	27.5900	15.1900	91.5700	36.9950	0.6000	0.0000	2.9000	14.9300
75%	31.1125	18.1300	93.9325	50.0000	0.8000	0.2500	4.1200	19.7425
Max	44.0400	25.9300	98.4000	95.3900	2.0400	171.7000	67.4400	34.1800

Table 3.8: Statistical description for Mutale

The output or target variable plot (Global solar radiation) for this study is shown in Figure 3.6. This is a very warm area and it receives an average maximum temperature of 31.286^{o}C and an average solar radiation of $18.321MJ/m^2$

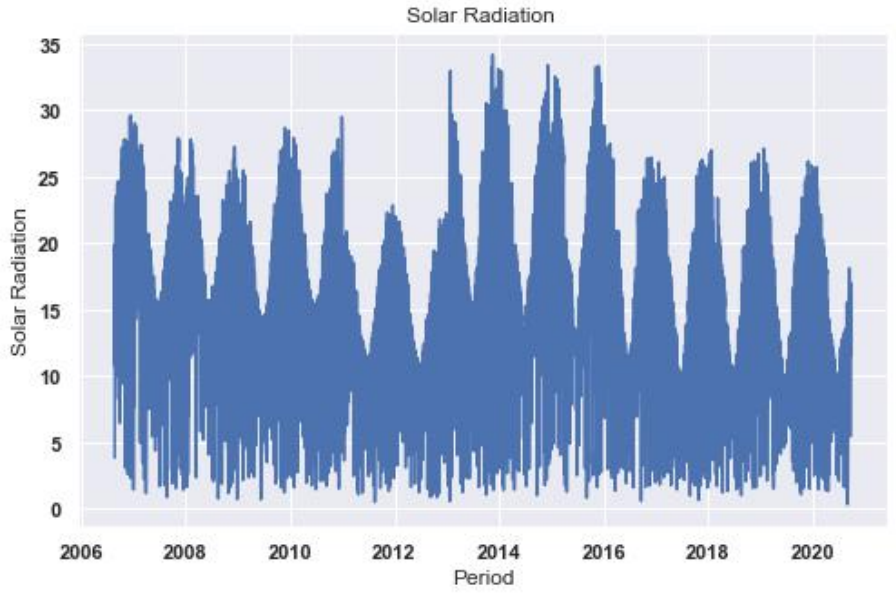


Figure 3.6: Global solar radiation observed for Mutale.

The data was processed in Python, utilising Jupyter notebook and its libraries to study the correlation between different variables. A correlation analysis was conducted in order to comprehend the linear and non-linear relationships between the meteorological prameters of the dataset and the target variable. Correlation coefficient always lies between -1 and 1. The correlation analysis was conducted to understand liner and non-linear between the meteorological parameters of the dataset and the target variable. T_x, T_n, RHx, RHn, Rs, $U2$, $Rain$, and ETo for each station were all examined.

Figure 6.1 shows the correlation analysis relationship for Mutale area between different input variables and the target variable. The strongest positive correlation comes from the total relative evapotranspiration and solar radiation. As evapotranspiration increases, so does solar radiation. It can also be seen that there is a negative correlation between maximum temperature and solar radiation.

It is observed from Figure 6.2 that the inout variables, i.e., $U2$, RHn, RHx, and T_n and solar radiation gives weak correlation. However, from ETo it can be observed that there is a moderate positive relationship between this variable and solar radiation.

The Figure 6.3 display the relationship between the input variables and output variables for Musina area. The correlation analysis was done for each variable to see which variables have a strong correlation relationship when compared with the target variable. According to the observations there is a strong positive correlation between ETo and Rs.

Figure 6.4 demonstrate the correlation analysis for Burgersfort area and from the observa-

tion we can see that there is no strong relationship between the variables. Some variables like RHx and T_n shows no correlation.

Figure 6.5 shows the correlation analysis for Marble Hall area and the observations shows weak relationship between the input variables and also target variable. The maximum temperature shows a linear relationship between itself and other variables. There is even a moderate positive correlation from ETo

Chapter 4

Research Methods

4.1 Introduction

This study used data gathered from five distinct areas in Limpopo province. The Clemence and Hargreaves-Samani empirical models were used to analyze and estimate global solar radiation for each location. In section 4.3 the Clemence and Hargreaves-Samani models are discussed. Again, machine learning algorithms were utilized for further predictions and a full description of different machine learning techniques are also described in section 4.4. The empirical and machine learning models were compared, and the performances of empirical models and ML algorithms for estimating daily solar radiation were further evaluated using statistical equations in various areas.

4.2 Data partition

4.2.1 Training data

The training data is the most significant subset of the original data set, and it is used to train or fit the machine learning model (Yu et al., 2015). The training data is initially fed into the ML algorithms, which allows them to learn how to generate predictions for the particular task. The training data differs depending on whether supervised or unsupervised learning algorithms are used. The type of training data we supply to the model has a significant impact on the model's accuracy and prediction capabilities. The higher the quality of the training data, the better the model's performance.

In this study, the original dataset was split into 80% for training and 20% for testing. Data

splitting is done to avoid model overfitting. For example, if a machine learning model fits its training data too well, it will fail to fit additional data correctly. Depending on the type of data utilized in the studies, multiple methods of splitting, including as k-folds splitting, time series splitting, and blocking time series splitting, can be performed when separating the training data. Each model underwent hyper parameter tuning.

Figure 4.1 shows the K-Fold Cross validation split. K-Fold split is a technique for estimating the power of machine learning models that uses resampling (Saud et al., 2020). After training a model, it is not certain that the model will perform well on unseen data, hence cross validation is essential to determine its performance accuracy. Validation of the model is required. This validation is a typical data analysis process that is usually used to assess the dependability of machine learning algorithms. To evaluate the model's capability, the data set was partitioned into various k-folds. A value of K can be chosen to avoid considerable variance and significant distortion in the model. The value of k is proportional to the size of the dataset. K-Fold are stable accuracy and validate model performance. One of the problem of using this validation is that one cannot work on an imbalanced dataset. It also increases training time since it requires training the model on multiple training sets. It is expensive as it needs to be trained on multiple training set. It can be seen from the figure that the validation occures before the training set.

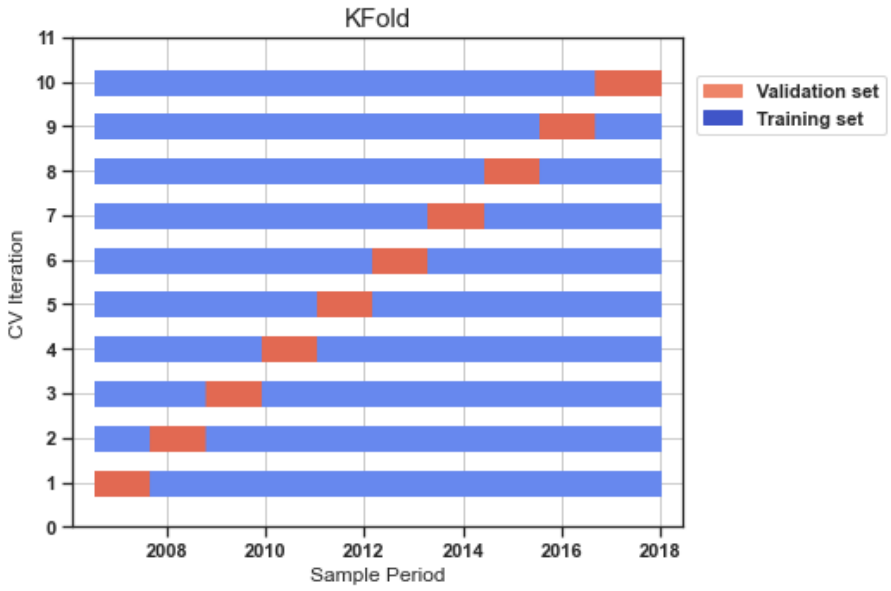


Figure 4.1: KFold Split

Blocked cross validation works by adding margins at two points. The first is between the training and validation folds to prevent the model from witnessing lag values that are utilized twice, once as an estimator (regressor) and once as a responder. The second is between

the folds utilized at each iteration to prevent the model from memorizing patterns from one iteration to the next. This validation eliminates data leakage and doesnot reuse data, that makes it to be one of the mostly used in time series data. Figure 4.2 below shows BlockingTimeSeries split. To tune the parameters, we divide the training data into ten divisions (folds). The randomised search cross-validation method was chosen because it requires less computer resources when there are fewer dimensions on the parameters to tune.

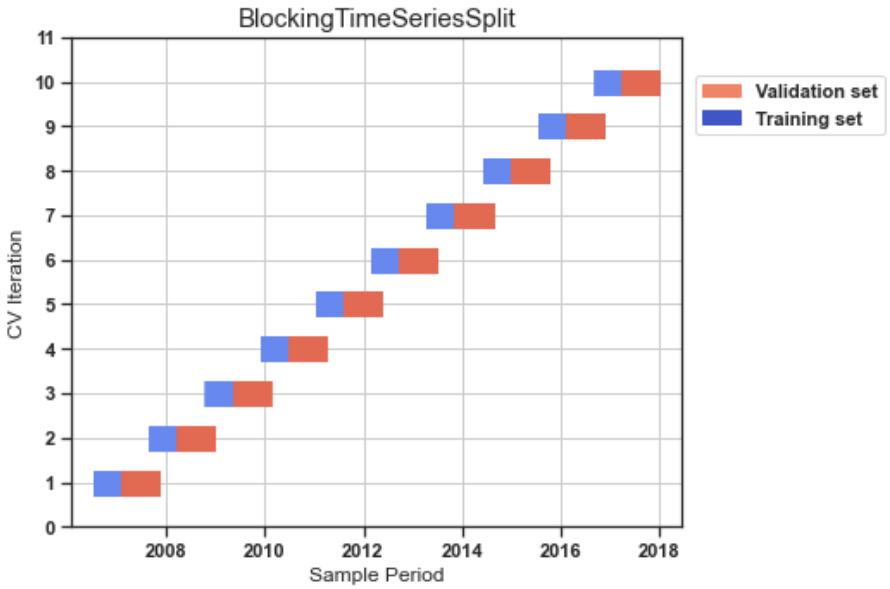


Figure 4.2: Blocked cross validation split

In Figure 4.3, the time series split algorithm divides the training set into two folds at each iteration, with the validation set always ahead of the training set. In the first iteration, the model is trained from 2006 and validated on 2008 data, and the following iteration trains on data from 2006 and validates on 2008 data, and so on until the training is completed. The advantage of using this splitting it helps understand the factors influencing specific variables in certain periods. It also helps with the historical datsets and its pattern. The limitation of this trype of validation is that it does not support the missing values on the dataset and that may lead to inaccurate results in the long term (Kang, 2013). Data transformations are compulsory and using this technique can be expensive.

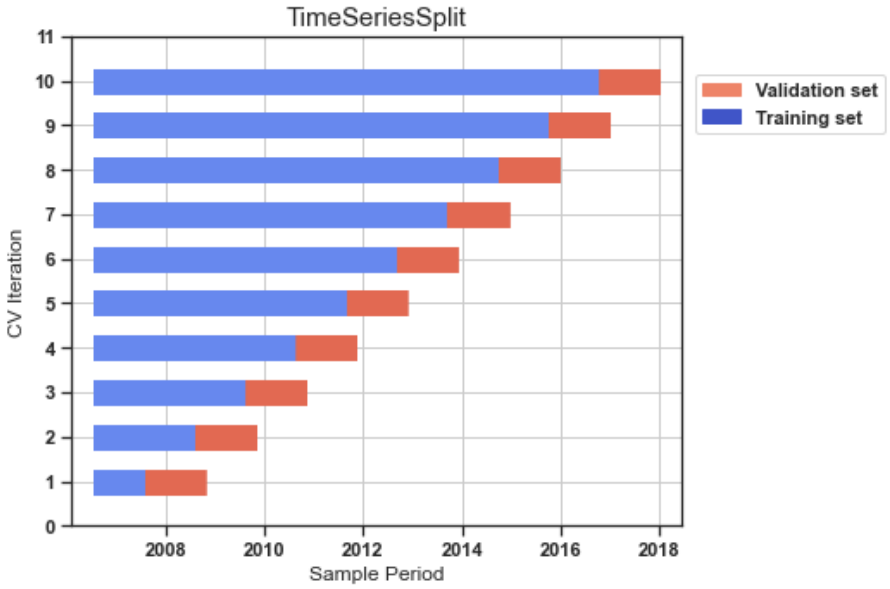


Figure 4.3: TimeSeriesSplit

4.2.2 Testing data

The remaining 20% of the original dataset is kept aside for testing. Once the machine learning model has been developed using training data, we need unseen data to test the models. This data is referred to as testing data, and it may be used to evaluate the performance and progress of the algorithms training and alter it to achieve better outcomes. Using the test data, several models were used to test the model performance using varied hyperparameter tuning for each model.

4.3 Empirical Models

Empirical models exist to evaluate global solar radiation based on meteorological and geographical characteristics such as sunshine duration, the difference between the maximum, $Tmax$, and minimum, $Tmin$, daily temperatures, and latitude (Bayrakçı et al., 2018). Two temperature-based models, that is Clemence and Hargreaves-Samani, will be used in this work to calculate the daily mean value of global solar radiation on a horizontal surface. Empirical models are built on correlations discovered through experimental data analysis(Shi et al., 2016). There is no experimental sun radiation data for all sites. As a result, numerous approaches for estimating total solar radiation have been developed. It is more precise when modelling global solar radiation at specific locations (Hassan et al., 2016).

4.3.1 Hargreaves-Samani Model

Hargreaves Samani is a temperature-based model that has recently been used to evaluate and estimate global solar radiation in various areas by calculating the difference between the daily mean value of maximum temperatures $(Tmax)$ and the daily mean value of minimum temperatures $(Tmin)$ on a horizontal surface. Hargreaves Samani proposed a simple technique for estimating global solar radiation by subtracting the maximum and minimum air temperatures from the extraterrestrial solar radiation (Hargreaves and Samani, 1982). This model is based on the relationship between the daily thermal amplitude (the difference between the maximum and minimum temperature) and the global atmospheric transmittance. Cloudiness in a specific location and season affects transmittance. Small values of maximum temperature are expected when cloud cover increases, reflecting a decrease in the intensity of global solar radiation. Minimum temperatures are affected by cloudiness because it increases the intensity of long-wave radiation emitted toward the surface at night(Groot and Carlson, 1996).

