PREDICTION OF SOLAR RADIATION IN TIRANA USING MACHINE LEARNING ALGORITHMS

A THESIS SUBMITTED TO
THE FACULTY OF ARCHITECTURE AND ENGINEERING
OF
EPOKA UNIVERSITY

BY

ALBI TRASHANI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

February, 2021

Approval sheet of the Thesis

This is to certify that we have read this thesis entitled "Prediction of Solar Radiation in Tirana Using Machine Learning Algorithms" and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

_____
Dr. Ali Osman Topal
Head of Department
Date: Februrary 20, 2021

Examining Committee Members:

Dr. Ali Osman Topal (Computer Engineering) _____

…. _____

…. _____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name: Albi Trashani

Signature:

ABSTRACT

PREDICTION OF SOLAR RADIATION IN TIRANA USING MACHINE LEARNING ALGORITHMS

Trashani, Albi

M.Sc., Department of Computer Engineering

Supervisor: Dr. Ali Osman Topal

The application of machine learning in everyday life is growing and extending to every aspect of life. Not only is it becoming more accurate but also is showing great improving results in many different industries. Nowadays the machines have the ability to learn from past experiences just like humans of the real world can, and that's a relieve, they have changed the way we think about problems and they have changed the solutions by making them easier to implement. In this study we are giving a closer look to renewable energy. Also, we will check what are the meteorological factors that affect the most solar radiation. This thesis will explain in details models to predict the Solar Radiation for Tirana city, the capital of Albania. Such prediction has never been performed for Albania, neither for Tirana nor any other city for that matter. The resources available to have valuable data for this study are very few near to none for Tirana. There is no adequate equipment to measure radiation, and neither any meteorological station for such observations. Part of this paper are also econometric models and statistical considerations. The measured date, temperature, wind speed, pressure relative humidity and also solar radiation are used for measuring the accuracy of the forecasting model.

Data used are past time series of meteorological data. This thesis shows a detailed analyze of forecasting methods using statistical means were we can say that pressure(press) and temperature (Temp) have a positive relation with Solar radiation (SR) while relative humidity (RH) and wind speed (WS) have negative relation and all are statistically significant. The main algorithm that is explained in details is ARIMA (3,1,1). This algorithm performed very well with an R =0.92 and RMSE equal to 71.67(Wh/m^2). And the best performance is made by Random Forest with an R equal to 0.93 and RMSE 68.76(Wh/m^2). In addition to this and prediction of next 30 days is made.

Keywords: Machine Learning, Solar Radiation, Prediction, Accuracy, Algorithm model, Econometrics Models, Statistical Considerations.

.

ABSTRAKT

PARASHIKIMI I RADICONIT SOLAR DUKE PËRDORUR ALGORITME TË MËSIMIT AUTOMATIK

Trashani, Albi

Master Shkencor, Departamenti i Inxhinierisë Kompjuterike

Udhëheqësi: Dr. Ali Osman Topal

Aplikimi i Machine Learning është duke u bërë pjesë e shumë aspekteve të jetesës së përditshme, çdo ditë dhe më shumë. Saktësia dhe vërtetësia e të mësuarit automatik vjen së një ndryshim pozitiv në një sërë indurstrish. Ashtu si njerëzit mësojnë nga ekperiencat e tyre të kaluara po në të njëjtën formë dhe mësimi automatik ka fituar aftësinë të mësojë prej tyre, një risi kjo që ka ndryshuar mënyrën sesi ne perceptojmë zgjidhjen e problemit si dhe kanë lehtësuar implementimin e këtyre zgjidhjeve. Studimi më poshtë është i përqëndruar tek rrezatimi diellor dhe tema shpjegon në mënyrë të detajuar modelet e përdorura për të parashikuar rrezatimin diellor/orë në Tiranë, kryeqyteti i Shqipërisë. Për sa i takon Shqipërisë një studim i tillë nuk është kryer më përpara, në Tiranë ose ndonjë qytet tjetër. Burimet e nevojshme për një rezultat të vlefshëm janë shumë të pakta për qytetin e Tiranës. Deri më tani mungojnë pajisjet e duhura për të matur rrezatimin, si dhe mungon një stacion i mirëfilltë metereologjik për një studim të tillë. Në këtë punim përfshihet gjithashtu vlerësimi i modeleve

ekonometrike dhe njohuri statistikore. Numri i ditëve, temperaturës, shpejtësisë së erës, relativitetin e lagështirës si dhe sasia e rrezatimit diellor janë përdorur për të përllogaritur vërtetësinë dhe saktësinë e modelit të parashikimt. Të dhënat e përdorura janë të dhëna në seri kohore të kaluara nga databazat e meteorologjisë. Kjo temë do të prezantojë një analizë të detajuar të metodave të parashikimit me anë të mesatares statistikore ku ne mund të themi se trysnia (press) dhe temperatura kanë një lidhje pozitive me rrezatimin diellor (SR), ndërkohë lagështia relative (RH) dhe shpejtësia e erës (WS) kanë një lidhje negative por të katër këta koefiçientë kanë rëndësi statistikore. Algoritmi kryesor i shpjeguar në detaje është ARIMA (3,1,1). Ky algoritëm ka treguar performancë të lartë me një koefiçient R=0,92 dhe RMSE= 71,76 (Wh/m2). Performanca më e saktë është arritur me modelin Random Forest ku R=0,93 dhe RMSE 68,76 (Wh/m2) . Në vijimësi të këtyre vlerësimeve gjithashtu kam performuar një parashikim për 30 ditët e ardhshme të rrezatimit diellor.

Fjalët kyçe: Mësimi Automatik, Rrezatim Diellor, Parashikim, Skatesi, Model Ekonometrik, Njohuri statistikore.

Dedicated to Albanian Republic

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

SR	Solar Radiation
ML	Machine Learning
Temp	Temperature
RH	Relative humidity
Pres	Pressure
WS	Wind Speed
WD	Wind Direction
RF	Rainfall
SF	Snowfall
SD	Snow depth
Std	Standard deviation

Chapter 1

Introduction

The thesis is separated into four Chapters and each chapter will present the whole understanding and explain what needed related to the topic. This chapter there will give an overview of what is going to be covered in this work. In Chapter one the importance of solar radiation energy, research questions, objective of this study and the overall structure is discussed.

1.1 Importance of Solar Radiation Energy

One of the most mentioned topics of the recent years has been pollution. The reason behind this is the energy being produced by polluting resources like fossil fuel. This use of fossil fuel is causing many problems on our landscapes and ecosystems. From the use of these resources we are having degradation of the land, pollution of water and air, global warming also acidification of ocean.

The above-mentioned reasons ought to be more than enough to understand why we should use renewable sources of energy like wind or solar and minimize the use of fossil fuel energy. There are countless reasons why solar energy should be used, it's the best resource offered right now. For this thesis we are going to mentions some of them as follows. Solar energy is renewable, it doesn't pollute environment, doesn't create global warming, solar energy is a free source, reduces imports of energy, inexhaustible, it creates new jobs as well as a growth in economy, it is used for large scale and for small scale energy generation.

Solar radiation provides heat and light which is converted to energy. This energy which is a clean resource can be used for lighting, heating, cooling, generating electricity, for water heating and other uses of a small house to an entire city or state. The resource of this energy is enormous and according to previous researchers the amount of sunlight that hits the ground, entire earth in ninety minutes is enough for all year consumption. Also, another thing to be

mentioned is that only for full eighteen days of sunlight contains energy equal to all reserves of fossil fuel energy in our planet. Last ten years building of these systems to create and store energy has progressed rapidly. Now we just need to find the most suitable systems and places to get the most of it. In our near future we will see an enormous growth of solar energy.

Albania thanks to its location in the western part of the Balkan Peninsula also having a Mediterranean climate, results in hot dry summers, plenty days with enough sunlight. There has been estimation of the areas most exposed to sunlight, about 1500 kWh/m2 per year, variation between 1185, and 1690 kWh/m2 for the past years. However as already mentioned, there in not enough data to give exact valuation for the city of Tirana, but I can say that Tirana is part of the hot areas. The government of Albania is currently still on the process of decision making for choosing the incentive mechanism to encourage near-term investments in renewable energy. In these conditions this study for Tirana is becoming even more valuable.

1.2 Thesis objective

The task and objective of this thesis is separated in two parts. First part is to prove my model with econometric tests that my model is valid for further analysis and second part is to compare algorithms so we check which works best for this case and predict solar radiation in Tirana, the capital city of Albania while the testing are in python and E – views. Furthermore, a 30-day forecast will be made.

This thesis main research questions are:

1. What are the meteorological factors that affect the performance of solar radiation in Albania?

2. What is the statistical and causal relationship between solar radiation and these meteorological factors?

3. How can we achieve a more accurate forecast of solar radiation values in order to develop alternative methods of providing electricity?

4. Which of the machine learning algorithms performs better and how can we improve them?

1.3 Thesis Structure

Chapter 1: Chapter 1 is the overview of what the study is about. It will describe importance of solar radiation and give details of the thesis objective and structure.

Chapter 2: Chapter 2 provides the basic and general knowledge to understand this whole work. Here is going to be explained the already existing methods, the used machine learning algorithms, and the statistical ways to test a model.

Chapter 3: Chapter 3 introduces the econometric tests done to the model and the implementation of ML algorithms by using the tested model and shows their comparison.

Chapter 4: Finally, in this chapter the conclusions derived from chapter 3 will be written down and the limitations from this work and future one will be cited.

Chapter 2

Background

This chapter introduces you to the basic and general knowledge to understand this whole work. Background information on Solar Radiation (SR), Machine Learning, Econometrics model and statistical considerations, existing forecasting methods used on this thesis.

2.1 Solar Radiation

One of the main objectives of this thesis is prediction of solar radiation. So, having some important information for SR is necessary in order to understand further the implementation and the conclusion of this project. Is important to have this information for SR because it will be easier in understanding of the use of the data in order to approach the problem and the model in a correct way. Solar radiation has an essential role to life on earth. SR has a key role in determining the environment where we live since it has a direct influence on weather processes. We don't need to take a closer look at this to know how vital it is for food consumption for mankind. Also, because of the reasons that are mentioned is previous chapter it is important to know the amount of the energy released in one place.