To apply the Hargreaves Samani model, an empirical coefficient of proportionality (k_r) between transmittance and the difference between maximum and minimum temperatures, which is dependent on the station's location and altitude, must be given(Thornton et al., 2000). When calculating k_r, Hargreaves only considered coastal or interior stations. Allen advised to suggested a threshold of 20km from the coast to distinguish between coastal and interior settings, hence Hargreaves offered $k_r = 0.190$ for sites within 20km of the coast. Otherwise, $k_r = 0.162$ was suggested for the inner stations. The non-linear relationship between transmittance and the difference between the maximum and minimum temperature, on the other hand, is more complicated because it is affected by factors such as humidity, proximity to water masses (such as coastal zones, regions near lakes and rivers), cloudiness, topography, and altitude(Xiang et al., 2017).

These intervening elements cause k_r to change in order to maintain an optimum balance of transmittance and thermal amplitude. Because of this, concentrating only on the station's distance from the coast may be an arbitrary norm and lead to significant errors when calculating global solar radiation under various air conditions. According to Annandale et al. (2002), the authors proposed altitude correction models, so, it is still necessary to consider whether the location is coastal or interior when calculating k_r. This explains why, in the Hargreaves-Samani model, using altitude-corrected k_r values does not always result in more accurate estimates of global solar radiation for various climatic conditions. Aside from the subjective nature of the criterion, it is critical to note the difficulties (presence of discontinuities) in simulating the geographical distribution of global solar radiation using geoprocessing

tools when the 20km threshold is used.

The model calculated the site's global solar radiation by subtracting the daily mean value of maximum temperatures $(Tmax)$ from the daily mean value of minimum temperatures $(Tmin)$ on a horizontal surface. Hargreaves and Samani proposed a simple equation to compute solar solar radiation using the temperature difference based on this idea(Hargreaves and Samani, 1982), ΔT and calculated as

$$H = H_o[K_r(T_{max} - T_{min})^{0.5}], \tag{4.3.1}$$

where, H represents daily mean value of global solar radiation ($MJm^{-2}day^{-1}$), Ho is daily mean value of extraterrestrial solar radiation mean value ($MJm^{-2}day^{-1}$), T_{max} and T_{min} are the daily mean value of maximum and minimum in (^oC), respectively and K_r is the empirical coefficient and it is given as (Maluta et al., 2014)

$$K_r = 0.00185(T_{max} - T_{min})^2 - 0.0433(T_{max} - T_{min}) + 0.04023. \tag{4.3.2}$$

These coefficients were calibrated using the data from different stations we are using in this study. Furthermore, the data used to perform the calibrations were recorded the time the equation was develped which will affect the performance of the model. Therefore, it is recommended to calibrate new coefficients for k_r using the current data and same station for the study to improve the results. The monthly average daily values of extraterrestrial irradiation (H_o) can be calculated from the following equation

$$H_o = \frac{24}{\pi} I_{sc} \left[1 + 0.033 \cos\left(\frac{360n}{365}\right)\right] \left[\cos\psi \cos\omega_s \sin\omega_s + \frac{\pi\omega_s}{180} \sin\psi \sin\delta\right], \tag{4.3.3}$$

where, I_{sc} is the solar constant $(1367Wm^{-2})$
ψ is the latitude of the site
δ is the solar declination
ω_s is the mean hour angle for the given month,
n is the number of days of a year starting from the first of January

The solar declination (δ) and the mean sunrise hour angle (ω_s) can be calculated from the following equations,

$$\delta = 23.45 \sin\left[\frac{360}{365}(n + 284)\right], \tag{4.3.4}$$

$$\omega_s = \cos^{-1}(-\tan\varphi \tan\delta). \tag{4.3.5}$$

4.3.2 Clemence Model

In this study, another type of temperature-based model was considered. Clemence developed this model to enable researchers to evaluate and estimate global solar radiation in various places by taking into account the difference between the daily mean values of maximum temperatures $(Tmax)$ and the daily mean values of minimum temperatures $(Tmin)$ on a horizontal surface. The establishment of this model used South African data, taking into account the country's various climatic conditions (Schulze et al., 1993).

Clemence temperature based equation for estimating global solar radiation is given by

$$H = a(b * H_0 * \Delta T + c * T_{max} - d * T_{max} * \Delta T + e), \tag{4.3.6}$$

where a, b, c, d and e are the parameters that were calibrated using the training dataset. ΔT is the difference between the maximum temperature (T_{max}) and the minimum temperature (T_{min}). The H_0 is given by the following equation

$$H_0 = \frac{24(60)}{\pi} * Gsc * dr[ws * \sin(\varphi) * \sin(\delta) + \cos(\varphi) * \cos(\delta) * \sin(\omega s)], \tag{4.3.7}$$

where dr is the inverse relative distance Earth-Sun, and is given by,

$$dr = 1 + 0.033\cos\left(\frac{2\pi}{365}J\right), \tag{4.3.8}$$

and,

$$\delta = 0.409\sin\left(\frac{2\pi}{365}J - 1.39\right), \tag{4.3.9}$$

where J is the number of the day in the year between 1 (1 January) and 365 0r 366 (31 December). The variable J can be determined for each day (D) of month (M) by using the following recursive algorithm(Allen and Lueck, 1998):

Algorithm 1: An algorithm for computing day number of the year

Data: Day (D), Month (M), Year (Y)

Result: Day number in the year

$J =$ **INTEGER** $\left(275 \times \frac{M}{9} - 30 + D\right) - 2$;

if $M < 3$ **then**

  $J = J + 2$;

end

if *(leap year and* $M > 2$ *)* **then**

  $J = J + 1$ ▷ This is a Leap year

end

The sunset hour angle, ωs is defined as:

$$\omega_s = cos^{-1}[-\tan(\phi)\tan(\delta)]. \tag{4.3.10}$$

The coefficient values in Clemence model were calibrated at the period the model was built using data for the weather conditions at the time the model was developed, however time has passed and the climatic conditions have also changed. This indicates that the calibrated coefficients used in the equation when the model was constructed may be invalid. To improve the efficiency of the model, the recent data in a particular study area has to be employed since this can reduce errors. The new calibrated coefficient were determined using the EXCEL solver and then compared to the initial coefficients to assess whether the parameters have been changed by climate change in recent years.

4.4 Machine Learning Algorithms

Machine learning is an artificial intelligence technology that is a branch of computer science(Rupali and Amit, 2017). Machine learning may be applied to many other fields, and the benefit of this technique is that a model can solve issues that explicit algorithms cannot(Busoniu et al., 2008). Machine learning models may uncover links between inputs and outputs even when representation is impossible, making them useful in a number of contexts such as pattern recognition, categorization, spam filtering, data mining, and forecasting. Because big data sets must be worked with and machine learning models are capable of data processing and preparation, classification and data mining are particularly appealing in this discipline. Following this level, machine learning can be used to forecast concerns.

These machine learning methods predicts future solar radiation values based on previously

observed values. These algorithms learn from databases of data. They discover patterns, gain insight, make judgments, and assess those decisions. The dataset was divided into two sections. The training data that was the greater chunk of our actual dataset that is fed into the machine learning model to uncover and learn patterns. The second dataset was test data. In this study 80 percent of the data was used for training and 20 percent was used for the test.

The models employed in global solar radiation predictions have three key applications: Time-series models that exclusively consider historically observed solar radiation data as input characteristics, structural models based on additional geographic and climatic variables, and hybrid models that treat external variables like solar radiation and other factors equally(Badescu, 2008). In the predictive learning problems, the system consist of a random output or response variable y and a set of random input or explanatory variables $x = x_1..., x_n$. Using a training sample y_i, x_i of known (x, y)- values, the goal is to obtain an estimate or approximation $f(x)$ of the function $f^*(x)$ mapping x to y, that minimizes the expected value E_{yx}, E_x and E_y of some specified loss function $L(y, f(x))$ over the joint distribution of all (y, x)-values:

$$f^* = \arg\min_f(E_{yx}(L(y, f(x)))) = \arg\min_f(E_x(E_y(L(y, f(x)))|x)) \tag{4.4.1}$$

The frequently employed loss functions $L(y, f(x))$ include squared-error $(y - f(x))^2$ and absolute error $|y - f(x)|$ for regression and negative binomial log-likelihood for classification. A common procedure is to restrict $f(x)$ to be a member of a parameterized class of functions $F(x, p)$, where $p = p_1, p_2, ...$ is a finite set of parameters whose joint values identify individual class members. Most frequently bias variance trade-off is a problem for all machine learning techniques especially supervised cases (Rashidi et al., 2019). Attempting to eliminate two sources of error at the same time causes a problem that precludes supervised learning algorithms from generalizing outside of their training set.

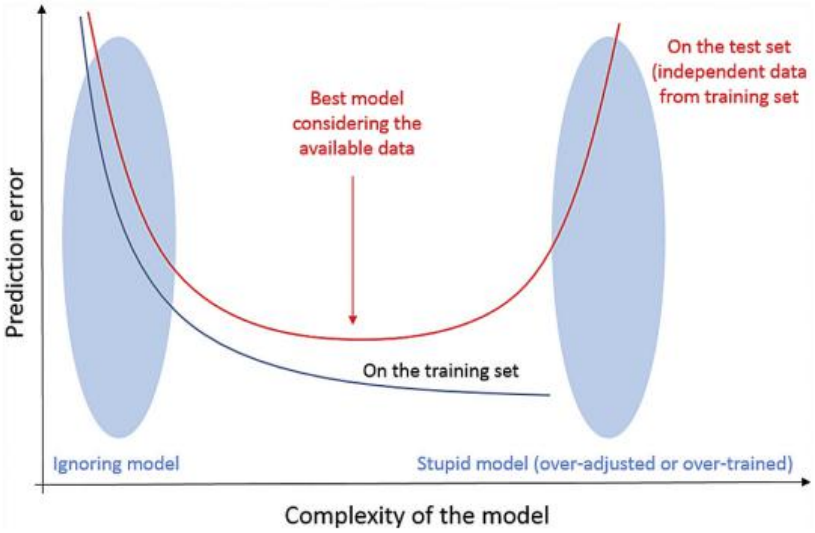


Figure 4.4: Machine learning model complexity (Voyant et al., 2017)

The bias is the inaccuracy caused by faulty assumptions in the learning algorithm. If an algorithm's bias is significant, it may become unable to build relationships between expected and actual results (under-fitting)(Jabbar and Khan, 2015). The variance is the error caused by catching minor oscillations in the training set. It should be noted that excessive variance may result in over-fitting, which models random noise in the training dataset rather than the desired output. The breakdown of the bias-variance link is used to analyze the expected generalization error of a learning approach for a certain task(Wolpert and Macready, 1999). This error is composed of three terms: bias, variance, and irreducible error, which are all induced by noise in the situation. The machine learning algorithms used in the study, particularly support vector machines, random forest, and artificial neural networks, are discussed in detail below.

4.4.1 The support vector machines

SVM learning is a type of supervised machine learning method that can be used to solve classification and regression problems(Shmilovici, 2009). The smallest portions of training data are used to determine the best support vector machine prediction model between two classes. However, because it was insufficient for multi-class estimating problems, the SVM-based SVR approach was created. Support vector regression (SVR) is a regression-based technique that calculates a linear regression function in a multidimensional feature collection(Cankurt and SUBAŞI, 2016). The main principle behind SVM is to perform non-linear data mapping in a different dot product space (referred to as the feature space) before run-

ning a linear method in the future space. Moreover, evaluating a dot product involves a several computer resources and time in the future space, it is high-dimensional. However, in some cases, a simpler kernel should be built and tested for effectiveness. Due to current linear learning machines are restricted by their computational superiority, complicated real-world situations necessitate a more expressive hypothesis space than linear functions. Target data cannot be defined as a simple linear combination of the provided attributes. One distinguishing feature, it is the ability of learning machines to be stated in a dual representation. It suggests that if the hypothesis is given as a linear combination of the training points, the decision rule can be evaluated using only the inner products between the test point and the training points. If a mechanism for directly computing the inner product in future space as a function of the original input points is available, a non-linear learning machine can be built(Twining and Taylor, 2001). This method is known as the direct calculation method of the kernel functions, which implicitly map the data to the higher dimensional feature space, and it is responsible for the flexibility of the support vector machine. In the higher dimensional feature space, a non-linear solution. As a result, SVM is a feasible solution for dealing with a wide range of naturally non-linear circumstances.

Support vector regression was offered as an optional, insensitive loss function. The goal of SVR is to discover the function that deviates the most from the actual destination vector while preserving the highest amount of flatness across all training data(Clarke et al., 2005). Non-linear support vector regression is a concept that recognizes the kernel function for SVR requires less defining kernel specific parameters. It is necessary to determine the ideal values for the legalisation argument and size errors in sensitive areas. One of the most important advantages of SVR is its algorithm, which solves the quadratic programming function and produces a unique, optimal, and thorough solution.

Support-based approaches which were first created to address classification problems, can also be successfully used to address regression issues. In contrast to classification situations where the desired outputs are discrete values: $y\epsilon[+1, -1]$, the system responses in this case are continuos values. The general regression learning problem's set is as follows:

$$D : \{(\boldsymbol{x}^{(i)}, y^{(i)}), ..., (\boldsymbol{x}^{(N)}, y^{(N)})\}; \boldsymbol{x}\epsilon\Re^n, y\epsilon\Re \tag{4.4.2}$$

where inputes $\boldsymbol{x}$ are n-dimensional vectors and scalar output y has continous values.

In support vector regression, the main objective is to find a function $\hat{y} = f(x)$ that has at most ε evaluation (when ε is a prescribed parameter) from the actually obtained targets y for

all the training data. The researcher employed an ε-sensitive loss function to account for the regression error in our support vector formulation:

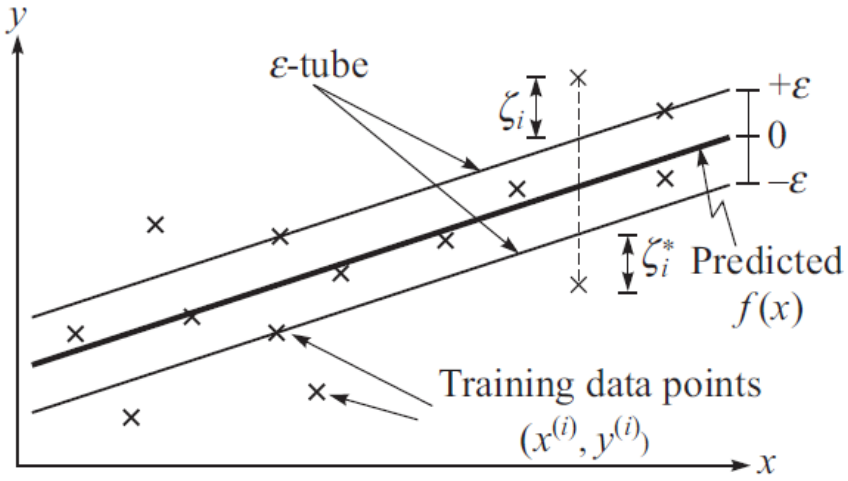


Figure 4.5: One-dimensional support vector linear regression. (Gopal, 2019)

$$\|y - f(x)\|_\epsilon \underset{=}{\Delta} \begin{cases} 0 & i \quad |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & otherwise \end{cases} \tag{4.4.3}$$

This loss(error) function defines an ε-sensitivity zone (ε-tube). Errors are tolerated up to ε (data points $(\boldsymbol{x}^{(i)}, y^{(i)})$ within the ε-insitivity zone) and errors beyond (above/below the ε-tube) have linear effect. The parameter ε determines the requirements for approximation accuracy. An increase in ε results in smoothing effects when modelling very noisy polluted data and lower accuracy requirements conversly, a decrease in ε might lead to a sophisticated model that overfits the data.

The objective of formulating a support vector algorithm for regression, is to minimize the error (loss) and $||\boldsymbol{w}||^2$ simultaneously. The role of $||\boldsymbol{w}||^2$ in the ojective function is to reduce model complexity, thereby preventing problem of overfitting, thus improving generalization.

Additionally, the support vector formulation for linear regression can be described by the function $f(\cdot)$ in the form

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0, \qquad \boldsymbol{w} \epsilon \Re^n, w_0 \epsilon \Re, \tag{4.4.4}$$

The following formulation of linear regression is comparable to the support vector formulation for the soft margin liner classifiers.

$$\text{minimize} \qquad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}(\zeta_i + \zeta_i^*),$$

$$\begin{aligned} \text{Subject to} \qquad y^i - \boldsymbol{w}^T\boldsymbol{x}^{(i)} - w_0 &\leq \varepsilon + \zeta_i; & i &= 1, ..., N \\ \boldsymbol{w}^T\boldsymbol{x}^i - w_0 - y^i &\leq \varepsilon + \zeta_i^*; & i &= 1, ..., N \\ \zeta_i, \zeta_i^* &\geq 0. & i &= 1, ..., N \end{aligned} \tag{4.4.5}$$

where C is the regularization, $||\boldsymbol{w}||$ is the weight, ζ^i and ζ^* is the errors. The constrained optimization problem in Eq.(4.4.5) can be solved by forming lagrangian equation:

$$\begin{aligned} L(\boldsymbol{w}, w_0, \zeta, \zeta^*, \lambda, \lambda^*, \mu, \mu^*) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}(\zeta_i + \zeta_i^*) - \sum_{i=1}^{N}\lambda_i(\varepsilon + \zeta_i - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0) \\ - \sum_{i=1}^{N}\lambda_i^*(\varepsilon + \zeta_i^* + y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)} - w_0) - \sum_{i=1}^{N}(\mu_i\zeta_i + \mu_i^*\zeta_i^*), \end{aligned} \tag{4.4.6}$$

where $\boldsymbol{w}$, w_0, ζ_i and ζ_i^* are the primal variables, and λ_i, λ_i^*, μ_i, $\mu_i^* \geq 0$ are the dual variables. The Karush-Kuhn-Tucker(KKT) conditions for optimality are as follows:

$$\begin{aligned} (i) \quad \frac{\partial L}{\partial \boldsymbol{w}} &= \boldsymbol{w} - \sum_{i=1}^{N}(\lambda_i - \lambda_i^*)\boldsymbol{x}^{(i)} = 0 \\ \frac{\partial L}{\partial w_0} &= \sum_{i=1}^{N}(\lambda_i^* - \lambda_i) = 0 \\ \frac{\partial L}{\partial \zeta_i} &= C - \lambda_i - \mu_i = 0, & i &= 1, ..., N \\ \frac{\partial L}{\partial \zeta_i^*} &= C - \lambda_i^* - \mu_i = 0, & i &= 1, ..., N \end{aligned} \tag{4.4.7}$$

$$\begin{aligned} (ii) \quad \varepsilon + \zeta_i - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0 &\geq 0, & i &= 1, ..., N \\ \varepsilon + \zeta_i^* + y^{(i)} - \boldsymbol{w}^T\boldsymbol{x}^{(i)} - w_0 &\geq 0, & i &= 1, ..., N \\ \zeta_i, \quad \zeta_i^* &\geq 0, & i &= 1, ..., N \end{aligned} \tag{4.4.8}$$

$$(iii) \qquad \lambda_i, \lambda_i^*, \mu_i, \mu_i^* \geq 0, \qquad i = 1, ..., N \tag{4.4.9}$$

$$
\begin{aligned}
(iv) \quad \lambda_i(\varepsilon + \zeta_i - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} - w_0) &= 0, \qquad i = 1, ...N \\
\lambda_i^*(\quad \varepsilon + \zeta_i^* - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} - w_0) &= 0, \qquad i = 1, ...N \\
\mu_i\zeta_i &= 0, \qquad i = 1, ...N \\
\mu_i^*\zeta_i^* &= 0, \qquad i = 1, ...N
\end{aligned} \tag{4.4.10}
$$

Substituting the relations in condition (i) of KKT conditions into Lagrangian Eq.(4.4.6), which yields to the dual objective function. The resulting dual optimization problem is

$$
\begin{aligned}
\text{maximize} \quad L_*(\lambda,\ \lambda^*) = &-\varepsilon\sum_{i=1}^{N}(\lambda_i + \lambda_i^*) + \sum_{i=1}^{N}(\lambda_i + \lambda_i^*)y^{(i)} \\
&- \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}(\lambda_i - \lambda_i^*)(\lambda_k - \lambda_k^*)\boldsymbol{x}^{(i)T}\boldsymbol{x}^{(k)},
\end{aligned} \tag{4.4.11}
$$

subject to,

$$
\sum_{i=1}^{N}(\lambda_i + \lambda_i^*) = 0, \qquad \lambda_i, \lambda_i^*\ \epsilon\ [0, C] \tag{4.4.12}
$$

From condition (i) of KKT conditions , we have,

$$
\boldsymbol{w} = \sum_{i=1}^{N}(\lambda_i + \lambda_i^*)\boldsymbol{x}^i \tag{4.4.13}
$$

Thus, the weight vector $\boldsymbol{w}$ is completely decribed as a linear combination of the training patterns $\boldsymbol{x}^i$. For $|\hat{y}^{(i)} - y^{(i)}| < \varepsilon$, the second factor in the following KKT conditions (iv):

$$
\lambda_i(\varepsilon + \zeta_i - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0) = \lambda_i(\varepsilon + \zeta_i - y^{(i)} + \hat{y}^{(i)}) = 0 \tag{4.4.14}
$$

$$
\lambda_i^*(\varepsilon + \zeta_i^* - y^{(i)} + \boldsymbol{w}^T\boldsymbol{x}^{(i)} + w_0) = \lambda_i^*(\varepsilon + \zeta_i^* - y^{(i)} + \hat{y}^{(i)}) = 0 \tag{4.4.15}
$$

From conditions (i) and (iv) of KKT conditions, it follows that

$$
(C - \lambda_i)\zeta_i = 0 \tag{4.4.16}
$$

$$
(C - \lambda_i^*)\zeta_i^* = 0 \tag{4.4.17}
$$

Thus, the only samples $((\boldsymbol{x})^{(i)}, y^{(i)})$ with corresponding $\lambda_i, \lambda^* \epsilon (0, C)$, we have $\zeta_i, \zeta_i^* = 0$ and the second factor of Eq.(4.4.14) and Eq.(4.4.15) has to vanish. Hence, w_0 can be computed as follows:

$$\begin{aligned} w_0 &= y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)} - \varepsilon \qquad for \;\; \lambda_i \;\; \in \;\; (0, C) \\ w_0 &= y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)} + \varepsilon \qquad for \;\; \lambda_i^* \;\; \in (0, C) \end{aligned} \tag{4.4.18}$$

All the data points with $\lambda_i, \lambda^* \in (0, C)$ may be used to compute w_0, and then their average taken as the final value for w_0.

Once this solve the quadratic optimization problem for $\lambda$and λ^* , it can be observed that all insatnces that fall in the $\varepsilon-$tube $\lambda_i = \lambda_i^* = 0$, these are the instatnces that are fitted with enough precision. The support vectors either satify $\lambda_i > 0$ or $\lambda_i^* > 0$.

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = \sum_{i \; \epsilon \; svindex} (\lambda_i - \lambda_i^*) \boldsymbol{x}^k \boldsymbol{x} + w_0 \tag{4.4.19}$$

where $svindex$ denotes the set of indices of support vectors. w_0 is defined using Eq.(4.4.20) below as

$$w_0 = \frac{1}{|svmlindex|} \left[\sum_{k \; \in \; svmlindex} \left[y^k - \varepsilon - \left(\sum_{i \; \in \; svindex} (\lambda_i - \lambda^*) \boldsymbol{x}^k \boldsymbol{x}^k \right) \right] \right] \tag{4.4.20}$$

4.4.2 The Random Forest

Breiman (2001) invented the ensemble machine learning technique of regression trees, known as random forests, which is a bagging-based approach. Random Forest is a supervised ensemble machine learning technique that is utilised to solve classification and regression problems(Ren et al., 2016). The random forests are a combination of several decision trees that were built using the bootstrapping technique, which involved selecting randomly at each node from samples in the learning dataset for the predictors. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers. This is a significant advance because decision trees produced by splitting are typically not adaptable to diverse data samples (Strobl et al., 2009). Consequently, using two sub samples of the data, one can produce two classification trees that are dramatically different for cross validation and this poses serious difficulties. A large number of decision trees are built during the training phase of ensemble learning.

The classification and regression trees model technique is used to apply the random forests concept (Chen et al., 2017). Two variations can be seen in the interim. A learning datasets is first randomly chosen for each split node in random forests, and the best split is only calculated in this phase. Secondly, there is no trimming done, therefore, all the alleged trees in the forest are maximum trees. According to a variable important measure that takes into account the impact of the input variables on the output based on prediction accuracy, random forests technique compares the prediction error with out-of-bag data term to determine the relevence of the variables. When the random forests are built during the training phase, which is also used to establish the variable importance,this out-of-bag data term is employed to obtain a running, unbiased estimate of the prediction error (Ibrahim and Khatib, 2017). The random forests algorithm is a combination of training and testing stages.

The bootstrap samples are randomly formed from the learning dataset with replacement (Kohavi et al., 1995). The number of the created samples equal to the number of the trees. There is no mathematical method to determine the ideal amount of trees to plant in a forest, therefore, Breiman (2001) proposes attempting the default (500 trees), half of the default, and twice the default before selecting the best option. Different patterns from the dataset are employed in each sample to build an independently generated random forests model. About one-third of these data are missing, which is referred to as out-of-bag data, while the remaining data called in-bag data.

At each node in a regression tree, a number of split is selected randomly to create the binary rule (Buntine and Niblett, 1992). To choose th optimal number, of splits in each tree, also known as the number of leaves per tree, the mean squared error is assessed in each split and compared with the collected out-of-bag data. The number of leaves per tree is maintained at 5, as specified by the algorithms creators, throughout the expansion of the forest.

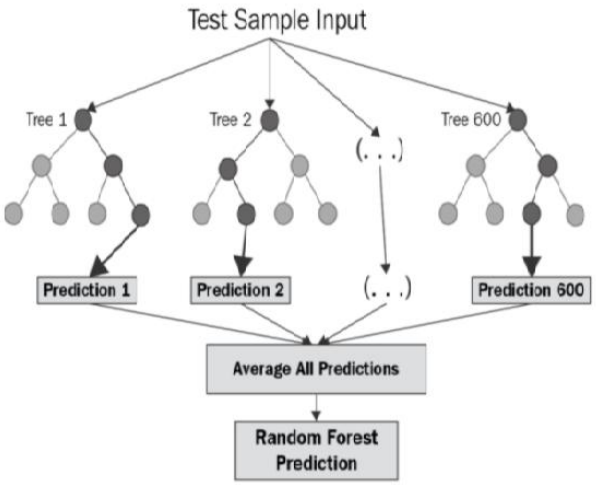


Figure 4.6: Random Forest architecture (Blakely et al., 2018)

In a random forest algorithm, there are many different decision trees. The random forest algorithm creates a forest that is trained through bagging or bootstrap aggregation (Gislason et al., 2006). The accuracy of machine learning algorithm is increased by bagging. The random forest algorithms establishes the outcome based on the prediction of the decision trees. It makes predictions by averaging or averaging out the results from different trees. Increasing the number of trees increases the precision of the outcome (Pal and Mather, 2003). It eradicates the limitations of a decision tree algorithm. It also reduces the overfitting of datasets and increases precision.

The random forest algorithm predicted value is generated using the following mathematical equation (Kor, 2021):

$$\hat{Y} = \frac{1}{N}\sum_{n=1}^{N} T_n(X) \tag{4.4.21}$$

where the average values of Y comes from n,N and $T_n(X)$. The X input parameters refers to the number of decision trees T_n; $n = 1, 2, \cdots, N$ for the input X with the aim of obtaining a robust prediction.

4.4.3 Artificial Neural Networks

Artificial neural networks are the neural networks that are based on the design of a human neuron. In the domains of artificial intelligence, machine learning, and deep learning, and neural networks enable computer programs to identify patterns and resolve common issues

by recognising the behaviour of the human brain (Tyagi and Chahal, 2022). Deep learning techniques are based on neural networks, sometimes referred to as artificial neural networks (ANNs) or simulated neural networks (SNNs), which are subset of machine learning. Their name and structure are inspired by the human brain, mirroring the communication between organic neurons. ANNs are comprised of a node layers, containing an input layer, one or more hidden layers and an output layer. Each node or artificial neuron is connected to others and has a weight and threshold that go along with it. Any node whose output exceeds the defined threshold value is activated and works uppermost layer. Otherwise, no data is transmitted to the network's next tier.

Training data is essential for neural networks to develop and enhance their accuracy over time (Han et al., 2015). However, these learning algorithms become effective tools in computer science and artificial intelligence once they are adjusted for accuracy, enabling us to quickly classify and cluster data. The purpose of the input layer is to receive as input the values of explanatory attributes for each observation. In an input layer, there are typically as many input nodes as there are explanatory variables. The network receives the patterns from the input layer and transmit them to one or more hidden layers through the network. The input layers nodes are passive, which means they do not alter the data (Nahar, 2012). They receive one value as an input and replicate it across all of their outputs. It copies each value from the input layer and sends it to every hidden node.

Figure 4.7 represent the Articial neural networks showing three layers that make up the neural network namely: Input layer, hidden layers and output layer. Let $x = \{x_1, x_2, \cdots, x_n\}$ represnt the number of input data in the network, $a = \{a_1, a_2, ..., a_m\}$ represent the number of hidden layers in the neural network and y represent the output data in the neural netwok. In Figure 4.7, the neurons are grouped into layers. The first and last Layers are called input and output layers respectively, because they represent inputs and outputs of the overall network. The remaining layers are called hidden layers.