Solar radiation, is also known by name the solar resource which is a general term for electromagnetic radiation received from the sun. SR is taken and transformed in different form of energy which we need like electricity or heat. However, every location on earth receive different amount of energy during the year and this amount of SR varies from: location, time, season, and also weather. Depending on the location the sun hits the ground differently since earth is round.

It can be 0 to 90 degrees. The most amount of energy gathered is when the sun is vertical to earth location. There are places that don't receive sun at all like north and south pole were the living doesn't exist. Also, there are places that receive the most energy and other that receive less.

2.2 Forecasting Methods That Already Exists

The availability of correct data for accurate solar radiation works is very important and so hard. This is because it is very expensive to gather these data and calibrate them. This happens expensive because the equipment that measure these data are very difficult to be made also very expensive. Another struggle is to calibrate them. In order to fix this issue a lot of engineers and researchers during these years have developed alternative ways to generate the data. Various methods are created to predict the data. The forecasting techniques that are developed and used are separated into four main categories: regression techniques, methods of artificial intelligence, the statistical approaches, and the satellite imagery techniques. Most widely used for forecasting of solar energy are regression techniques. However, they are not very accurate because they depend on meteorological data that are gathered before in order to predict solar radiation. The statistical methods are also very used because they have a linear structure which is used for forecasting purposes. But this technique also has it own pros and cons, this because it is used mostly for large data. The satellite imagery techniques are used when we have data for the location that we want to predict. Artificial intelligence methods have been and still are the most fitting techniques used for solar radiation prediction. This is because they are very capable to perform challenging tasks in a very efficient and accurate way. One thing that is already known is that each category has its own advantages and difficulties.

The regression models are grouped into four sub models: A. Sunshine Hour Based Regression Models; B. Temperature Based Regression Models, C. Multi Regression Models, and D. Cloud Cover Based Regression Models. However, the most widely and commonly used models for solar radiation prediction are sunshine hours, cloud cover, and temperature. [1]

The artificial intelligence models are grouped into two main categories: The Artificial Neural Network (ANN) models and other Artificial Intelligence (AI) models. There are many different methods developed like Particle Swam Optimization (PSO), Artificial Neural Network (ANN), adaptive neuro-fuzzy inference system (ANFIS), fuzzy logic (FL) methods, support vector machine (SVM), neuro-fuzzy network, and evolutionary optimization algorithms. [2] From the different of AI models the most widely used for the prediction of solar energy is artificial neural network method.

The statistical models are grouped mostly in time series (TS) models and other models. TS methods include: Persistence method, Auto-Regressive Moving Averages (ARMA), ARIMA, AMRAX, k-Nearest Neighbors, Markov chains and Bayesian inference. [3] [4] Other methods we can bring up are: Sunshine number [5], Clearness index [6], Exponential smoothing state space model (ESSS) [7] and other methods.

The satellite imagery techniques are another form of forecasting technique being used to predict solar radiation, which are increasing during the recent years even though equipment and experts to use them are low. Some of the models that we can mention are: Multidimensional (M-D), Solar Irradiation from Cloud Image Classification (SICIC), Multifunctional Transport Satellite (MTSAT) and other models.

2.2.1 Statistical tests

Machine learning quite often needs to borrow statistical functions/model for support and functionality. In this case we have used statistical methods to test our data set before implementing the algorithm that will run our prediction. But why do we need statistical test? And how can we use them? The tests are very useful to examine if the outputted variable has any significant relationship with the predicted variable, and to estimate differentiations between groups. First of we need to choose the correct and adequate test according to our dataset. We should know size and complexity of it so the test used can be the most efficient possible. We can choose a parametric test, regression, comparison, or correlation. The non-parametric test are not very strict about the type of data needed to be used therefore they are not much preferred as the results might not be as accurate as from a parametric test. The regression test will search for the effect of one continuous variable on other variables. They are called test-cause and effect relationship tests. Comparison tests will find the differences between the group mean. Correlation test will identify if there is any relation between the variables or not.

2.3 Machine Learning Overview

Considering that artificial intelligence is the broad science of mimicking the human abilities to perform and execute precise tasks, machine learning is a subset of artificial intelligence which trains machines how to learn. These machines look for patterns and data so that they can try to draw human like conclusions, a method behind how machines can learn from data. We can simply define machine learning as any technology that uses algorithms to create some repeatable results, where we provide many examples that will specifically indicate them what to do. With given input we produce the desired output. The main goal of these algorithms is to discover the pattern of the data and then use it to predict, answer, detect and analyze complex problems. [8]. These methods are being used in many industries which have very large amount of data needed to be processed. By deriving insights of this data even in real time, organizations and companies can work more efficiently and gain advantage over competitors beforehand.

Machine learning algorithms are used in various sectors such as financial institutions, government agencies, health care institutions, retail, oil and gas, transportation. [9] For instance:

- In Financial institutions – Identifies insights in data & prevents fraud.
- In Government Agencies – Analyzes sensitive data, increases efficiency, saves money, minimizes identity theft.
- In health care institutions – Indicates trends or red flags to improve patients diagnoses.
- In oil and gas industry – Produce realistic and useful predictions, improves the monitorization of renewable energy.
- In Transportation – Making routes more efficient and predicts possible problems to increase profitability.
- In energy – Making accurate predictions on demand

It is crucial to emphasize that machine learning is the process that powers the services we use every day. Statistics are used to find the pattern in massive data. Advertising recommendations you see non-stop on platforms like Facebook, YouTube, search engines of google etc. Frankly this process is quite basic; find the pattern, apply the pattern. [10] The output depends solely on the extensivity and accuracy of the dataset. The machines need granular data as much

detailed as possible, extremely diverse data, and very large volumes of data. Challenges and requirements for machine learning systems. [11]

- Quantitative targets/labels
- Explainability
- Automation and iterative processes
- Scalability
- Algorithms – basic and advances
- Ensemble modeling

How can computers solve problems on their own? How can they predict?

Instead of writing programs by hand to do specific tasks we collect large number of examples and give them to the machine to produce the program that will do the job. The machine learning program will contain millions of numbers and differs significantly from the hand-written program. [12]

A simplified way to describe the lifecycle of machines learnings.

Figure 1 Machine learning lifecycle

2.4 Algorithms

There are different available machine learnings methods to be used. Although it is important to know exactly how and what you want your model to be like. The more widely used methods are supervised learning, unsupervised learning, and reinforcement learning.

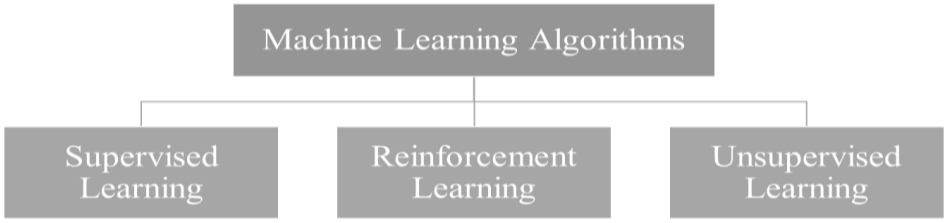

Figure 2 Machine Learning Algorithms

2.4.1 Supervised Learning

Supervised learning machines are designed to learn by example. The name itself derives from the fact that training this type of algorithm is like having a mentor supervising the whole process. The training data will consist of multiple inputs pairing with the correct output. [13]. During this process the algorithm will search for the pattern in the data related to the desired output. This is the most frequently used method; the data is tagged to indicate to the machine the precise pattern it should be looking for. After the data has been trained the supervised algorithms will take in the new unknown inputs and will determine which target these new inputs will be labeled, based on the previous inputs. The aim of this process is for the algorithm to predict the correct label for the newly introduced input data. [14]. Supervised learning is regularly used in applications where historical data predicts possible future events. The most basic form it would be where Y is the predicted output which is determined by a mapping function that assigns a new label for input X. [15].

However, there are two types of supervised learning regression, and classification. The two categories consist of an input (X) and a target output (Y). There exist a few differences between the two of them which will be explained separately.

2.4.1.1 Linear Regression

Under the category of supervised learning the linear regression model is the simplest method. This model was firstly used for statistics, so it can be confusing but as we already know that machine learning can be also recognized as the field of predictive modelling, we say that its main concerns is to provide the most accurate predictions. For this reason, machine learning will borrow and connect with other fields of study as well in this case statistics. This type consists in application of the relationship between a depended variable else called responsible variable on one or more independent variables else called explanatory variable. Linear regression is useful when the relationship between the variables is known and can be shown in the form of a straight line. It will model continuous variables and do a prediction which will have proximality to true relationships. Some examples of where we can used for this purpose are: Prediction of prices in real-estate, forecasting of stocks, forecasting of sales etc. This model

is useful to gain knowledge about data analysis processes. A disadvantage is to this is when we need to work with non-linear relationships. [16]

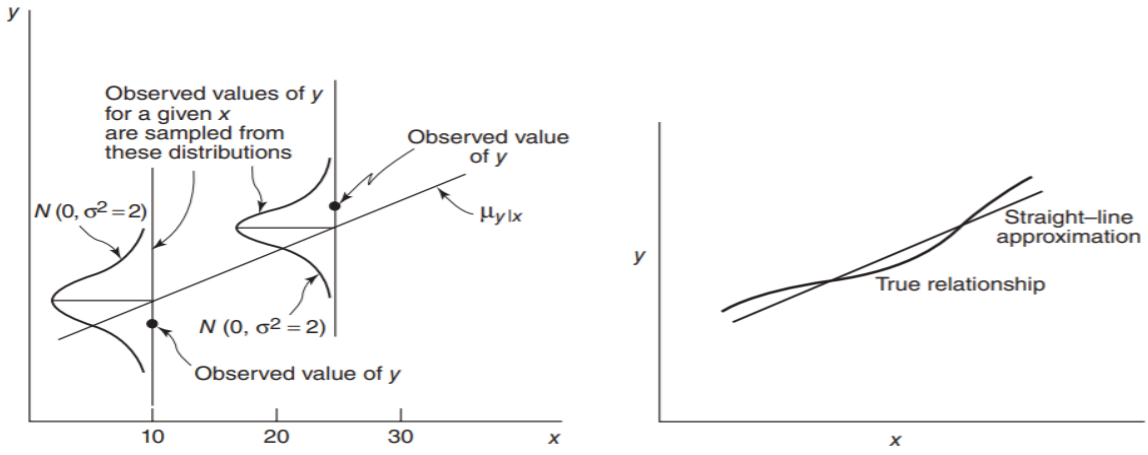

Figure 3 Generation of observations on linear regression and proximality

When working with linear regression we establish a linear prediction function where the model parameters are identified by the data.