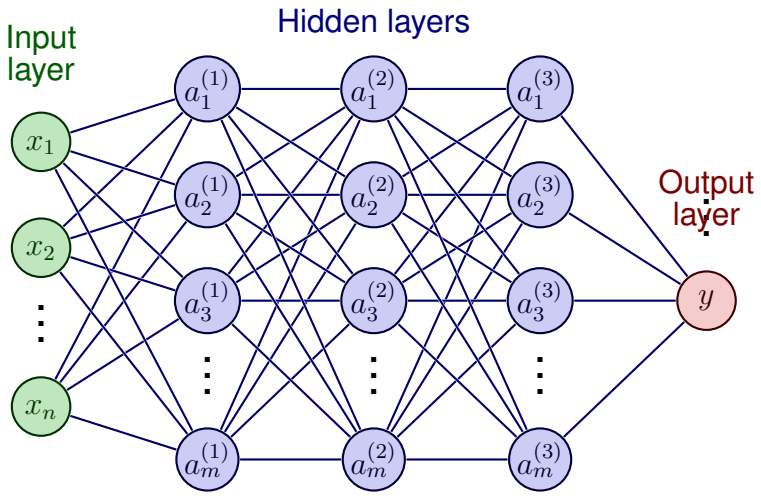


Figure 4.7: Artificial Neural Networks architecture

Suppose the total number of layers is L. The first layer is the input layer, the lth is the output layer, and layers 2 to $L-1$ are hidden layers. Let the number of neurons in the lth layer be $N_l, l = 1, 2, \cdots, L$. Let w_{ij}^l represent the weight of the link between jth neuron of $l-lth$ layer and ith neuron of the Lth layer, $1 \leq j \leq N_{L-1}$, $1 \leq j \leq N_L$. Let x_i represent the ith external input to the MLP, and z_i^l be the output of ith neuron of the lth layer. The researcher introduces an extra weight parameter for each neuron, w_{i0}^l, representing the bias for ith neuron of lth layer. As such, w of MLP includes w_{ij}^l, $j = 0, 1, \cdots, N_{l-1}$, $i = 1, 2, \cdots, N_l$, $l = 2, 3, \cdots, L$, that is,

$$w = [w_{10}^2,\ w_{11}^2,\ w_{1,2}^2 \cdots, w_{N_L,\ N_{N-L}}^L]^T. \tag{4.4.22}$$

In a neural network, each neuron with exception of neurons at the input layer receives and process inputs from other neuron (Shamseldin et al., 2002). As an example, a neuron of the lth layer receives input from the neurons of $l-l$th layer, that is, z_1^{l-1},z_2^{l-1},$\cdots$,$z_{N_{l-1}}^{l-1}$. Each input is first multiplied by the corresponding weight parameter, and the resulting products are added to produce weighted sun γ. This weighted sum is passed through a neuron activation function $\sigma(\cdot)$ to produce the final output of the neuron. This output z_i^l can, become the input for neurons in the next layer.

The most commonly-used hidden neuron activation function is the sigmoid function given by

$$\sigma(\gamma) = \frac{1}{(1+e^{-\gamma})}, \tag{4.4.23}$$

Other possible hidden neuron activation functions are the arc-tangent function and given

by

$$\sigma(\gamma) = \left(\frac{2}{\pi}\right)\arctan(\gamma), \tag{4.4.24}$$

and the hyperbolic function given by

$$\sigma(\gamma) = \frac{(e^{\gamma} - e^{-\gamma})}{(e^{\gamma} + e^{-\gamma})}. \tag{4.4.25}$$

Given the inputs $\boldsymbol{x} = [x_1,\ x_2 ... x_n]^T$ and the weights w, neural network feedforward is used to compare the outputs $\boldsymbol{y} = [y_1,\ y_2 ... x_m]^T$ from MLP neural network. In the feedforward process, the external inputs are first fed to the input neurons, the outputs from the input neurons are fed to the hidden neuronss of the 2nd layer and finally the outputs of $L-1$th layer are fed to the output neuron. The comparison is given by,

$$z_i^l = x_i, i = 1, 2, \cdots, N_1, n = N_1, \tag{4.4.26}$$

$$z_i^l = \sigma\left(\sum_{j=0}^{N_{l-1}} w_{ij}^l z^{l-1} j\right), i = 1, 2, \cdots, N_l, l = 1, 2, \cdots, L. \tag{4.4.27}$$

The outputs of the neural network are extracted from the output neurons as

$$y_i = z_i^L, i = 1, 2, \cdots, N_L, m = N_L. \tag{4.4.28}$$

The main objective in neural model development is to find an optimal set of weight parameters $\boldsymbol{w}$, such that $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x}, \boldsymbol{w})$ closely represent the original problem behaviour. This can be archieved through training process that is optimization in $\boldsymbol{w}$-space. During training, the neural network performance is evaluated by computing the difference between actual neural network outputs and desired outputs for all training samples (Sharma and Venugopalan, 2014). The difference, also known as the error is given by

$$E = \frac{1}{2}\sum_{k \in T_r}\sum_{j=1}^{m}(y_j(\boldsymbol{x}_k, \boldsymbol{w}) - d_{jk})^2, \tag{4.4.29}$$

where d_{jk} is the jth element of d_k, $y_i(x_k, w)$ is the jth neural network output for input x_k, and T_r is an index set of trining data. The weight parameter $\boldsymbol{w}$ are adjusted during training, such that this error is minimized.

During the training, $\boldsymbol{w}$ is updated along the negative direction of the gradient of E, as

$$\boldsymbol{w} = \boldsymbol{w} - \eta \frac{\partial E}{\partial w}. \tag{4.4.30}$$

The parameter η denotes the learning rate. Using one training sample at the time to update w, then a per-sample error function E_k given by

$$E_k = \frac{1}{2} \sum_{j=1}^{m} (y_j(\boldsymbol{x}_k, \boldsymbol{w}) - d_{jk})^2, \tag{4.4.31}$$

is used and $\boldsymbol{w}$ is updated as

$$\boldsymbol{w} = \boldsymbol{w} - \eta \frac{\partial E_k}{\partial w}. \tag{4.4.32}$$

In the present study, back propagation process was used to compute the gradient information $\frac{\partial E_k}{\partial w}$.

Using the definition of E_k in Eq.(4.4.30), the derivative of E_k with respect to the weight parameters of the lth layer can be computed by

$$\frac{\partial E_k}{\partial w_{ij}} = \frac{\partial E_k}{\partial z_i^l} . \frac{\partial z_i^l}{\partial w_{ij}^l}, \tag{4.4.33}$$

and,

$$\frac{\partial z_i^l}{\partial w_{ij}^l} = \frac{\partial \sigma}{\partial \gamma_i^l} . z_j^{l-1}. \tag{4.4.34}$$

The gradient $\frac{\partial E_k}{\partial z_i^l}$ can be initialized at the output layer as

$$\frac{\partial E_k}{\partial z_i^L} = (y_j(\boldsymbol{x}_k, \boldsymbol{w}) - d_{ik}), \tag{4.4.35}$$

using the error between neural network outputs and desired outputs(training data). Subsequent derivatives $\frac{\partial E_k}{\partial z_i^l}$ are computed by back-propagating this error from $l+1th$ layer to lth layer as

$$\frac{\partial E_k}{\partial z_i^l} = \sum_{j=1}^{N_{l+1}} \frac{\partial E_k}{\partial z_i^{l+1}} . \frac{\partial z_i^{l+1}}{C}. \tag{4.4.36}$$

If the multi-layer perceptron uses sigmoid as a hidden neuron activation function,

$$\frac{\partial \sigma}{\partial \gamma} = \sigma(\gamma)(1 - \sigma(\gamma)), \tag{4.4.37}$$

$$\frac{\partial z_i^l}{\partial w_{ij}^l} = z_i^l(1 - z_i^l)z_j^{l-1}, \tag{4.4.38}$$

and,

$$\frac{\partial z_i^l}{\partial z_i^{l-1}} = z_i^l(1 - z_i^l)w_i^l ij. \tag{4.4.39}$$

Let δ_i^l be defined as $\delta_i^l = \frac{\partial E_k}{\partial \gamma_i^l}$ representing local gradient at ith neuron of the lth layer. The back propagation process is then given by

$$\delta_i^L = (y_j(\boldsymbol{x}_k, \boldsymbol{w}) - d_{ik}), \tag{4.4.40}$$

$$\delta_i^l = \left(\sum_{j=1}^{N_{l+1}} \delta_j^{l+1} w_{ji}^{l+1} \right) z_i^l(1 - z_i^l), \qquad l = L-1, L-2, \cdots, 2 \tag{4.4.41}$$

and the derivatives with respect to the weigths are

$$\frac{\partial E_k}{\partial w_{ij}^l} = \delta_i^l z_j^{l-1}, \qquad l = L, L-1, \cdots, 2. \tag{4.4.42}$$

4.5 Model performance evaluation

The statistical errors will be used to evaluate and compare the accuracy and effectiveness of the empirical and machine learning techniques to estimate the global solar radiation for each area. The following statistical errors: coeffiient of determination (R^2), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE) will be employed for statistical analysis.

R^2 is used to inspect how well observed results are reproduced by the model depending on the ration of total deviation of results described by the model. This measure is represented as a value between 0.0 and 1.0, where a value of 1.0 indicates a perfect fit, and is thus a highly reliable model for future prediction, while a value of 0.0 would indicate that the model fails to accurately model the data at all. The equation for R^2 is given by

$$R^2 = 1 - \frac{\sum (x_i - y_i)^2}{\sum (x_i - \bar{x_i})^2} \tag{4.5.1}$$

MSE mearures the average of the squares of the errors between the estimated values and what is estimated. The equation for MSE is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.5.2}$$

The MAE denotes how accurate the predictions are and what is the amount of deviation from the actual values. MAE is given by the equation below

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{4.5.3}$$

$RMSE$ tells the average distance between the predicted values from the model and the actual values in the dataset. The lower the $RMSE$, the better the given model is able to fit a dataset and this mearsure range from 0 to ∞. The equation for $RMSE$ is then given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - p_i)^2} \tag{4.5.4}$$

In linear regression, correlation coefficients are used to assess the strength of a relationship between two variables. This correlation ranges between -1 and 1, where a value of 1 shows that there is a strong positive relationship, a value of -1 shows a strong negative relationship and zero shows that ther is no relationship at all. The equation for the correlation is described below

$$r = \frac{\sum(x_i - \bar{x_i})(y_i - \bar{y_i})}{\sqrt{\sum(x_i - \bar{x_i})^2 \sum(y_i - \bar{y_i})^2}} \tag{4.5.5}$$

Chapter 5

Results and discussion

5.1 Introduction

This chapter presents the findings from the empirical and machine learning techniques employed to estimate the daily global solar radiation for different stations in Limpopo province. A Python software programming was utilised for all the computations to this work, while Excel solver was also used to calibrate the empirical models when determining the model coefficients. The calibration results or coefficients for the clemence and hargreaves-samani models will be provided in Section 5.2 and also the plots for these models. The plots for all different stations for Machine learning techniques are also given in Section 5.3 respectively. A comparison of the estimated and observed global solar radiation will be provided. The performance of the models below are evaluated using the statistical error analysis.

5.2 Results for Empirical Models

5.2.1 Hargreaves-Samani model

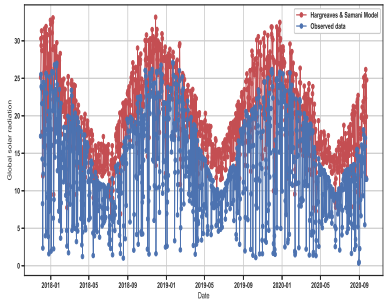
Table 5.1 below displays the original Hargreaves-Samani coefficients of the model and the calibrated results obtained from Excel solver. The parameters a, b and c are used as the coefficients for the model and their values for each station are given in Table 5.1 below respectively. These parameters were developed by the time the model was developed using the data that were collected by that time. Hence, due to climate change and other factors like global warming these parameters need to be changed using the current data. To improve the effiency of the model, these parameters will be calibrated using the current data. From Table 5.1 it can be seen that there is a difference between values for the calibrated and

original parameters and this tells us that how old is the data can affect our results due to climate conditions. The new calibrated values has dropped when compared to the original model. A scaler value of 0.95 was used to increase the efficiency of this coeffcients for each area.

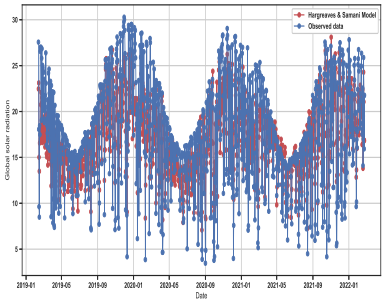
$Calibrated\ Model$	a	b	c
$Mutale$	0.0000176	-0.0004000	0.1145500
$Mussina$	0.0004623	-0.0136770	0.2323700
$MarbleHall$	0.0002760	-0.0071070	0.9572200
$Ammondale$	0.0000240	-0.0026300	0.0827500
$Burgersfort$	0.0000768	-0.0013500	0.1370000
$Original\ Model$	0.0018500	-0.0433000	0.0402300

Table 5.1: Coefficients for original Hargreaves-Samani and calibrated Hargreaves-Samani model

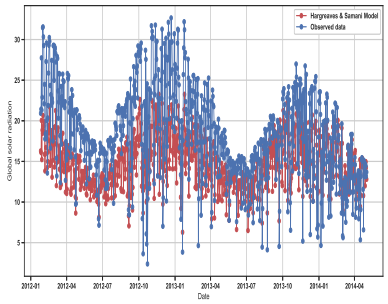
Figure 5.1 shows the estimated and observed global solar radiation for the original Hargreaves-Samani model. A python software programming were used to plot the figures for all different study area. The original parameters for this model that were developed with the model were used to estimate global solar radiation. Correlation between the estimated and observed data for global solar radiation varies with the study area. In Figure 5.1, it can be observed that between the observed and estimation for the mutale area, Hargreaves-Samani model is overestimating the observed data. The model was overestimating throught the whole year since it can be seen that it happened all the seasons. It can be seen that in Mussina area, Hargreaves-Samani model and the obsereved data are correlating very well in this area. In Ammondale and Marble Hall it can be observed that between the observed and the estimated data, the observed data is overestimating over the Hargreaves-Samani model through out all seasons of the year. In Burgersfort it can be seen that Hargreaves-Samani model was overestimating when compared to observed data during winter seasons.



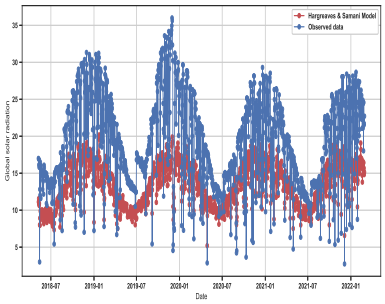
(a) Mutale



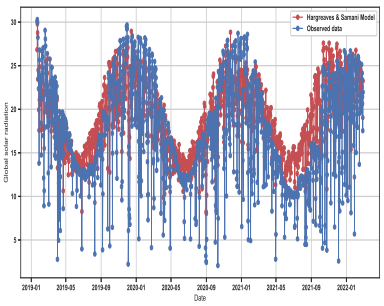
(b) Mussina



(c) Ammondale



(d) Marble Hall

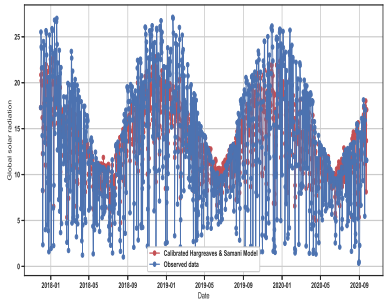


(e) Burgersfort

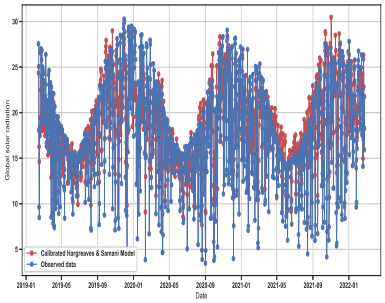
Figure 5.1: Estimation of the original Hargreaves-Samani compared with the observed data for five different study area

Below is Figure 5.2 that presents the estimated and observed global solar radiation for the calibrated Hargreaves-Samani model. An Excel solver was used to perform all the calibration process for each station in the study area. A python software programming were used to plot the figures for all different study area. Good correlation between the estimated and observed data can be seen-observed from the plots. The graphical figures show that the

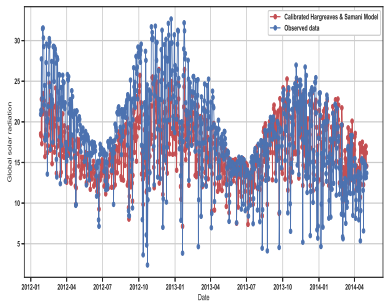
overestimation has decreased and this confirms that the calibration using current data gives out a better performing model. There is an improvement of correlation between the original model and the calibrated model of Hargreaves-Samani.



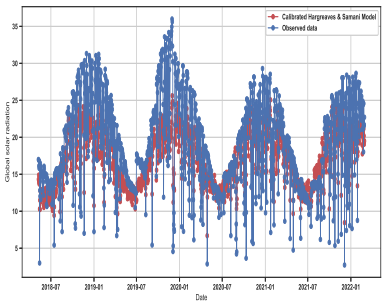
(a) Mutale



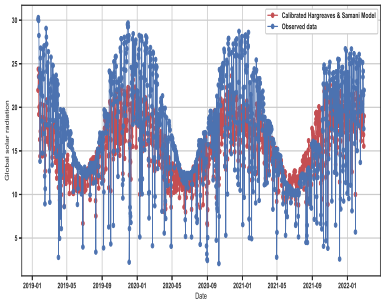
(b) Mussina



(c) Ammondale



(d) Marble Hall



(e) Burgersfort

Figure 5.2: Estimation of the calibrated Hargreaves-Samani compared with the observed data for five different study area

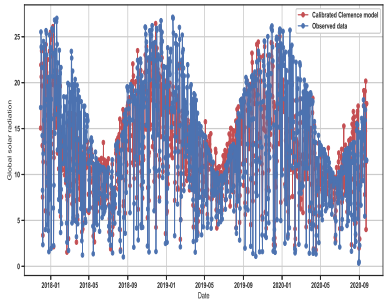
5.2.2 Clemence model

Table 5.2 provide the coefficients of the original Clemence model for all five stations. The clemence model equation contains coefficients, which were calibrated at the time the model was built using data for the weather conditions at the time the model was developed. However, time has passed, and climatic conditions have also changed. This indicates that the calibrated coefficients used in the equation when the model was created may not be reliable. To find a model that fits the area under study's current climate. The new calibrated coefficient will be computed using the Excel solver and compared to the initial coefficients to ascertain whether the parameters have been impacted by climate change in recent years and to date. The parameters a,b,c,d and e are used as the coefficient for the Clemence model and and their values for each station are given in Table 5.2 below respectively.

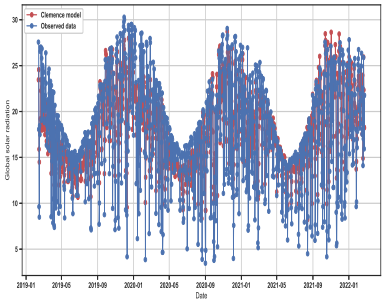
$Callibrated\ Model$	a	b	c	d	e
$Mutale$	0.01439	1.94632	0.00000	0.04989	0.00000
$Mussina$	0.03184	1.23300	8.59300	−0.70000	11.6000
$MarbleHall$	0.02794	1.50950	8.65785	−0.77000	11.9000
$Ammondale$	0.04063	1.15267	9.31147	−0.70000	11.0000
$Burgersfort$	0.02176	1.92631	8.95300	−0.88000	11.5480
$Original\ Model$	0.0318	1.2330	8.5930	−0.7130	16.5480

Table 5.2: Coefficients for original Clemence model

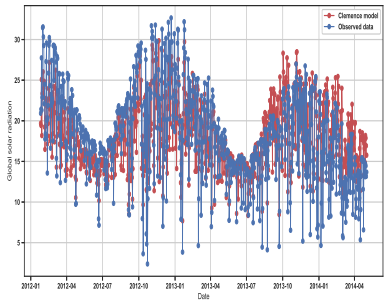
A comparison between the estimated and observed data for the Clemence model is graphically shown in Figure 5.3 below for five different study areas. A python programming was used to compute all the plots for the model. Between the estimated data and the observed data, it can be seen that the Clemence model was overestimating compared to the observed data during winter seasons for Mutale, Marble Hall and Ammondale area. But for Mussina and Ammondale area the correlation between the observed and the Clemence model is satisfying.



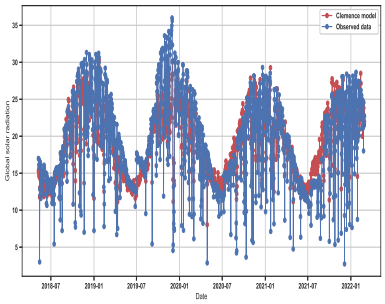
(a) Mutale



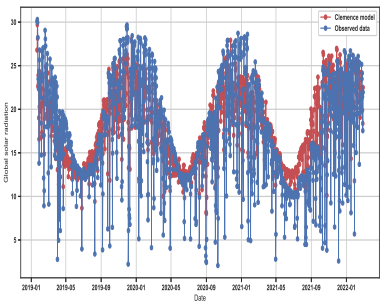
(b) Mussina



(c) Ammondale



(d) Marble Hall

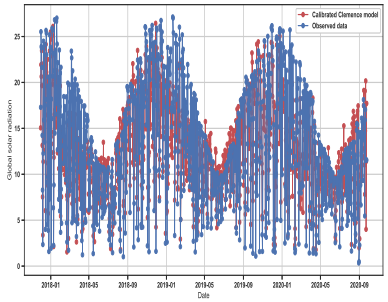


(e) Burgersfort

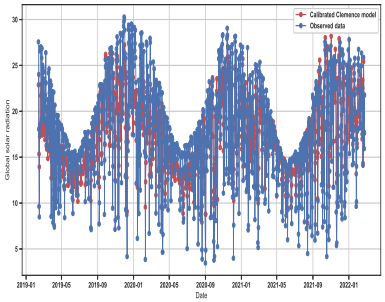
Figure 5.3: Estimation of Clemence compared with the observed data for five different study area

Below is Figure 5.4 presents the estimated and observed global solar radiation for the calibrated Clemence model. An Excel solver were used to perform all the calibration process for each station. A python software programming were used to plot the figures for all different study area. Good correlation between the estimated and observed data can be seen from the plots. The graphical figures shows that the overestimation has decreased and this con-

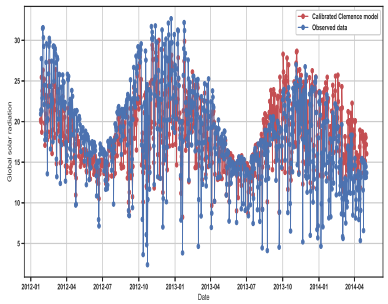
firms that the calibration with a recent data gives out a better performing model. There is an improvement of correlation between the original model and the calibrated model of the Clemence. A scaler value of 0.95 was to increase the efficiency of this coefficients.



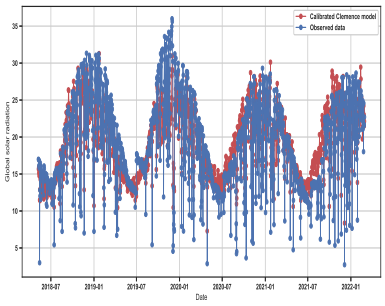
(a) Mutale



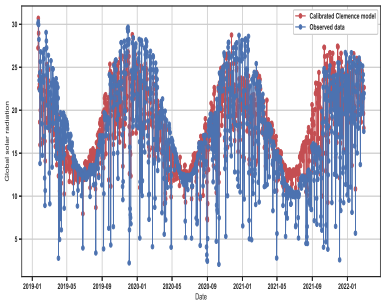
(b) Mussina



(c) Ammondale



(d) Marble Hall



(e) Burgersfort

Figure 5.4: Estimation of the callibrated Clemence compared with the observed data for five different study area

5.3 Results for Machine Learning Models

In this section, results for different machine learning techniques will be provided for each station respectively. A python software was used to perform all the graphically plots under all subsections. Subsection 5.3.1 provides results for the support vector machines and the hyper-parameter tunning for the model. Subsection 5.3.2 provides detailed results for the random forest model and its hyper-parameter tunning. A detailed explanation and graphically plots for the artificial neural networks is presented under Subsection 5.3.3.

5.3.1 Support vector machine

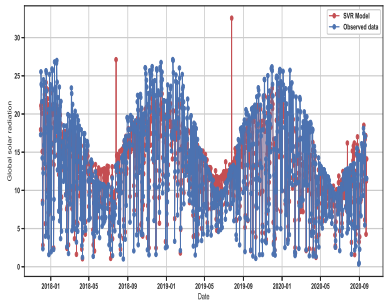
Figure 5.5 displays the estimated and observed global solar radiation plots for the support vector machines model. A python software was used to perform all the plots. Machine learning uses algorithms to learn from data in datasets. This finds patterns, develop understandings, make decision and evaluate those decisions. In machine learning, datasets are split into two subsets using blocking time series split. The first one is training data and the other subset is the testing data. Since this machine learning models are not intelligent enough to know what hyperparameters would lead to the highest possible accuracy on the given dataset, a hyperparameter tunning was done for all five stations to obtain a better machine learning models. Kernels, Regularization and Gamma are the hyperparameters used for the model. A randomized search was used to give the best parameters for the model. One of the common well known hyperparameter for the support vector machines is Kernel. With the use of randomized search, the best kernel for this model chosen was radial basis function. Table 5.3 displays the summary of best hyperparameter from the randomized serach.

$Hyperparameters$	$Mutale$	$Mussina$	$Marble\ Hall$	$Burgersfort$	$Ammondale$
$Kernel$	rbf	rbf	rbf	rbf	rbf
$Regularization$	1.00	1.00	1.00	1.00	1.00
$Gamma$	0.81	0.42	0.76	0.29	0.71

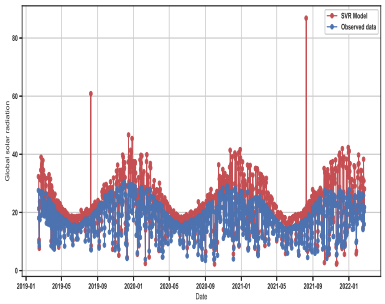
Table 5.3: Coefficients for original Clemence model

In Figure 5.5 a comparison between the observed and support vector machines model, it can be observed that in Mutale area the support vector machine was overestimating and overfitting during winter when compared with the observed data, There is again an oversestimation of the model and overfitting during winter seasons in Mussina area over the observed data. In Burgersfort the model was overestimating during winter seasons throughout over the ob-

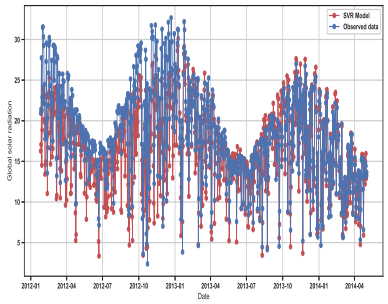
served data and there was overfitting of support vector machine.



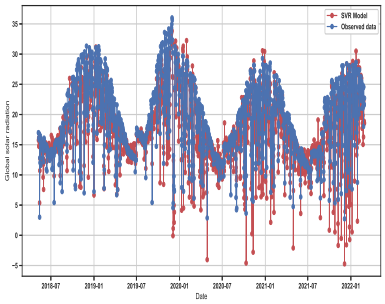
(a) Mutale



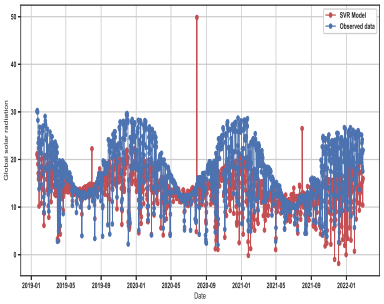
(b) Mussina



(c) Ammondale



(d) Marble Hall



(e) Burgersfort

Figure 5.5: Estimation of Support vector machines compared with the observed data for five different study area

5.3.2 Random forest

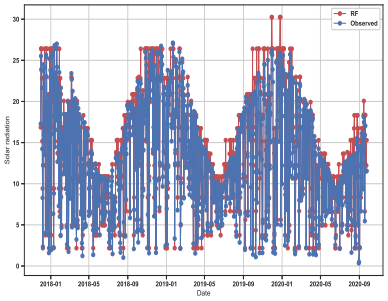
Figure 5.6 presents the estimated and observed global solar radiation plots for the random forest model. Hyperparameter tunning was done to give the best parameter and a ramdom-

ized search was used to give the best parameters for the model. Maximum number of trees, maximum depth, minimum sample split, minimum sample leaf, maximum features and bootstrap sample are the hyperparameters used for random forest model. Table 5.4 summarises the best parameters from the model and a randomized search was employed to provide those hyperparameters.