Take into consideration the below dataset:

$$\mathcal{D} = \{(x^1, y^1), \dots, (x^l, y^l)\}, x \in \mathbb{R}^n, y \in \mathbb{R}$$

Equation 1 Data set of Linear function (Gunn, 1998)

The linear function:

$$f(x) = \langle w, x \rangle + b$$

Equation 2 Linear Function (Gunn, 1998)

In this case the optimal regression function will be given by the minimum of the function:

$$\Phi(w,\xi) = \frac{1}{2} \parallel w \parallel^2 + C\sum(\xi_i^- + \xi_i^+)$$

Equation 3 Optimal Regression Function (Gunn, 1998)

Here *C* identifies the pre-specified value, and $\xi -$, $\xi +$ the slack variables respectively upper and lower constraints of the systems output. [17].

There are a few subcategories for this type of model here mentioned:

- Simple Linear Regression

There is only one explanatory variable.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Equation 4 Simple Linear model

- Multiple Linear Regression

There are two or more explanatory variables.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

Equation 5 Multiple Linear Regression Model

2.4.1.2 Generalized Linear Models

Noted as GLM was first used to unify all linear models, it is a flexible generalization of the usual linear regression that permits the response variables to have error distributions despite the normal distribution. The linear model in this case is permitted to create a link function that relates the response variables to size of the variances so there will be a prediction for each measurement. [18]

2.4.1.3 Non-Linear Regression

Non-linear regression model is like linear model, but this type predicts depending on one or more unknown parameters, and the relationship between these parameters is not on a straight line, it's a curved relationship. Dependent and non-dependent variables here should be quantitative. Non-Linear regression model is more useful as it can help predict more accurate results than linear when we have an unknown relationship between the parameters of which we know its not a linear relationship. It can be shown in the simple form as below: [19]

$$Y_i = f(\mathbf{x}_i, \theta) + \varepsilon_i, i = 1, \ldots, n$$

Equation 6 Non-Linear Regression, basic form

Yi are the responses, f the known function of vector Xi, and εi is a random error.

2.4.1.4 Support Vector Machines/Support Vector Regressions

This learning technique is used both in classification and regression tasks. Support Vector Regression shortly noted as SVR and first introduced by Vapnik in 1986. SRV can be defined as a computer algorithm that learns by example and tags objects. For instance, if we need to identify fraudulent credit cards this algorithm will examine thousands of fraudulent and non-fraudulent activities from these credit cards and detect the fraudulent ones. By using mathematical functions in relation to given data sets this algorithm will maximize the output of the function itself. [20]. We will have a closer look to the SVR related to regression tasks. The model is largely used in regression problems related to time series forecasting. So, if we have a wide dataset in high dimensional spaces its an effective algorithm to use, its also safe to use when the number of dimensions is bigger than the number of samples. SVR uses a subset of training targets for the deciding function (called support vectors) so it is also memory efficient. [21]. The parameters in this model are derived with specific conditions. SVR works for continuous variables. There are few terms we need to be familiar with when referring to SVR:

Kernel – Function to map lower dimension data to higher dimension data

Hyperplane – Line that assists to predict the continuous variable

Boundary Line – The lined which create the margin of the datasets

Support Vectors – The data nearest to the hyperplane

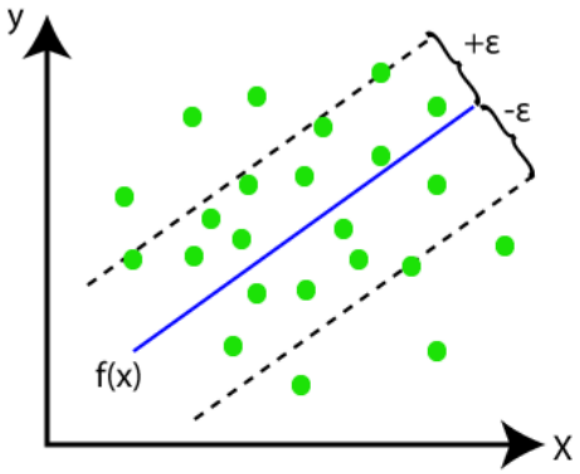

Figure 4 SVR graphical illustration of the hyperplane and closest datasets

2.4.1.5 Decision Tree Learning

This is one of the easiest models to explain and understand. The goal is always to predict, but how? In decision trees the algorithm starts from the roots and stops to each node individually, it follows the branches and jumps to the next node. For every jump there is a specific decision rule followed to locate the targeted value. For easier understanding we can explain as the following steps:

Step 1 – Root Node (contains all the dataset)

Step 2 – Locate the best attribute in the dataset

Step 3 - Separating the root into subsets that may have the best attributes

Step 4 – Generates the tree node with the best attribute

Step 5 - Go back to step 3 again and keep searching for the best attribute nodes until it cannot find one anymore.

One of the many reasons why decision trees are good to use is because of their structure to identify all possible outcomes of a problem, and it requires less data than the other algorithms. Decision trees can contain numerical data as well as categorical data (Yes/No). [22]

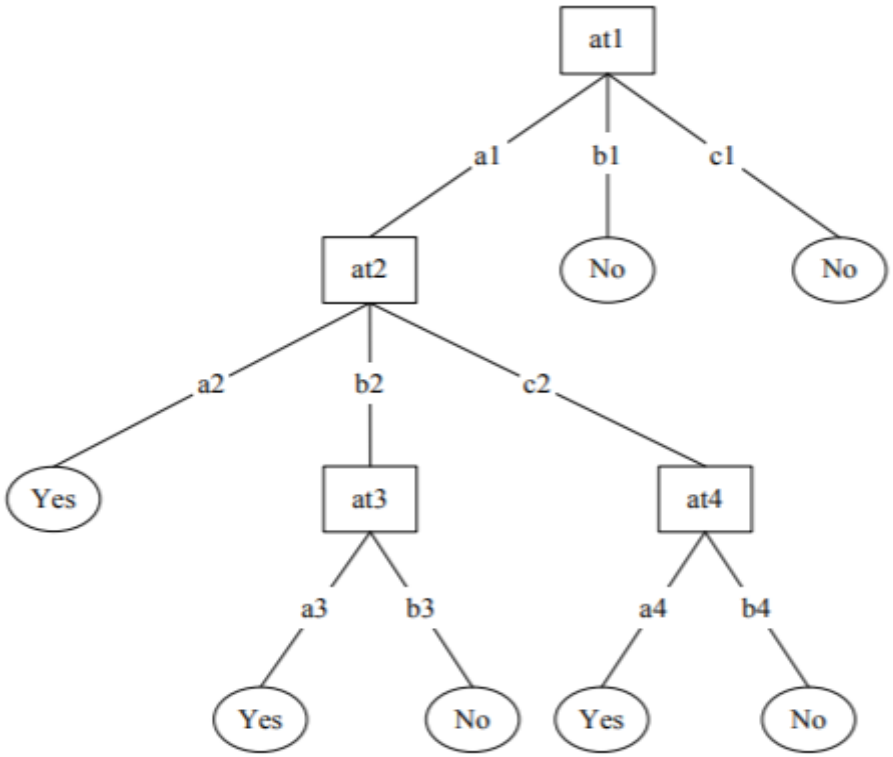

Figure 5 Example for decision tree structure

The most commonly used methods of Decision Trees is Random Forest, also for our model described in the next chapter. This model is also based on the concept of ensemble learning, shortly defined as the process of combining several classifiers to solve a more complicated problem which aims to improve the prediction and performance. It contains a few decision trees from different subsets for any know dataset and gets the average to improve the prediction from this dataset. Random forest will rely on the prediction of different trees and based on them it will give a final output. So, by simple logic the bigger the number of trees involved and the more predictions it gets the more accurate the final output will be, therefore it avoids also the problem of overfitting.

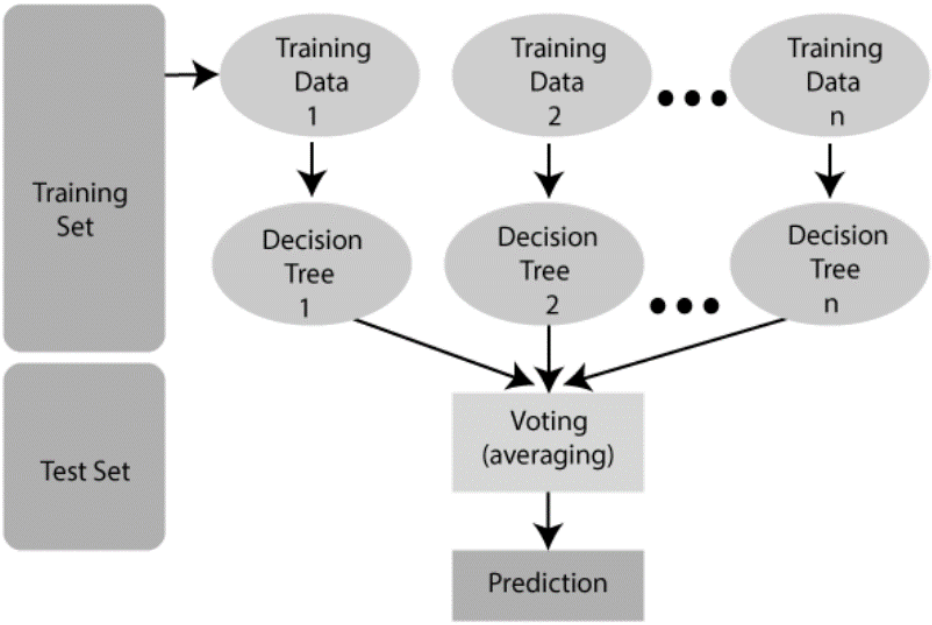

Figure 6 Diagram for Random Forest Algorithm

This algorithm takes data from the training set randomly, builds the subsets, chooses the n number of the decision trees it wants to build, repeats again until the final output. The training set will gather data and the most voted pattern from each decision tree and repeat this process until there is one patter the most voted from all trees. [23]

Random forest is now a days used in different industries like banking, health care, marketing etc. It's a very useful model for large datasets with high dimensions and has better accuracy.

2.4.1.6 K-Nearest Neighbor

K-nearest neighbor or else called nearest neighbor is one of the many algorithms which applies to both classification and regression, simple and easy to implement. This algorithm presumes that similar patterns/objects exist in proximality with each other.

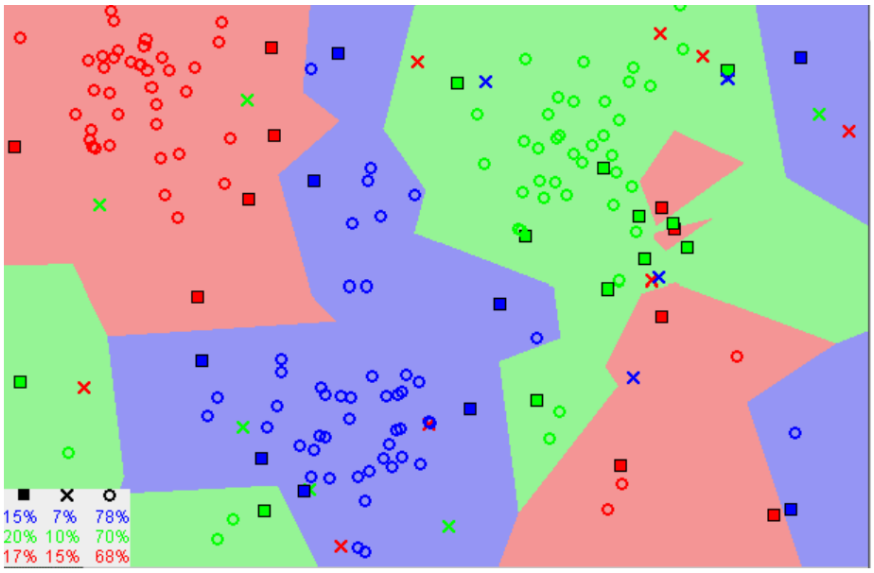

Figure 7 Similar data are position near each other

In the above photo we can clearly notice how similar data/patterns are close to each other, and this is what KNN assumes and tends to prove. Sometimes we call this measuring similarity, distance, proximity, closeness etc. K is the number of neighbors. The algorithm loads the data and measures it by calculating the distance of the current example and the query example. After putting them in ascending order according to an indexed distance it labels the data and in the case of regression, which is the model, we are using it returns the mean of K labels. However, this model is not recommended for large datasets as it gets slower the bigger the volume is. Good usage of this algorithm is when we must process data that we know beforehand to be similar. [24]

2.4.1.7 Markov Chain

Markov Chain is known to the mass in different names such as Markov Property or Markova Process named after the Russian mathematician A.A. Markov in the 80'. The essence of this model is that even if we know the present state of any system and we do not possess any further detailed data from past states of this it we can predict the future state of the same system therefore we ignore its past states. Markov process claims the past is clearly immaterial for future predictions. Markov processes are distinct in time and space. Despite that this algorithm can be widely and easily used in many areas and industries we notice a deficiency of quantity and quality when used by mathematicians or scientists. [25]

$$X = (X_n)_{n \in \mathbb{N}} = (X_0, X_1, X_2, \dots)$$

Equation 7 Markov Properties mathematically shown.

2.4.2 Unsupervised Learning

Unlike the supervised learning model this one does not need a "mentor" to interfere and supervise the whole process. This type of model can find the output without any previous labeled inputs. The purpose and aim of this learning model is to explore and find the pattern by itself. The algorithms here must find out what is being showed to them. In this case the unsupervised model does not require any historical data to be provided as the model itself discovers the pattern. It's worth mentioning that unsupervised learning has been widely ignored for about 40 years by the scientists as they found it difficult to identify its purpose. There are two main tasks within this model. First task is to identify the elementary dispersal of the input data. Secondly is to locate natural grouping/clusters in the data provided. The most commonly used algorithms are k.means clustering, hierarchal clustering, principal component analysis, auto encoders and Parzen windows. [26]

2.4.2.1 K-means

K-means is the most know and used algorithm from the unsupervised category. This model clusters the instances into k clusters, where k in this case is a positive number which the user provides. The aim in this process is to extract useful information from the data itself and the reason why this is performed is to model the behavior of time series and to point out the pattern of the inputs clustering the data. Clustering data term refers to the task of splitting the data into several groups. The created groups will contain data with similar traits assigned to the clusters.

2.4.2.2 Hierarchical clustering

This method of cluster analyzing attempts to create and build a hierarchy of clusters. We can represent this hierarchy in the form of a tree structure often called the dendrogram where both roots and leaves are included. The root of the tree will represent the singular cluster with many observations and the leaves correspond to individual observations. This process can occur in two ways; Agglomerative the process starts from the leaves by merging the cluster together or divisive where it starts form the root and repetitively splits the cluster. A function of a linkage criteria of the pairwise distances among observations decides which cluster to merge or split. [27]

2.4.2.3 Gaussian Models

The gaussian process is a collection of random variables indexed by time or space in a way that every finite collection has a multidimensional normal distribution. Furthermore, this process is entirely defined by a mean function *m(x)* and a covariance function *k (x, x').* The mean is a vector and covariance is a matrix. This model is relatively new in the non-linear modellings.

$$f \sim GP(m, k)$$

Equation 8 Gaussian function

2.4.2.4 Cluster evaluation

Measuring the quality of clustering algorithms can be as crucial as the algorithm itself, we need to be able to choose the cluster which would perform better for the input data. As clusters main focus is to calculate the similarities of different objects without any previous information about the exact distribution of the data its highly recommend and required that these models are evaluated thoroughly. The validation process is a stand-alone task therefore not included in the clustering tasks. There are two types of clustering validation methods: Internal validation and external validation. Internal validation consists of comparing the evaluation with the end results, comparison of the structure found results and their relation to each other. External validation stands for comparing the results with other reference results which are considered the "ground truth". If the obtained result is almost like the reference result the output of this cluster would be identified as "good" cluster. This evaluation is mainly suggested and used for synthetic data. [28]

2.4.3 Reinforcement learning

First thing to mention about reinforcement learning are the three components needed to explain this case; agent which is the learner, or the decision makes, environment consists of everything that interacts with the agent and the last component is the action what the agent does. [21] The learning data in this case interacts with the environment in discrete steps of time. Every time the learning data receives an observation it chooses an action from the set of the available set of actions. The objective is this case is for the agent to find the action that maximizes the expected reward at a specified time sequence. In other words, these algorithms aim to learn the best policy. [29]

2.5 Ensemble Learning

The combined diverse models are called ensemble systems, also known as classifier systems or committee-based learning. Ensemble learning is the process in which several different models are deliberately created and combined to solve a particular computational problem. This model is used to improve the classification, prediction, and proximality. [30] We can also refer to this learning algorithm as training of various multiple learning machines, combining their outputs as one and treating them as "group" decision makers. The goal in this case is for the model to have better overall accuracy than the individual models. Ensemble method can be quite simple to implement as it can be extremely difficult for the many hundreds of hours that has to be spent in their study. [31] The term itself ensemble is used to name the method that generates several hypotheses always using the same base learner. Base learners is a concept used to identify multiple learners commonly trained by a base learning algorithm which can be decision tree, neural network etc. The model is usually useful for these mentioned fast algorithms such as decision trees, but nevertheless it can be also for slow algorithms. The struggle with this model is at the time spent to do the evaluation of the combined hypotheses, its time consuming to decide which of these hypotheses is the most accurate one. Some models will use a specific learning algorithm that will produce a homogenous ensemble, and others that use multiple learning algorithms that will produce heterogenous ensembles. Why do we use ensembles? Easy, they are one of the simplest to implement and researchers have managed to achieve success using it. [32].

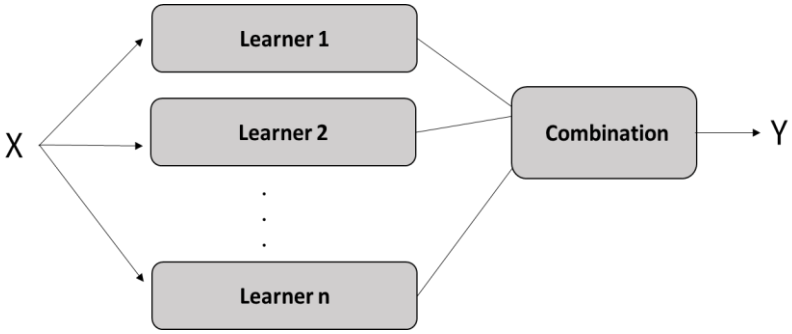

Figure 8 Common Ensemble Architecture

2.5.1 Boosting

This method first emerged to combine several classifiers to produce a powerful "committee". Boosting is a generic algorithm rather than a particular model. It is used to boost the accuracy of any given algorithm. [33] The way this model works differs from others as in this type of learning method you need to identify the "problematic" model so that it can improve it. Simple planning results in wide improvements of the classification performance. Two major algorithms we can mention are AdaBoost and Gradient Boosting. [34]

$$\boldsymbol{H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)}$$

Equation 9 AdaBoost Algorithm by Freud & Schapire,1995

2.5.2 Bagging

A prime example of the ensemble learning can be the bagging algorithm. It is easily implemented with just a few lines of code. It is used and specifically designed to improve the stability and accuracy of machine learning algorithms. It is extensively implemented in statistical classification and regression. This model is also designed to reduce the variance and to prevent overfitting. [21]. Bagging is the short form of "bootstrap aggregation" first introduced by Breiman in 1996 as the device to reduce the prediction error in learning algorithms. This model is executed by outlining bootstrap samples from the training sample. The algorithm is applied to each bootstrap and so averaging the resulting predictions. As already mentioned, Breiman was the first one to actually provide empirical evidence that the prediction error was in fact reduced. [35]. Bagging is mostly applied in decision tree methods. It is a special method of the model averaging approach.