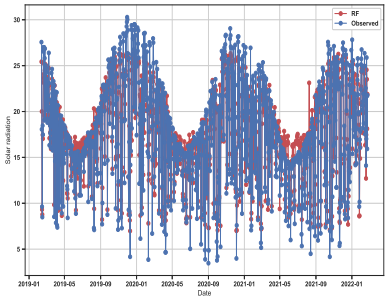
$Hyperparameters$	$Mutale$	$Mussina$	$Marble\ Hall$	$Burgersfort$	$Ammondale$
$N\_estimators$	253	466	446	253	253
$Min\_samples\_split$	32	2	2	32	32
$Min\_samples\_leaf$	22	2	2	22	22
$Max\_features$	$auto$	$sqrt$	$sqrt$	$auto$	$auto$
Max_depth	5	5	5	5	5
$Bootstrap$	$False$	$False$	$True$	$False$	$False$

Table 5.4: Hyperparemters for the random forest model

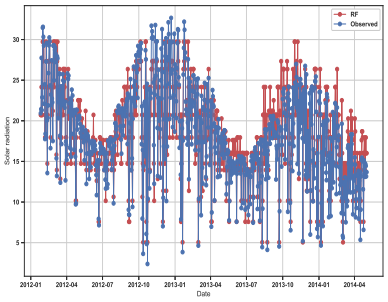
In Figure 5.6 the random forest model was compared with the observed data. It can be seen that in Mutale area and Mussina the data is showing that there is a good agreement and correlation between the obseved data and the estimated one. In Ammondale the data was responding well but there was a bit of overestimation throughout the seasons and an overfitting occured during winter season. In Marble Hall during winter seasons the random forest model data was high over the observed data. There is a good agreement between the estimated data and the observed data though there is an overfitting of a model during winter seasons.



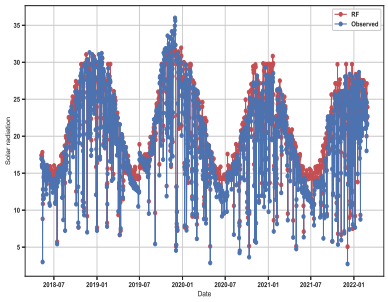
(a) Mutale



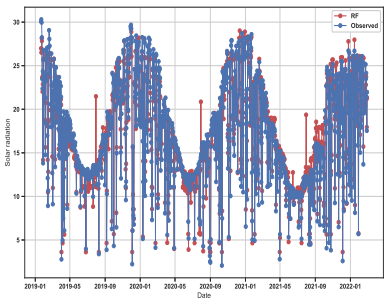
(b) Mussina



(c) Ammondale



(d) Marble Hall



(e) Burgersfort

Figure 5.6: Estimation of Random forest compared with the observed data for five different study area

5.3.3 Artificial neural network

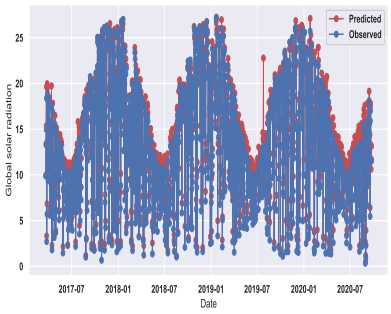
The observed and estimated global solar radiation for random forest model are graphically represented in Figure 5.7 below. A randomized search was used for the best model parameter. Number of hidden layer, activation function, optimizer, learning rate and number of

neurons are the different parameters for artficial neural network. Table 5.5 below summarizes the best hyperparameters for the model.

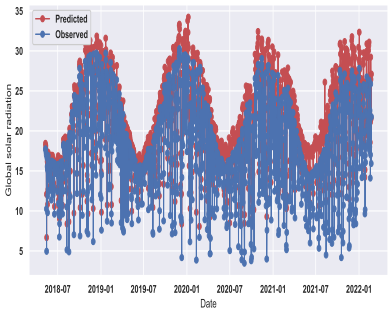
$Hyperparameters$	$Mutale$	$Mussina$	$Marble\ Hall$	$Burgersfort$	$Ammondale$
N_hidden	2	3	5	5	1
$Activation$	$selu$	$tahn$	$tahn$	$sigmoid$	$sigmoid$
$Optimization$	$RMSprop$	$RMSprop$	SGD	$RMSprop$	SGD
$Leraning\ rate$	1	0.01	0.01	1	0.01
$Neurons$	53	46	39	39	50

Table 5.5: Hyperparameter for the artificial neural network

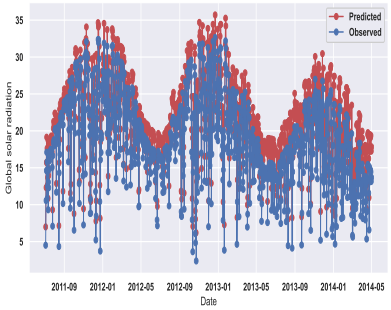
From Figure 5.7 it can be observed that in Mutale area that the observed data and the artificial neural network model are in good agreement and a good correlation. It can be seen that there is overfitting and overestimation during winter seasons of the model when compared with the observed data. In Ammondale and Marble Hall the artificial neural network and the observed data shows a good correlation and that gives an evident that the data is also corresponding.



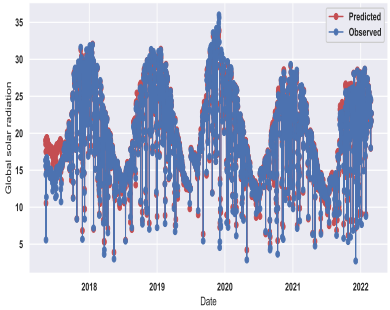
(a) Mutale



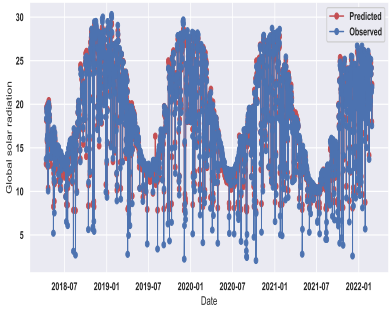
(b) Mussina



(c) Ammondale



(d) Marble Hall



(e) Burgersfort

Figure 5.7: Estimation of Artificial neural network compared with the observed data for five different study area

5.4 Comparison of the methods

A comparison between the empirical and machine learning techniques are graphicaly shown from the Figures above respectively for global solar radiation of the estimated and observed

data for all stations. The tables below show the statistically analysis for global solar radiation comparing observed and estimated data for both empirical and machine learning techniques. Table 5.6 shows the statistical analysis for the original and calibrated Hargreaves-Samani model. When comparing the original Hargreaves-Samani and the calibrated model, The statistical analysis shows and confirms that the calibration of the Hargreaves-Samani model with a recent data gives better results. The coefficient of determination for the Mutale area shows that the original Hargreaves-Samani was -0.25 and for Marble Hall was -0.27 which shows a poor fit for the data. The calibrated model shows a huge improvement of 68% for the Mutale and 57% for Marble Hall and that is a good indication of how calibration is important since the coefficient of the original model were calibrated using the data of the time the model was developed.

Original Model	MSE	R^2	$RMSE$	MAE
Mutale	52.05	-0.25	7.21	6.49
Mussina	10.81	0.67	3.29	2.64
Marble Hall	47.17	-0.27	6.87	5.83
Ammondale	27.39	0.16	5.23	4.25
Burgersfort	17.34	0.51	4.16	3.32
Calibrated Model	MSE	R^2	$RMSE$	MAE
Mutale	13.32	0.68	3.67	2.89
Mussina	10.36	0.68	3.22	2.44
Marble Hall	16.01	0.57	4.00	3.08
Ammondale	16.72	0.48	4.09	3.24
Burgersfort	14.25	0.60	3.78	3.01

Table 5.6: Statistical analysis for original Hargreaves-Samani and calibrated Hargreaves-Samani model

Table 5.7 presents the statistical analysis for the original Clemence model and calibrated Clemence model. From the summary presented in the table it can be observed that the coefficient of determination tells that the original Clemence shows that there were a good correlation between the estimated data and the observed data but it can be also taken into account that the recent data are more reliable since there is a difference of model performance.

Original Model	MSE	R^2	$RMSE$	MAE
Mutale	23.38	0.44	4.83	3.83
Mussina	10.05	0.69	3.17	2.45
Marble Hall	10.60	0.72	3.26	2.41
Ammondale	14.11	0.56	3.76	2.95
Burgersfort	13.30	0.62	3.65	2.79
Calibrated Model	MSE	R^2	$RMSE$	MAE
Mutale	10.58	0.74	3.25	2.46
Mussina	10.05	0.69	3.22	2.42
Marble Hall	10.43	0.72	3.23	2.42
Ammondale	14.04	0.57	3.75	2..93
Burgersfort	12.72	0.64	3.57	2.80

Table 5.7: Statistical analysis for original Clemence and calibrated Clemence model

To check if the model is performing well a coefficient of determination is used. From the table above it can be seen that the coefficient of determination for the original model is not satisfactory since it should be between zero and one for a better performance. Mussina data seems to show a good correlation between and relationship between the observed and the estimated data. After the calibration process the model seems to improve the performance with notable margin. The performance ranges from 49% to 69% percent which shows a better performance.

The performance metrics for the support vector machine is presented in Table 5.8 below. MSE, R^2, RMSE and MAE are the different measures used to calculate the performance of the model. The model responded well with the data and there is a good relationship between the dataset for all the staions. The coefficient of determination ranges from of 65% to $81\%1$ percent which indicate a good relationship from the observed and estimated data.

Support vector machine	MSE	R^2	$RMSE$	MAE
Mutale	7.87	0.81	2.81	2.12
Mussina	8.16	0.75	2.86	2.36
Marble Hall	8.75	0.76	2.96	2.41
Ammondale	9.26	0.71	3.04	2.17
Burgersfort	12.32	0.65	3.51	2.93

Table 5.8: Statistical analysis for support vector machines model

Table 5.9 shows the performance metrics for the random forest model. The model was responding very well in correspondence to the dataset. The correlaton between the estimated and the observed data was excellent. The coefficient of determination for random forest ranges from 0.89 to 0.97 percent which shows a good the model.

Random Forest	MSE	R^2	$RMSE$	MAE
Mutale	1.15	0.97	1.07	0.72
Mussina	1.08	0.97	1.04	0.72
Marble Hall	3.81	0.89	1.95	1.50
Ammondale	2.99	0.91	1.73	1.31
Burgersfort	1.24	0.96	1.12	0.85

Table 5.9: Statistical analysis for random forest model

In Table 5.10 a summary of statiscal calculations are shown below for artificial neural network model for each area. The model reponded very well and the coefficient of determinantion gives a perfect results for the model when compared with observed and estimated data. The model gives the best respond to the model.

Support vector machine	MSE	R^2	$RMSE$	MAE
Mutale	0.27	0.99	0.52	2.12
Mussina	0.17	0.99	0.41	2.36
Marble Hall	0.61	0.99	0.78	2.41
Ammondale	0.51	0.98	0.72	2.17
Burgersfort	0.30	0.98	0.55	2.93

Table 5.10: Statistical analysis for the artificial neural network model

5.5 Discussion

In this research project, two empirical models and three machine learning techniques were employed to predict global solar radiaton by comparing the observed and estimated data for five different stations in Limpopo. Calibrations was done to improve the deficiencies of the empirical models in estimating global solar radiation. The coefficients for the Hargreaves-Samani and Clemence models were calibrated at the time the model was developed and time has now changed and climate condition as well and that tells us that the calibrated coefficients that were used in the equation when the model was developed might not be valid. From Table 5.6 for performance evaluation it can be seen that a calibrated model gives the better performance than the original model.

Three different machine learning techniques were investigated and their comparison between the observed and estimated are graphically shown in the Figures 5.3, 5.4 and 5.5 above. It can be observed from the Figures that the global solar radiation reaches its peak and lowest points throughout the summer and winter seasons respectively. Summer is the season

with the highest levels of global solar radiation which concides with the highest temperatures. From the Figures 5.4 and 5.5 above it can be clearly seen that random forest and artificial neural networks, these two models are following the same pattern, however their temperatures are close to each other.

In comparison of machine learning techniques and empirical models, from the performance evaluation tables we can be able to conclude that machine learning techniques are the most effective method for predicting global solar radiation. Reading from the tables, the coefficient of determination for the machine learning are very close to 1 which proves the effectiveness of the models.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this study, empirical models and machine learning techniques were employed to estimate the global solar radiation for different areas in Limpopo province. Empirical models have been employed to predict global solar radiation for different locations. They use Mathematical formulas to calculate the solar radiation. These models show some deficiencies due to the used data. The coefficients that were used to generate the empirical equations were done using the data that were recorded by the time the model was developed. Due to the climate change and other factors these models may not be effective for the current use data. A calibration was done to improve the models deficiencies and model performance since the empiricals were providing poor and unsatisfactory results. Excel solver was used to perform the calibrations for each station. A hyper-parameter tunning was also taken into consideration in purpose for obtaining better machine learning and a randomized search was used to give the best model. Statistical analysis was also utilised to determine the performance measures of the models. The researcher have observed that the performance of the empirical models is low compared to machine learning techniques. Artificial neural networks shows to be the most reliable technique since the R^2 is 0.98 and 0.99. Random forest also gives out the best results since the R^2 for this model ranges from 0.89 to 0.97 and support vector machines shows to be the better technique reading that R^2 ranges from 0.65 and 0.81. If the values of R^2 is close to 1 then the model is said to be a perfect fit.

6.2 Limitation of the study

The main problem that encountered during the prediction was that the data used in this study were having numerous of missing values that causes the unsatisfactory model performance. This could be due to network miscommunication, parameters replacement or failure. Having lot of missing values from the collected data affect these model since they fail to handle and follow the pattern due to these missing values. Furthermore, they diminish predictive model accuracy and precision, resulting in erroneous estimations and inaccurate conclusions. The best way to deal with these nulls is to drop them completely which also affect the size of the data that is left for use. In machine learning techniques the more the data the better the results. Outliers were also giving alot of deficiencies for this study and needed to be taken into account. The best way of working with the outliers is to drop them off and this also affect the size of data. The recording instrument will also need frequent maintenance service or after a certain period of time need to be changed for a reliable data.

6.3 Future Work

For future studies, Empirical equations need to be calibrated using MATLAB curve fitting to improve the performance of the model. MATLAB can produce better results because is a software specifically designed for high level computing and iterative environment for algorithm development, data visualisation, data analysis and numerical computing. For the data to be more useful and accurate, explanatory data analysis need to be thoroughly done such as handling missing values and outliers since droping them completely from dataset reduces the size of data. For prospective studies, these nulls and outliers should be done by calculating mean or median. Deep learning techiques such as long short term memory (LSTM) and recurrent neural networks (RNN) can also be investigated using the same datasets. More statistical analysis such as mean absolute percentage error (MAPE), root mean square log error(RMSE) and mean absolute deviation (MAD) can be used to calculate the performnce measures and compare the results between machine learning and deep learning togeter with the empirical models to see which technique is the best or gives the reliable results. These statistical analysis mentioned above can significally improve the results because MAPE have capabilities of measuring the prediction accuracy of a forecasting method, while MAE gives idea about the variable in the data set. RMSLE incures a larger penalty for the underestimation of the actual variable than the overestimation, that is more penalty is incured when predicted value is less than the actual value which all this are huge boost to

statistical analysis used in this study.