2.5.3 Random subspace

Random subspace method is often referred to as attribute/feature bagging, meaning this one is also an ensemble learning which aims to reduce correlation in estimators of an ensemble by randomly training samples or features instead of the complete set/feature. This machine learning algorithm incorporates predictions from different decisions trees on different subsets of columns from the training dataset. Features in these algorithms are known are else called as subspaces, so in our case different subspaces will provide different point of views to the data. This method works well in practice fact this proven when the datasets have large number of subspaces and samples. [36]

2.6 Evaluation of model accuracy

One of the most crucial processes of this work is to confirm how accurate our models are. Evaluation models differ from classification models to regression models. So before proceeding with implementation of these models we need to know the used method. In our case as already stated throughout this thesis are used the Regression models. As we are trying to predict based and depending on continuous ranges of data instead of an exact dataset as in classification method, we need to use these alternative metrics. To measure the accuracy, we often borrow statistical formulas which in almost all cases result true. Other way to measure is by graphs. Pictures below along with short description will give an overview of these graphs.

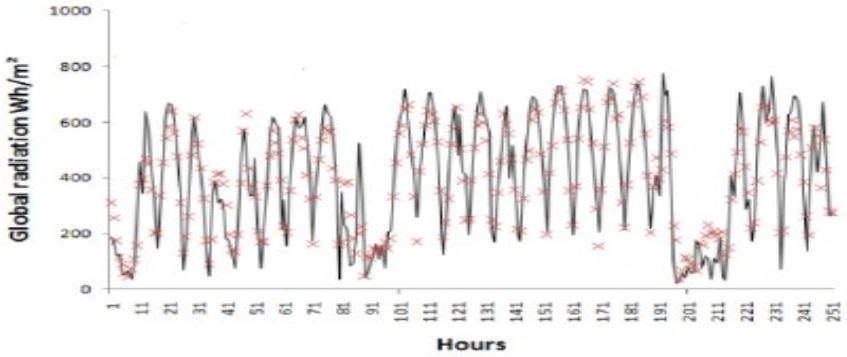

Figure 9 Time series graph

Figure 18 gives a representation of Time Series forecasting. This graph shows us the quality of the predicted data for radiation compared to the actual radiation.

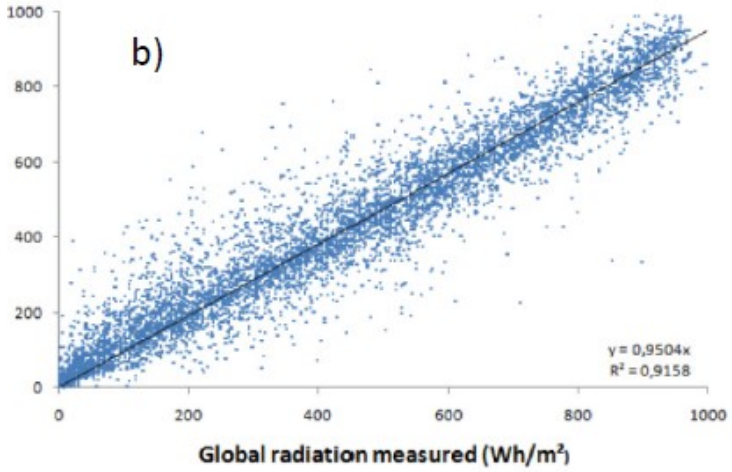

Figure 10 Scatter plot graph

Figure 19 is the example for scatter plot graphs, presenting the deviations and the systematic bias for the model. In this case the deviations depend on the radiation conditions.

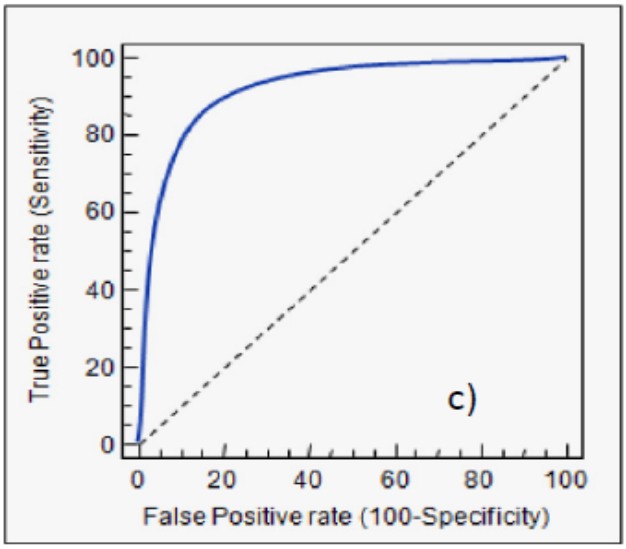

Figure 11 Receiver Operating Characteristics

Receiver Operating Characteristics (ROC) shows a curve with two valuations, *True positive rate* and *False positive rate.*

Evaluation models from statistics are detailed as below:

2.6.1 Explained Variance Model - EV

$$EV(y_{\text{true}}, y_{\text{pred}}) = 1 - \frac{\text{Var}(y_{\text{true}} - y_{\text{pred}})}{y_{\text{true}}}$$

Equation 10 Explained Variation Model

The variance model is designed to compare the predicted variance from the expected outputs and that of the variance error of our model. In other words what is presented above is the size of the variance that our model actually calculates from the original data.

2.6.2 Mean Biased Error – MBE

$$MBE = \frac{1}{N} \times \sum_{i=1}^{N} (\hat{y}(i) - y(i))$$

Equation 11 MBE

The mean biased error (MBE) formula as above will give the mean bias of the prediction.

Y – observed values

Y^ - predicted values

N – number of observations

This model will indicate the underestimated and overestimated observations. We cannot depend on this model fully, but we can understand the over and under estimations.

2.6.3 Mean Absolute Error – MAE

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} | \hat{y}(i) - y(i)$$

Equation 12 MAE

Mean Absolute Error (MAE)measures the squared difference between the observed and predicted values. Symbols previously described in figure 21 are valid for this model too.

2.6.4 Mean Squared Error (MSE)

$$MSE = \frac{1}{N} \times \sum_{i=1}^{N} (\hat{y}(i) - y(i))^2$$

Equation 13 MSE

Mean squared error is usually noted as the average of the mean differences between predicted and observed values. This model will just show if the prediction is correct or false no other details are provided.

2.6.5 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (\hat{y}(i) - y(i))^2}$$

Equation 14 RMSE

Root mean square error model will indicate the spread of error. So, if there is little error there will be no large spread. This model is more suitable for cases where small error spreads are expected.

2.6.6 Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{1}{N} \times \sum_{i=1}^{N} \left| \frac{\hat{y}(i) - y(i)}{y(i)} \right|$$

Equation 15 MAPE

Mean absolute percentage error is quite similar to MAE model, but in this case the gap that exists between the observed and predicted value is divided by the observed value to close the gap of error.

Last model to mention is: *Indexed Root Mean Square Error (nRMSE)*

$$nRMSE = \frac{\sqrt{\frac{1}{N} \times \sum_{i=1}^{N} (\hat{y}(i) - y(i))^2}}{\bar{y}}$$

Equation 16 nRMSE

The indexed model has a deficiency which is that it cannot determine a value when y is close to 0 as it has proven to be unstable. For this reason we cannot estimate error when y(i)=0.

2.6.7 R-squared Coefficient

$$R^2 = 1 - \frac{\sum_{w=1}^{N} (\hat{Y}(w) - Y(w))^2}{\sum_{w=1}^{N} (Y(w) - \bar{Y}(w))^2}$$

Equation 17 R-squared coefficient

This statistical measure will provide a close look to how close actually the data is to the regression line. In this case R is the explained variation divided by the total variation, it will always have a value between 0-100. If 0 it shows that the model can not explain the variability of the response variable around its mean. If 100 it shows that the data is actually fitted to the regression line. Easy conclusion out of this is the bigger the R-squared is the better is you model explained and fitted to the linear regression.

The reason why we have shown all if these evaluation models, statistically and graphically is to prove the statement that there is no standard evaluation model for accuracy for regression or classification. The user/applier shall choose base on the used dataset/variables/observations which of the model will estimate more efficiently. [21]

2.7 Problems that can occur

While working with machine learning algorithms, we must understand that we need to be very careful from the beginning when we have to choose the right algorithms form the whole of them. The only safe way is to fully know and understand the data set you have and what do you aim to achieve by this study. The algorithms will function correctly if the data chosen fits to its characteristics. There is no such thing as the best model for all problems, each problem needs to be treated with the appropriate model to get the most accurate output and prediction. In both classification and regression methods problems can occur if not the right algorithms are used. Machine learning itself as continuously mentioned on this study concentrates completely on data analysis, data evaluation, identifying, targeting, integrating, and pre-processing of the data. “No free lunch” (NFL) theorem from 1997 by David Wolpert & William Macready explains exactly the fact that one algorithm cannot be the best for all problems, or otherwise said that there is no singular algorithm model for machine learning to predict all models. The size of the problem can depend on the size of your data and its complexity, so know your case then choose the data and then the model. As our study is based on regression model, we will be taking a closer look to those. To treat these regression problems it is recommended and useful to firstly train the data with Stochastic Gradient Descent (SGD) first introduced by Robbins-Monro in the 1950’. Shortly explained, this method is used to optimize the targeted function with smoothness properties. For instance it will replace the actual gradient measured from the whole dataset with an estimated gradient chosen from subsets from the data itself. What are the pros and cons of this method? The SGD is a great help in linear models as its fast to act and if the data we possess it’s not complicated but on the other hand if the data is complex and very large in scale the difficult to handle if we do not know the type of relationship between data. Because of the nature and structure of regression method itself the problems can occur are in quantity or size that’s why the SGD is a good choice.

Chapter 3

Implementation

In this chapter it will be shown a brief description of the implementation of my thesis. The aim of this thesis is to analyze different methods, and more specifically the forecast of the solar radiation, and determine a model which provides us with the most accurate forecasted values.

The methodology of this thesis is separated in two main parts the first one will be an in implementation in E-views based on an econometric model which will be used to identify the relationships between different variables, including temperature, wind speed, pressure etc.

The second part is the comparison of different algorithms using Python our econometric model with the final purpose of predicting solar radiation for the period of 1/1/2020 – 1/31/2020.

Programs used to make the testing are EViews 11 and Spyder (Anaconda 3.7) and the main machine learning library used is Sklearn.

3.1 Introduction

Energy is one of the largest topics being discussed over the world last years and so is Solar radiation from being related to it. So, if we could predict Solar energy it would be a great information for a country, a company or even a family in order to invest or not in that type of energy. So here in this paper I am doing a study for that purpose.

First, I did an econometric modeling to determine the level of impact of meteorological factors on solar radiation and consequently in the forecasted values of the solar radiation. Furthermore, we use these this economical model on Machine Learning algorithms in order to compare them and see which of them has the best accuracy. In order to reach to a conclusion first we have to go through some stages.

3.2 Data Set used

The dataset for this study is provided by NASA (National Aeronautics and Space Administration) taken from Soda (Solar Radiation Data) website. Dataset has hourly data for three years from 1/1/2017 until 1/1/2020 obtained from MERRA 2 (Modern-Era Retrospective Analysis for Research and Applications version 2). The location of this dataset is Tirana with latitude 41.328 and longitude 19.789. Solar Radiation (Wh/m2) in this data is taken from GHI. (Global irradiation on horizontal plane at ground level) which is a calculation of sum of BHI (Beam irradiation on horizontal plane at ground level) and DHI. (Diffuse irradiation on horizontal plane at ground level)

Our data were presented on hourly frequency. However, during the night hours some indicators took the value 0 and during the day they took values different from 0. We have averaged the values to better reflect the average solar radiation per day and in order to remove the seasonal effect during the analysis.

Variables that our dataset contains are:

- Temperature (Temp) Is a physical quantity that expresses hot and cold. Unit of measurement in: Kelvin (K) - Temperature at 2 m above ground.
- Relative humidity (RH) Is the ratio of the water vapor pressure to the saturation vapor pressure at a given temperature - Unit of measurement (%). Relative humidity at 2 m above ground.
- Pressure (Pres) Is the force per unit of area exerted on the Earth's surface by the weight of the air above the surface. Unit of measurement: (hPa) - Pressure at ground level.
- Wind speed (WS-m/s) Is the rate at which wind covers distance. Unit of measurement: meters per second (m/s) - wind speed at 10 m above ground.
- Wind direction (WD-deg) at 10 m above ground (0 means from North 90 from East...);
- Rainfall (RF) the amount of water falling in rain. Unit of measurement: (kg/m2)
- Snowfall (SF) the quantity of snow falling. Unit of measurement: (kg/m2)
- Snow depth (SD) depth of the new and old snow remaining on the ground (m)
- Solar Radiation (SR)- Is a general term for the electromagnetic radiation emitted by the sun. Unit of measurement: Watt-Hours per square meters (Wh/m^2);

3.3 Data Preprocessing

Before going to model selection one of most important steps to do is data preprocessing since as a saying says if the data is good the model will be good as well.

So first let's observe the dataset using the python command:

```
print(dataset.head(10))
```

Index	# Date	UT time	Temperature	Relative Humidity	Pressure	Wind speed	Wind direction	Rainfall	Snowfall	Snow depth	Solar Radiatio
0	1/1/2017	1:00	271.76	74.36	974.45	2.99	90.32	0	0	0.002143	0
1	1/1/2017	2:00	271.89	72.14	974.07	2.78	91.96	0	0	0.002143	0
2	1/1/2017	3:00	272.1	69.82	973.58	2.6	93.35	0	0	0.002143	0
3	1/1/2017	4:00	272.3	67.97	973.21	2.47	95.4	0	0	0.002144	0
4	1/1/2017	5:00	272.5	66.42	973.2	2.36	96.65	0	0	0.002144	0
5	1/1/2017	6:00	272.72	64.84	973.36	2.23	96.38	0	0	0.002144	0
6	1/1/2017	7:00	273.12	64.05	973.53	2.02	94.7	0	0	0.002144	34.7529
7	1/1/2017	8:00	275.16	64.46	973.72	1.76	93.41	0	0	0.002144	168.25
8	1/1/2017	9:00	277.11	57.53	973.93	1.65	97.76	0	0	0.00211	299.663
9	1/1/2017	10:00	278.74	51.83	973.83	0.73	119.12	0	0	0.002006	396.116
10	1/1/2017	11:00	279.88	49.22	973.3	0.62	247.08	0	0	0.001851	439.24

Table 1 Dataset snapshot

From here we clearly can see the first 10 row. From this snapshot we can see that our dataset contains 11 columns. Temperature is in Kelvin so we want to change it in degrees Celsius. Also, there might be missing values so we have to check that too. If there is a missing value, we have to choose either to drop these values or to fill them with mean or mode. In order to check that we use the following command:

```
Print(dataset.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26304 entries, 0 to 26303
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   # Date              26304 non-null  object
 1   UT time             26304 non-null  object
 2   Temperature         26304 non-null  float64
 3   Relative Humidity   26304 non-null  float64
 4   Pressure            26304 non-null  float64
 5   Wind speed          26304 non-null  float64
 6   Wind direction      26304 non-null  float64
 7   Rainfall            26304 non-null  float64
 8   Snowfall            26304 non-null  float64
 9   Snow depth          26304 non-null  float64
 10  Solar Radiation     26304 non-null  float64
dtypes: float64(9), object(2)
memory usage: 2.2+ MB
None
```

Table 2 Dataset info

From this information Table we can see what our dataset contains. It shows that it has 26304 rows and 11 columns. From here we can see that we don't have any missing values so we don't have to remove values or fill them with the average. So now let see the dataset statistics

```
Print(dataset.describe())
```

Index	Unnamed: 0	Temperature	Relative Hum	Preasure	Wind Speed	Wind direction	Rainfall	Snowfall	Snow depth	Solar Radiation
0	count	26304	26304	26304	26304	26304	26304	26304	26304	26346
1	mean	15.2176	67.241	965.599	3.05854	164.501	0.133273	0.00516	0.001242	183.395
2	std	8.28318	17.6744	5.5428	1.88295	95.9474	0.458517	0.069598	0.005828	262.931
3	min	-12.73	15.62	938.48	0.04	0	0	0	0	0
4	25%	8.82	54.95	962.62	1.68	82.465	2e-06	0	0	0
5	50%	14.82	70	965.91	2.69	149.175	0.001151	0	0	7.6169
6	75%	21.42	81.43	969.02	3.99	250.55	0.050719	0	1e-06	320.247
7	max	37.56	103.41	982.46	14.87	359.9	10.4605	3.72669	0.076181	1008.51

Table 3 Dataset statistics preprocessed

At this table we can see mean, standard deviation, min value, max values, 25%, 50% and 75% percentiles for every column. From her we can see that our data is a bit right skewed.

3.4 Data Evaluation

Before doing our model is better to have more information of the data by visualizing and exploring the variables of the dataset so we later we can use this information to from hypothesis or models. To achieve we use plotting firstly let's see the histogram by using the command:

```
dataset2=dataset[['Temperature','RelativeHumidity','Pressure','W
indSpeed','Rainfall','Snowfall','SnowDepth','SolarRadiation' ]]

dataset2.hist(bins=25, color = '#ec838a')

plt.suptitle('Histograms of Variables\n', horizontalalignment=
"center", fontsize=20, fontstyle="normal")

plt.tight_layout(rect=[0, 0.03, 1, 0.95])

plt.savefig('albi.png')
```

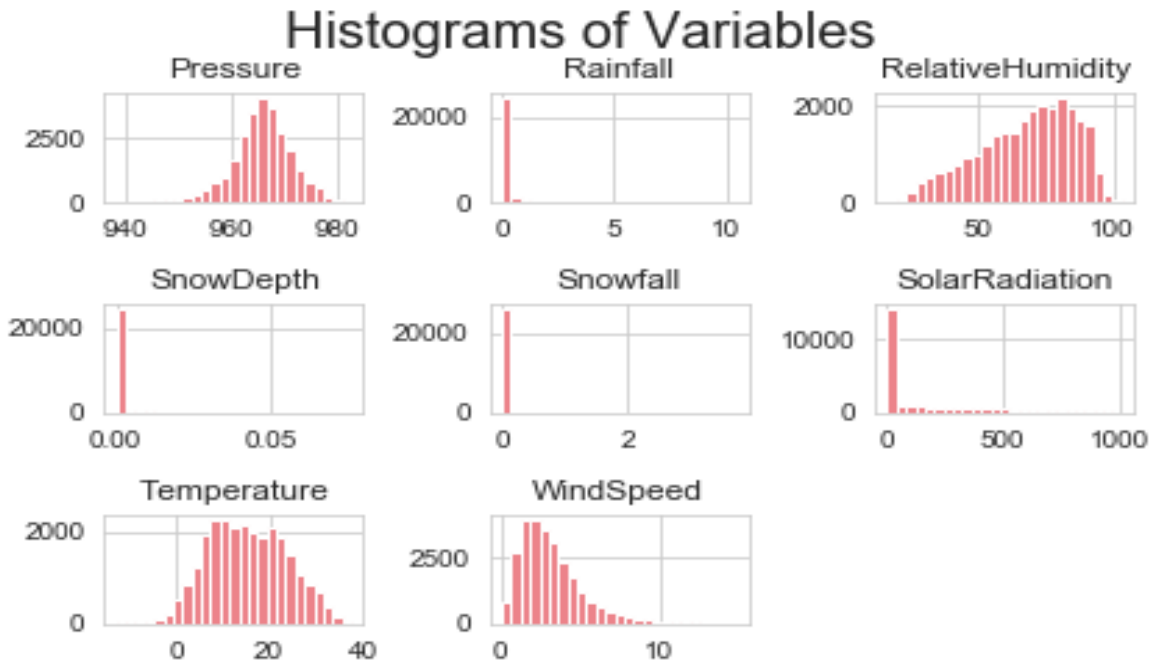

Figure 12 Histograms of Variables

From this histogram of variables, we can make some observations:

Pressure distribution is mostly concentrated from 960hPa to 970hPa.