References

G. Aad, E. Abat, J. Abdallah, A. Abdelalim, A. Abdesselam, B. Abi, M. Abolins, H. Abramowicz, E. Acerbi, B. Acharya, et al. The atlas experiment at the cern large hadron collider. *Journal of instrumentation*, 3:S08003, 2008.

Ü. Ağbulut, A. E. Gürel, and Y. Biçen. Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. *Renewable and Sustainable Energy Reviews*, 135:110114, 2021.

C. C. Aggarwal et al. Neural networks and deep learning. *Springer*, 10:978–3, 2018.

D. W. Allen and D. Lueck. The nature of the farm. *The Journal of Law and Economics*, 41 (2):343–386, 1998.

J. Almorox. Estimating global solar radiation from common meteorological data in aranjuez, spain. *Turkish journal of physics*, 35(1):53–64, 2011.

D. B. Ampratwum and A. S. Dorvlo. Estimation of solar radiation from the number of sunshine hours. *Applied energy*, 63(3):161–167, 1999.

K. Angela, S. Taddeo, and M. James. Predicting global solar radiation using an artificial neural network single-parameter model. *Advances in Artificial Neural Systems*, 2011, 2011.

A. Ångström. On the atmospheric transmission of sun radiation and on dust in the air. *Geografiska Annaler*, 11(2):156–166, 1929.

J. Annandale, N. Jovanovic, N. Benade, and R. Allen. Software for missing data error analysis of penman-monteith reference evapotranspiration. *Irrigation science*, 21(2):57–67, 2002.

I. Antonopoulos, V. Robu, B. Couraud, D. Kirli, S. Norbu, A. Kiprakis, D. Flynn, S. Elizondo-Gonzalez, and S. Wattam. Artificial intelligence and machine learning approaches to

energy demand-side response: A systematic review. *Renewable and Sustainable Energy Reviews*, 130:109899, 2020.

V. Badescu. Modeling solar radiation at the earth's surface. *SpringerVerlag, Berlin/Heidelberg*, 2008.

R. M. Balabin and E. I. Lomakina. Support vector machine regression (svr/ls-svm)—an alternative to neural networks (ann) for analytical chemistry? comparison of nonlinear methods on near infrared (nir) spectroscopy data. *Analyst*, 136(8):1703–1712, 2011.

H. C. Bayrakçı, C. Demircan, and A. Keçebaş. The development of empirical models for estimating global solar radiation on horizontal surface: A case study. *Renewable and Sustainable Energy Reviews*, 81:2771–2782, 2018.

M. Behrang, E. Assareh, A. Ghanbarzadeh, and A. Noghrehabadi. The potential of different artificial neural network (ann) techniques in daily global solar radiation modeling based on meteorological data. *Solar Energy*, 84(8):1468–1480, 2010.

L. Blakely, M. J. Reno, and R. J. Broderick. Evaluation and comparison of machine learning techniques for rapid qsts simulations. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

W. Buntine and T. Niblett. A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8(1):75–85, 1992.

L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

X. Cai, J. Magidi, L. Nhamo, and B. van Koppen. *Mapping irrigated areas in the Limpopo Province, South Africa*, volume 172. International Water Management Institute (IWMI), 2017.

S. Cankurt and A. SUBAŞI. Tourism demand modelling and forecasting using data mining techniques in multivariate time series: a case study in turkey. *Turkish Journal of Electrical Engineering and Computer Sciences*, 24(5):3388–3404, 2016.

J.-L. Chen, G.-S. Li, and S.-J. Wu. Assessing the potential of support vector machine for estimating daily solar radiation using sunshine duration. *Energy conversion and management*, 75:311–318, 2013.

W. Chen, X. Xie, J. Wang, B. Pradhan, H. Hong, D. T. Bui, Z. Duan, and J. Ma. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. *Catena*, 151:147–160, 2017.

J. Chen-Min Li. Physical chemistry of some microstructural phenomena, 1978.

S. K. Chow, E. W. Lee, and D. H. Li. Short-term prediction of photovoltaic energy generation by intelligent approach. *Energy and Buildings*, 55:660–667, 2012.

S. M. Clarke, J. H. Griebsch, and T. W. Simpson. Analysis of support vector regression for approximation of complex engineering analyses. 2005.

F. Dincer. The analysis on photovoltaic electricity generation status, potential and policies of the leading countries in solar energy. *Renewable and sustainable energy reviews*, 15(1): 713–720, 2011.

K. Djaman, L. Diop, K. Koudahe, A. Bodian, and P. Ndiaye. Evaluation of temperature-based solar radiation models and their impact on penman-monteith reference evapotranspiration in a semiarid climate. *International Journal of Hydrology*, 4(2):84–95, 2020.

J. P. Dorian, H. T. Franssen, and D. R. Simbeck. Global challenges in energy. *Energy policy*, 34(15):1984–1991, 2006.

H. K. Elminir, A. E. Ghitas, R. Hamid, F. El-Hussainy, M. Beheary, and K. M. Abdel-Moneim. Effect of dust on the transparent cover of solar collectors. *Energy conversion and management*, 47(18-19):3192–3203, 2006.

H. Ghasemi Mobtaker, Y. Ajabshirchi, S. F. Ranjbar, M. Matloobi, and M. Taki. Estimation of monthly mean daily global solar radiation in tabriz using empirical models and artificial neural networks. *Journal of Renewable Energy and Environment*, 3(3):21–30, 2016.

P. O. Gislason, J. A. Benediktsson, and J. R. Sveinsson. Random forests for land cover classification. *Pattern recognition letters*, 27(4):294–300, 2006.

F. Golestaneh, P. Pinson, and H. B. Gooi. Very short-term nonparametric probabilistic forecasting of renewable energy generation—with application to solar energy. *IEEE Transactions on Power Systems*, 31(5):3850–3863, 2016.

M. Gopal. *Applied machine learning*. McGraw-Hill Education, 2019.

A. Groot and D. W. Carlson. Influence of shelter on night temperatures, frost damage, and bud break of white spruce seedlings. *Canadian Journal of Forest Research*, 26(9):1531–1538, 1996.

S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

G. H. Hargreaves and Z. A. Samani. Estimating potential evapotranspiration. *Journal of the irrigation and Drainage Division*, 108(3):225–230, 1982.

G. E. Hassan, M. E. Youssef, M. A. Ali, Z. E. Mohamed, and A. I. Shehata. Performance assessment of different day-of-the-year-based models for estimating global solar radiation-case study: Egypt. *Journal of Atmospheric and Solar-Terrestrial Physics*, 149:69–80, 2016.

H. Herzog, B. Eliasson, and O. Kaarstad. Capturing greenhouse gases. *Scientific American*, 282(2):72–79, 2000.

I. A. Ibrahim and T. Khatib. A novel hybrid model for hourly global solar radiation prediction using random forests technique and firefly algorithm. *Energy Conversion and Management*, 138:413–425, 2017.

H. Jabbar and R. Z. Khan. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication and Instrumentation Devices*, 70, 2015.

H. Jiang, N. Lu, J. Qin, W. Tang, and L. Yao. A deep learning algorithm to estimate hourly global solar radiation from geostationary satellite data. *Renewable and Sustainable Energy Reviews*, 114:109327, 2019.

Y. Jiang. Computation of monthly mean daily global solar radiation in china using artificial neural networks and comparison with other empirical models. *Energy*, 34(9):1276–1283, 2009.

J. G. d. S. F. Junior, H. Ohtake, T. Oozeki, and K. Ogimoto. Local and regional hour-ahead forecasts of solar irradiance with training data selection and support vector regression. *IEEJ Transactions on Power and Energy*, 136(12):898–907, 2016.

K. Kaba, M. Sarıgül, M. Avcı, and H. M. Kandırmaz. Estimation of daily global solar radiation using deep learning model. *Energy*, 162:126–135, 2018.

H. Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402–406, 2013.

M. Kearns and Y. Nevmyvaka. Machine learning for market microstructure and high frequency trading. *High Frequency Trading: New Realities for Traders, Markets, and Regulators*, 2013.

I. N. Kessides. Chaos in power: Pakistan's electricity crisis. *Energy policy*, 55:271–285, 2013.

V. P. Khambalkar, S. S. Katkhede, S. Dahatonde, N. Korpe, and S. Nage. Renewable energy: an assessment of public awareness. *International Journal of Ambient Energy*, 31(3):133–142, 2010.

R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

H. Kor. Global solar radiation prediction model with random forest algorithm. *Thermal Science*, 25(Spec. issue 1):31–39, 2021.

I. Korachagaon and V. Bapat. General formula for the estimation of global solar radiation on earth's surface around the globe. *Renewable energy*, 41:394–400, 2012.

K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

R. Lindsey. Climate and earth's energy budget. *NASA Earth Observatory*, 680, 2009.

E. N. Maluta and S. T. Mulaudzi. Evaluation of the temperature based models for the estimation of global solar radiation in pretoria, gauteng province of south africa. *International Energy Journal*, 18(2), 2018.

E. N. Maluta, T. S. Mulaudzi, and V. Sankaran. Estimation of the global solar radiation on the horizontal surface from temperature data for the vhembe district in the limpopo province of south africa. *International journal of green energy*, 11(5):454–464, 2014.

R. Meenal and A. I. Selvakumar. Assessment of svm, empirical and ann based solar radiation prediction models with most influencing input parameters. *Renewable Energy*, 121: 324–343, 2018.

T. Melzer, M. Reiter, and H. Bischof. Appearance models based on kernel canonical correlation analysis. *Pattern recognition*, 36(9):1961–1971, 2003.

J. Mohtasham. Renewable energies. *Energy Procedia*, 74:1289–1297, 2015.

M. Montes, A. Abánades, J. Martínez-Val, and M. Valdés. Solar multiple optimization for a solar-only thermal power plant, using oil as heat transfer fluid in the parabolic trough collectors. *Solar energy*, 83(12):2165–2176, 2009.

K. A. Moore-O'Leary, R. R. Hernandez, D. S. Johnston, S. R. Abella, K. E. Tanner, A. C. Swanson, J. Kreitler, and J. E. Lovich. Sustainability of utility-scale solar energy–critical ecological concepts. *Frontiers in Ecology and the Environment*, 15(7):385–394, 2017.

K. Nahar. Artificial neural network. *COMPUSOFT, An international journal of advanced computer technology*, 1(2):25–27, 2012.

M. Pal and P. M. Mather. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of environment*, 86(4):554–565, 2003.

N. Panwar, S. Kaushik, and S. Kothari. Role of renewable energy sources in environmental protection: A review. *Renewable and sustainable energy reviews*, 15(3):1513–1524, 2011.

A. Pegels. Renewable energy in south africa: Potentials, barriers and options for support. *Energy policy*, 38(9):4945–4954, 2010.

F. A. PK. What is artificial intelligence? *"Success is no accident. It is hard work, perseverance, learning, studying, sacrifice and most of all, love of what you are doing or learning to do"*., page 65, 1984.

G. P. Podestá, L. Núñez, C. A. Villanueva, and M. A. Skansi. Estimating daily solar radiation in the argentine pampas. *Agricultural and forest meteorology*, 123(1-2):41–53, 2004.

N. Premalatha and A. Valan Arasu. Prediction of solar radiation for solar systems by using ann models with different back propagation algorithms. *Journal of applied research and technology*, 14(3):206–214, 2016.

V. H. Quej, J. Almorox, J. A. Arnaldo, and L. Saito. Anfis, svm and ann soft-computing techniques to estimate daily global solar radiation in a warm sub-humid environment. *Journal of Atmospheric and Solar-Terrestrial Physics*, 155:62–70, 2017.

P. Rani, C. Liu, N. Sarkar, and E. Vanman. An empirical study of machine learning techniques for affect recognition in human–robot interaction. *Pattern Analysis and Applications*, 9(1):58–69, 2006.

H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, and R. Green. Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology*, 6:2374289519873088, 2019.

D. T. Reindl, W. A. Beckman, and J. A. Duffie. Diffuse fraction correlations. *Solar energy*, 45 (1):1–7, 1990.

Y. Ren, L. Zhang, and P. N. Suganthan. Ensemble classification and regression-recent developments, applications and future directions. *IEEE Computational intelligence magazine*, 11(1):41–53, 2016.

M. Rupali and P. Amit. A review paper on general concepts of artificial intelligence and machine learning. *International Advanced Research Journal in Science, Engineering and Technology*, 4(Special Issue 4):79–82, 2017.

S. Saud, B. Jamil, Y. Upadhyay, and K. Irshad. Performance improvement of empirical models for estimation of global solar radiation in india: A k-fold cross-validation approach. *Sustainable Energy Technologies and Assessments*, 40:100768, 2020.

R. E. Schulze, G. A. Kiker, and R. P. Kunz. Global climate change and agricultural productivity in southern africa. *Global Environmental Change*, 3(4):330–349, 1993.

T. V. Segalstad. Carbon cycle modelling and the residence time of natural and anthropogenic atmospheric co2: on the construction of the" greenhouse effect global warming" dogma. In *Global Warming the Continuing Debate. Cambridge, UK. European Science and Environmental Forum*, pages 184–218, 1998.

A. Sfetsos and A. Coonick. Univariate and multivariate forecasting of hourly solar radiation with artificial intelligence techniques. *Solar Energy*, 68(2):169–178, 2000.

A. Y. Shamseldin, A. E. Nasr, and K. M. O'connor. Comparison of different forms of the multi-layer feed-forward neural network method used for river flow forecasting. *hydrology and earth system sciences*, 6(4):671–684, 2002.

B. Sharma and K. Venugopalan. Comparison of neural network training functions for hematoma classification in brain ct images. *IOSR Journal of Computer Engineering*, 16(1): 31–35, 2014.

L. Shi, G. Zhang, X. Cheng, Y. Guo, J. Wang, and M. Y. L. Chew. Developing an empirical model for roof solar chimney based on experimental data from various test rigs. *Building and Environment*, 110:115–128, 2016.

A. Shmilovici. Support vector machines. In *Data mining and knowledge discovery handbook*, pages 231–247. Springer, 2009.

A. K. Shukla, K. Sudhakar, and P. Baredar. Design, simulation and economic analysis of standalone roof top solar pv system in india. *Solar Energy*, 136:437–449, 2016.

C. Strobl, J. Malley, and G. Tutz. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods*, 14(4):323, 2009.

G. Tassey. Standardization in technology-based markets. *Research policy*, 29(4-5):587–602, 2000.

P. E. Thornton, H. Hasenauer, and M. A. White. Simultaneous estimation of daily solar radiation and humidity from observed temperature and precipitation: an application over complex terrain in austria. *Agricultural and forest meteorology*, 104(4):255–271, 2000.

A. Trabea and M. M. Shaltout. Correlation of global solar radiation with meteorological parameters over egypt. *Renewable Energy*, 21(2):297–308, 2000.