From Rainfall variable we can see that values mostly are 0 indicating that overall, there is not too much rain in Tirana.

Relative Humidity plot shows that the most of values range from 50% to 90% with a peak at 70%.

Snow Depth and snowfall are both lining at value 0 so we can say that there is now snow at this location.

Solar Radiation most of its values it has equal to 0 this makes sense due to night time and other values are spread up to 1000.

Most of Temperature values are from 8 to 25 degrees C.

Wind speed has most of its values from 2 to 5 m/s.

Another important step to check is correlations of variables since to check this we use the following code:

```
dataset2=datasest[['Time','Temperature','RelativeHumidity',
'Pressure','WindSpeed','Rainfall','WindDirection','Snowfall',
'SnowDepth', 'SolarRadiation']]

plt.rcParams['figure.figsize'] = (30, 20) # plot size

cmap = sn.diverging_palette(220, 12, as_cmap=True)

pd.set_option("display.max.columns", None)

sb.heatmap(dataset2.corr(), annot = True, cmap = cmap)

plt.show()
```

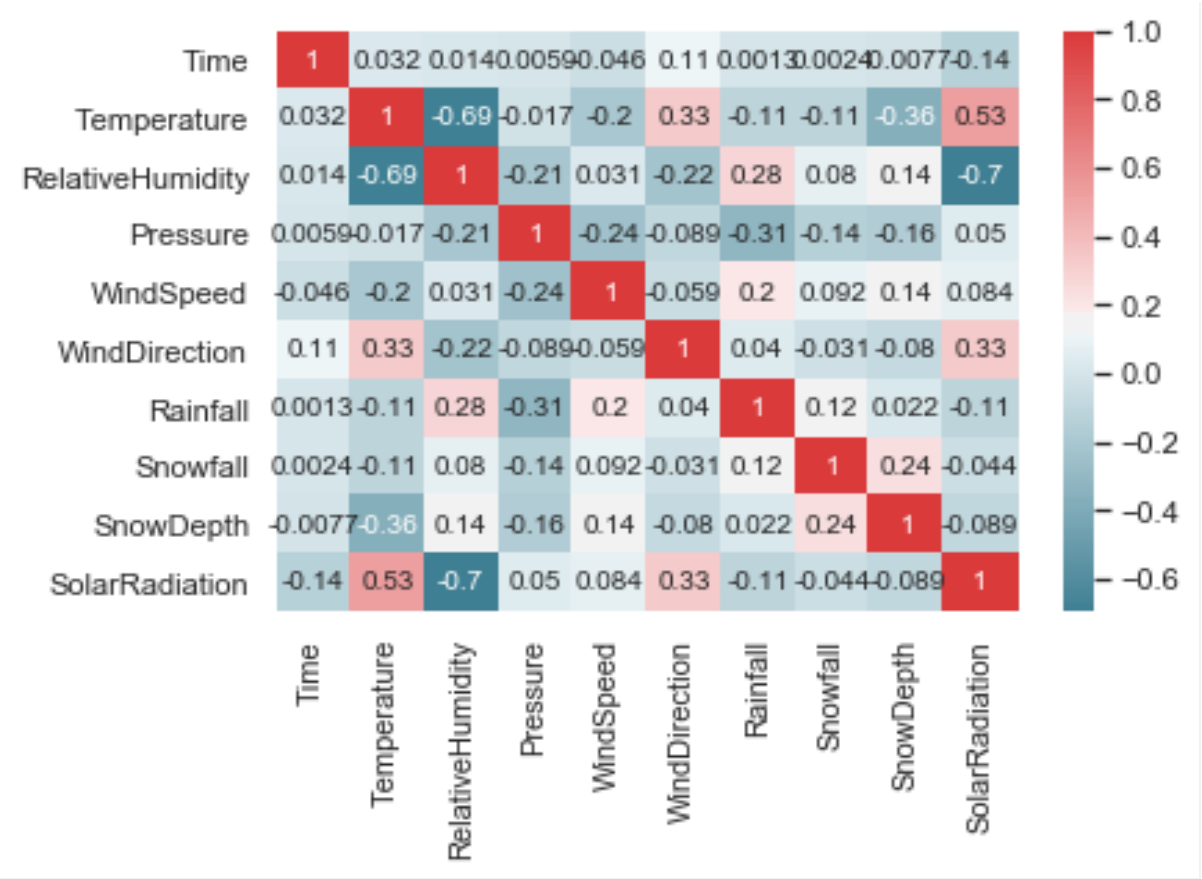

Figure 13 Correlation of Variables

From this plot we can check the correlation of all variables with each other. We can see all relationships of variables by seeing the positive or negative number respectively. But, in our study we are interested more in Solara Radiation so to have a clearer look we can use the following code:

```
corr = dataset2.corrwith(dataset2.SolarRadiation)

corr = correlations[correlations!=1]

positive_corr = corr[corr >0]. sort_values(ascending = False)

negative_corr =corr[corr <0].sort_values(ascending = False)

corr = dataset2.corrwith(dataset2.SolarRadiation)

correlations.plot.bar(figsize=(20,10),color='#e30918',fontsize=3
0, grid = True)
```

```
plt.text(7,-
0.7,'AlbiTrashani',fontsize=28,ha='center',va='center',
color='grey', alpha=0.5)

plt.title('Correlation with Solar Radiation \n', fontsize = "36",
horizontalalignment="center", fontstyle = "normal")
```

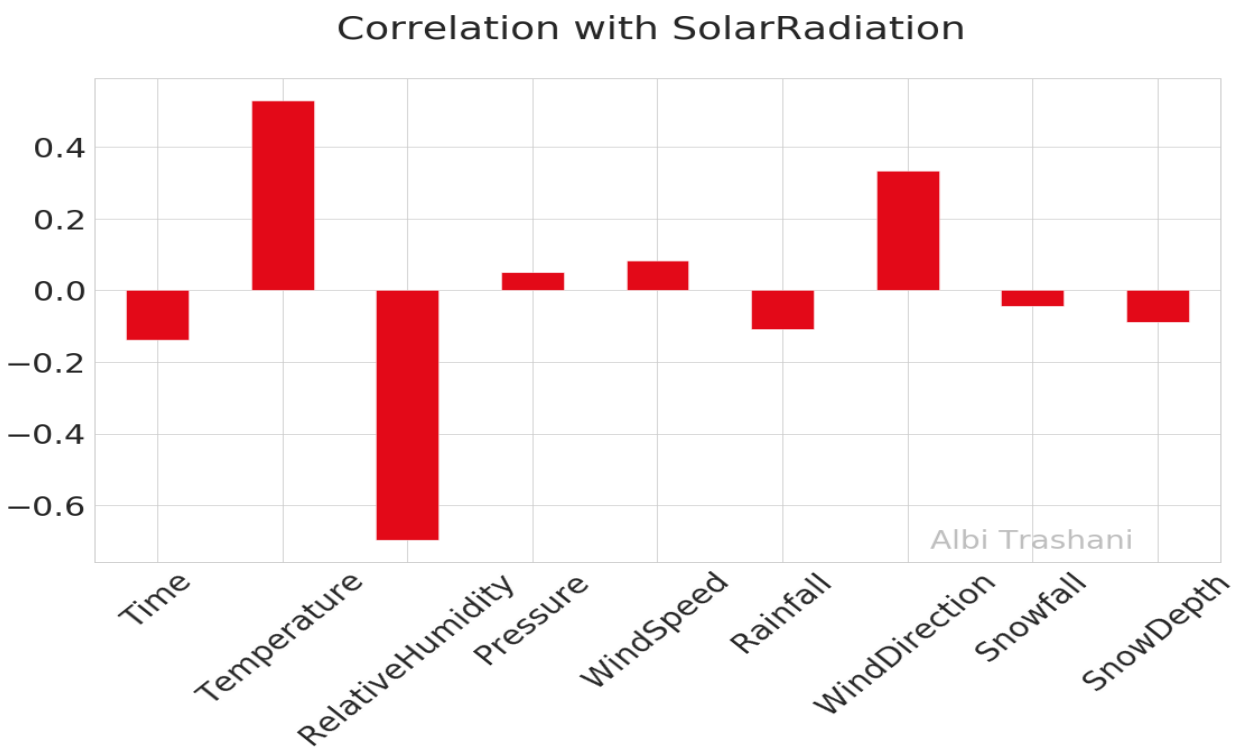

Figure 14 Correlation of variables with SR

From this figure we can clearly see which variables have positive or negative correlation to Solar radiation. Here from the plot it can easy notice variables from which we can mention Temperature with a value 0.5 and Wind Direction with value 0.35 having a positive correlation and Relative humidity with a value -0.7 with negative one.