A. Troncoso, S. Salcedo-Sanz, C. Casanova-Mateo, J. Riquelme, and L. Prieto. Local models-based regression trees for very short-term wind speed prediction. *Renewable Energy*, 81:589–598, 2015.

C. J. Twining and C. J. Taylor. Kernel principal component analysis and the construction of non-linear active shape models. In *BMVC*, volume 1, pages 23–32, 2001.

A. K. Tyagi and P. Chahal. Artificial intelligence and machine learning algorithms. In *Research Anthology on Machine Learning Techniques, Methods, and Applications*, pages 421–446. IGI Global, 2022.

M. Vakili, S. R. Sabbagh-Yazdi, S. Khosrojerdi, and K. Kalhor. Evaluating the effect of particulate matter pollution on estimation of daily global solar radiation using artificial neural network modeling based on meteorological data. *Journal of cleaner production*, 141: 1275–1285, 2017.

C. Voyant, G. Notton, S. Kalogirou, M.-L. Nivet, C. Paoli, F. Motte, and A. Fouilloy. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105:569–582, 2017.

L. Wald. Solar radiation energy (fundamentals). *Solar Energy Conversion and Photoenergy Systems, edited by Julian Blanco and Sixto Malato, in Encyclopedia of Life Support Systems (EOLSS), Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK,[http://www. eolss. net]*, 1:1–10, 2007.

D. H. Wolpert and W. G. Macready. An efficient method to estimate bagging's generalization error. *Machine Learning*, 35(1):41–55, 1999.

L. T. Wong and W. K. Chow. Solar radiation model. *Applied energy*, 69(3):191–224, 2001.

J. Xiang, A. Hansen, Q. Liu, X. Liu, M. X. Tong, Y. Sun, S. Cameron, S. Hanson-Easey, G.-S. Han, C. Williams, et al. Association between dengue fever incidence and meteorological factors in guangzhou, china, 2005–2014. *Environmental research*, 153:17–26, 2017.

F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

A. Zendehboudi, M. A. Baseer, and R. Saidur. Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of cleaner production*, 199:272–285, 2018.

APPENDIX

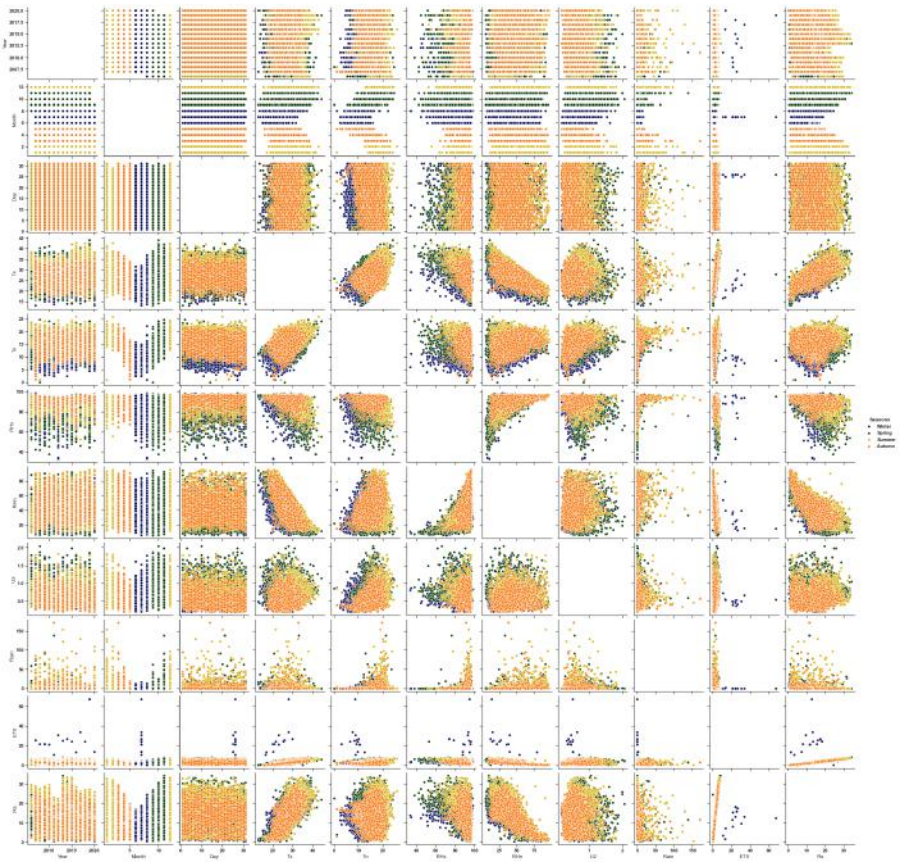


Figure 6.1: Correlation analysis variables: Mutale

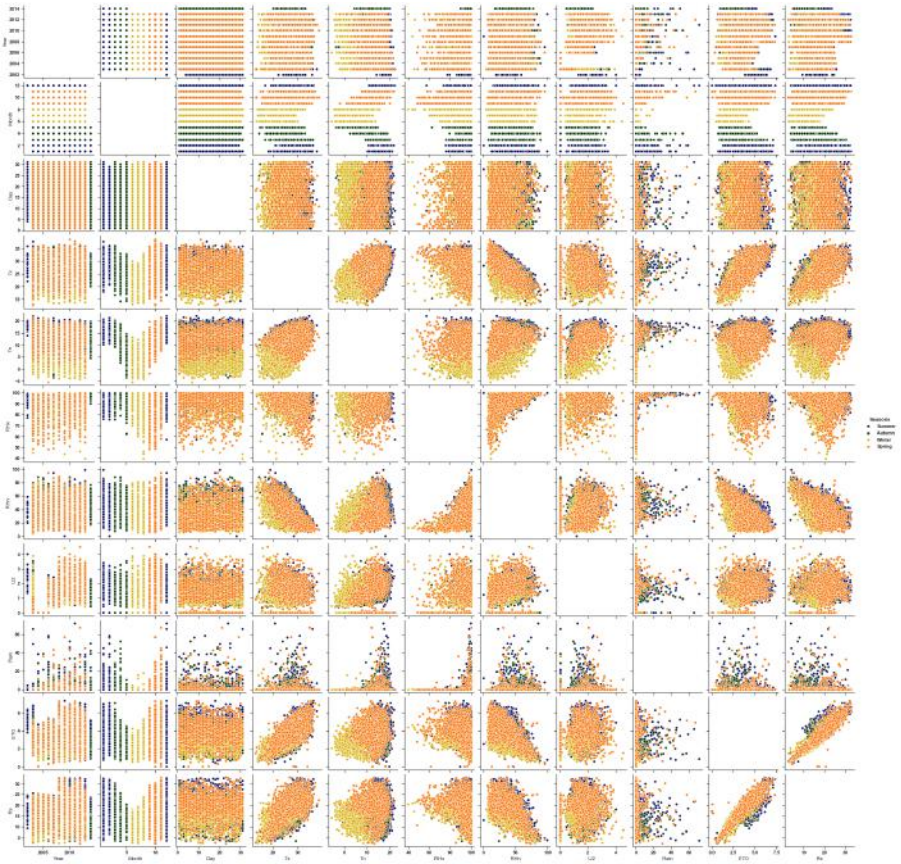


Figure 6.2: Correlation analysis variables: Musina

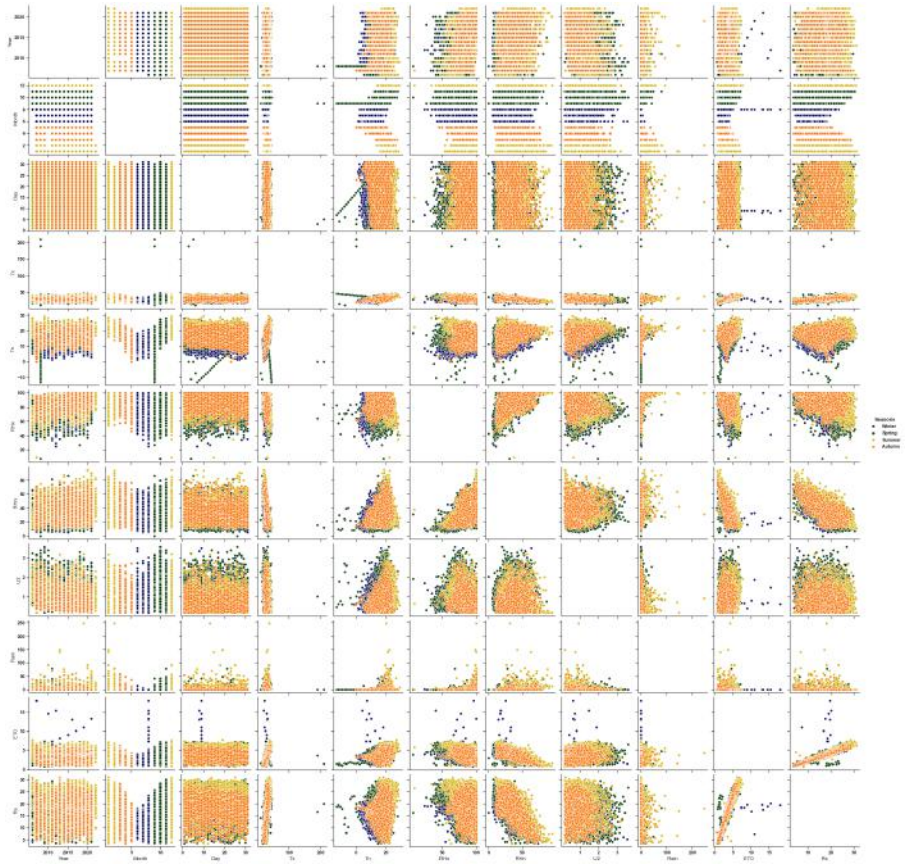


Figure 6.3: Correlation analysis variables: Marble Hall

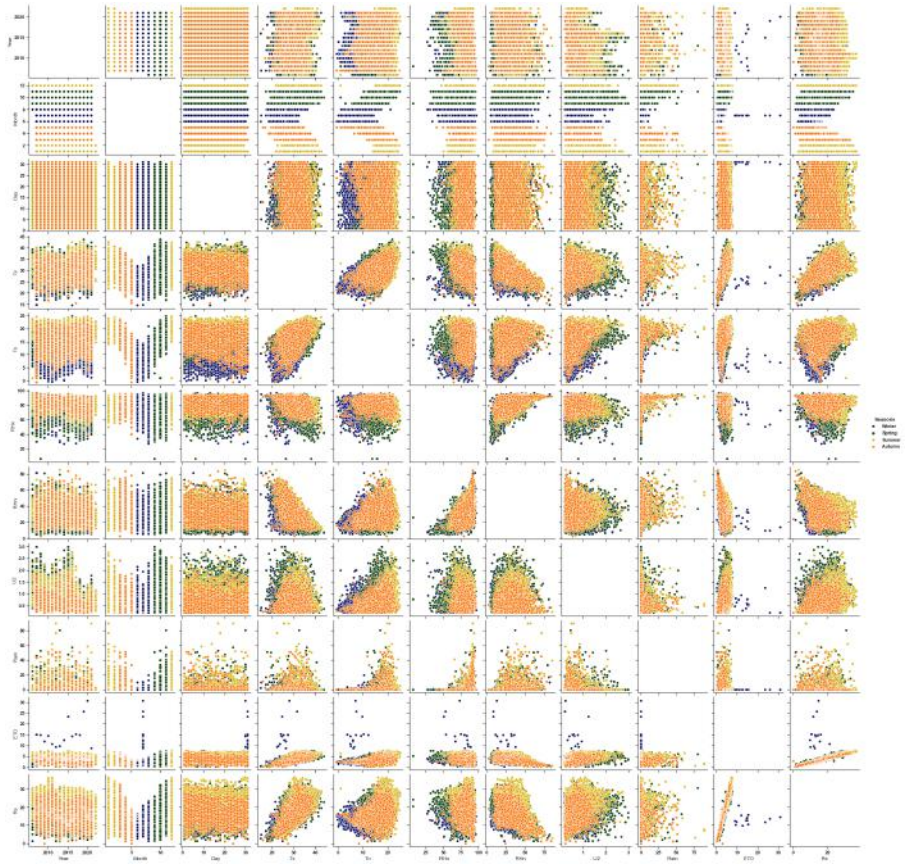


Figure 6.4: Correlation analysis variables: Burgersfort

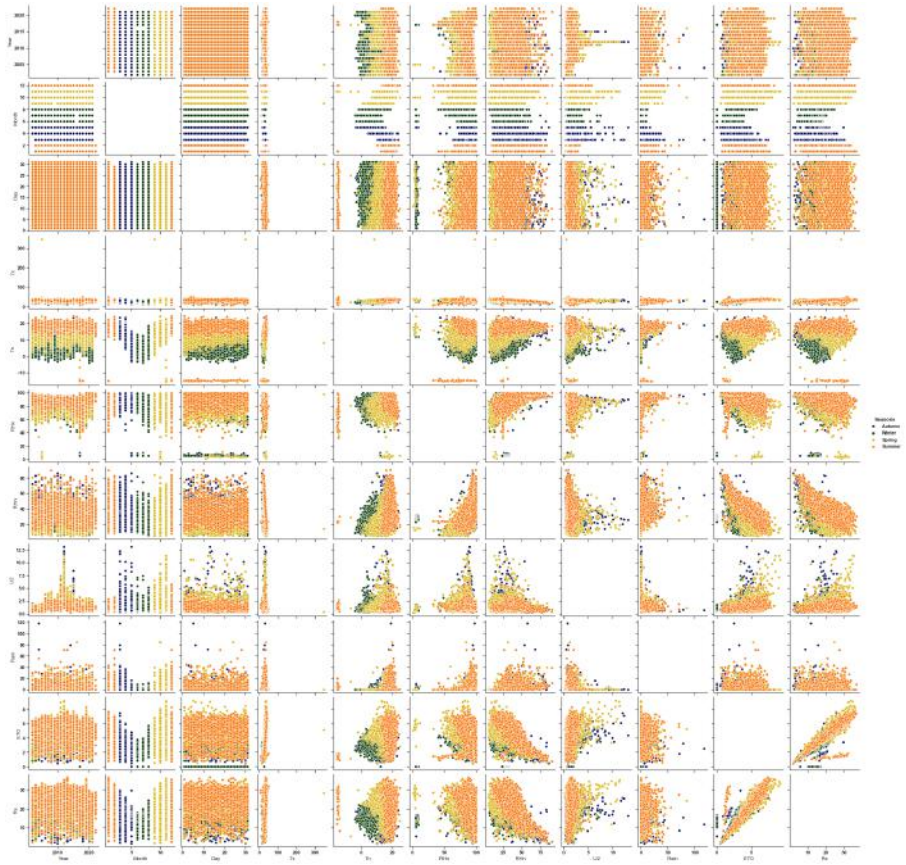


Figure 6.5: Correlation analysis variables: Ammondale