After observing all of these from here we can say that there most of values of snowfall, snow depth, and wind direction are 0 or near 0 so apparently, they don't have any effect on Solar Radiation. 0 values for Solar radiation can be explained because of the day and night so I found daily average value for each variable and after this each variable will have 1096 observations which are more than enough to build an accurate econometric model.

After these observations I decided to use for my model only 5 methodological factors since these independent variables would not be suited to be used on our model later.

Once more let's see the dataset information with the chosen variables that we will use on my model.

	Date	Date	Date	Temperature	RelativeHumidity	Pressure	WindSpeed	SolarRadiation
0	2017	1	1	2.86583	60.5721	972.977	1.46042	100.461
1	2017	1	2	4.49167	75.1763	971.034	1.4675	62.5681
2	2017	1	3	4.82458	89.4596	968.136	2.92417	36.1373
3	2017	1	4	4.72708	90.5771	962.344	2.31375	37.167
4	2017	1	5	3.19417	88.4	951.723	4.91958	33.1802
5	2017	1	6	-4.77292	72.8846	958.583	8.73625	57.0167
6	2017	1	7	-9.45208	75.0008	965.648	8.05667	80.0074
7	2017	1	8	-8.11542	72.7404	966.349	4.98708	95.5977
8	2017	1	9	-6.81875	69.4496	964.038	5.0575	75.7749

Figure 15 Dataset Preprocessed

```
MultiIndex: 1096 entries, (2017, 1, 1) to (2020, 1, 1)
Data columns (total 5 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   Temperature       1096 non-null   float64
 1   RelativeHumidity  1096 non-null   float64
 2   Pressure          1096 non-null   float64
 3   WindSpeed         1096 non-null   float64
 4   SolarRadiation    1096 non-null   float64
dtypes: float64(5)
memory usage: 46.5 KB
```

Figure 16 Info of our dataset preprocessed

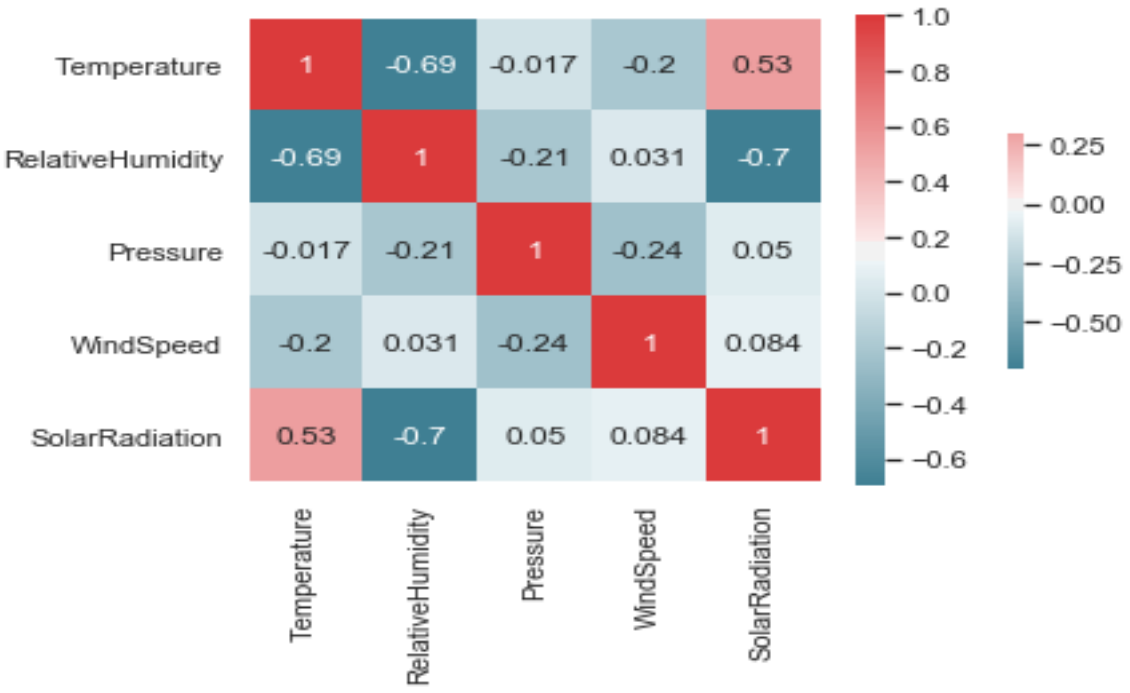

Figure 17 Correlation of selected variables

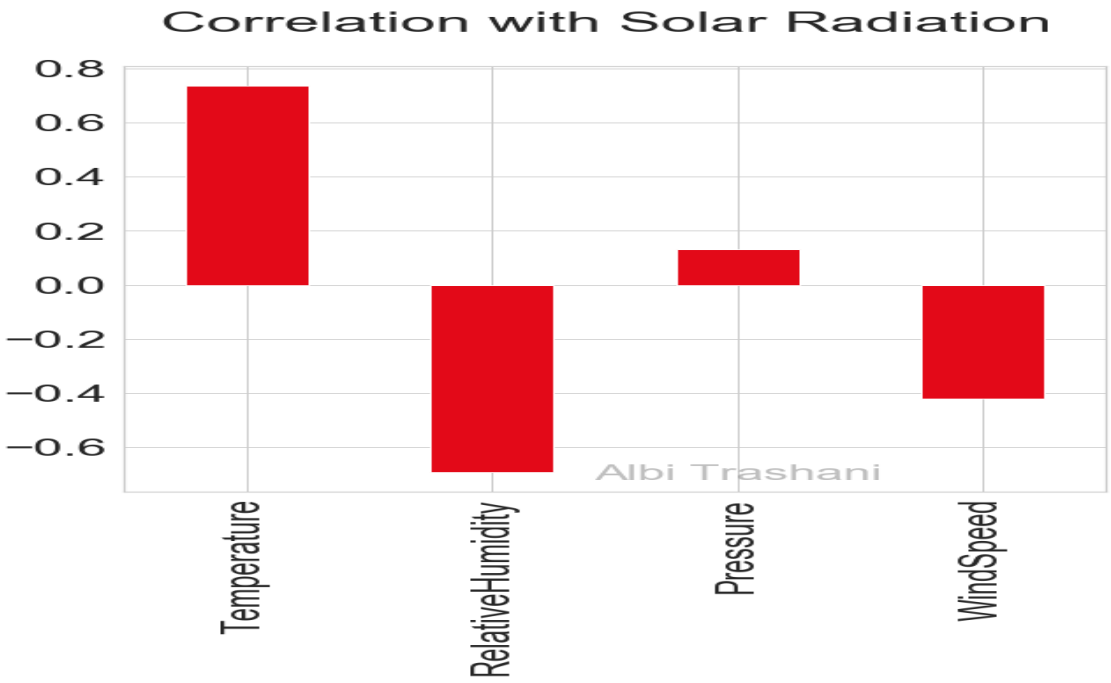

Figure 18 Correlation of SR with selected variables

A better way to observe linear relation of our chosen dataset variables is scatter plot. Using "SolarRadiation" attribute as y value and other variables as x value. In order to show these plots, we use the following code:

```
for x in range(0, 4):
        plt.figure(figsize=(10, 7))
        plot1 = sb.scatterplot(i[x],i[5],data =  dataset, color =
                                'orange', s = 250)
      plt.title('{} / Solar Radiation'.format(i[x]), fontsize =
16)
        plt.xlabel('{}'.format(i[x]), fontsize = 20)
        plt.ylabel('Solar Radiation', fontsize = 20)
        plt.xticks(fontsize = 16)
        plt.yticks(fontsize = 16)
        plt.show()
```

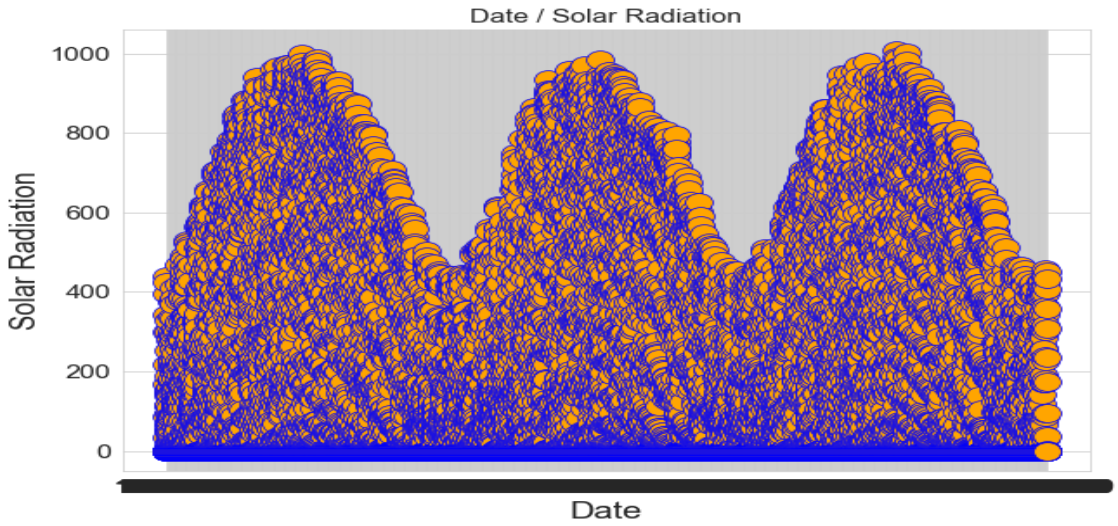

Figure 19 Plot between Date and Solar Radiation

From figure 19 we can see that there are three peaks, this because dataset contains values of three years form 2017 to 2020. The highest values are seen in the middle of each year, which

makes sense since are the summer values. Overall, it can be seen that there are high values ranging from 400 Wh/m^2 to 1000Wh/m^2.

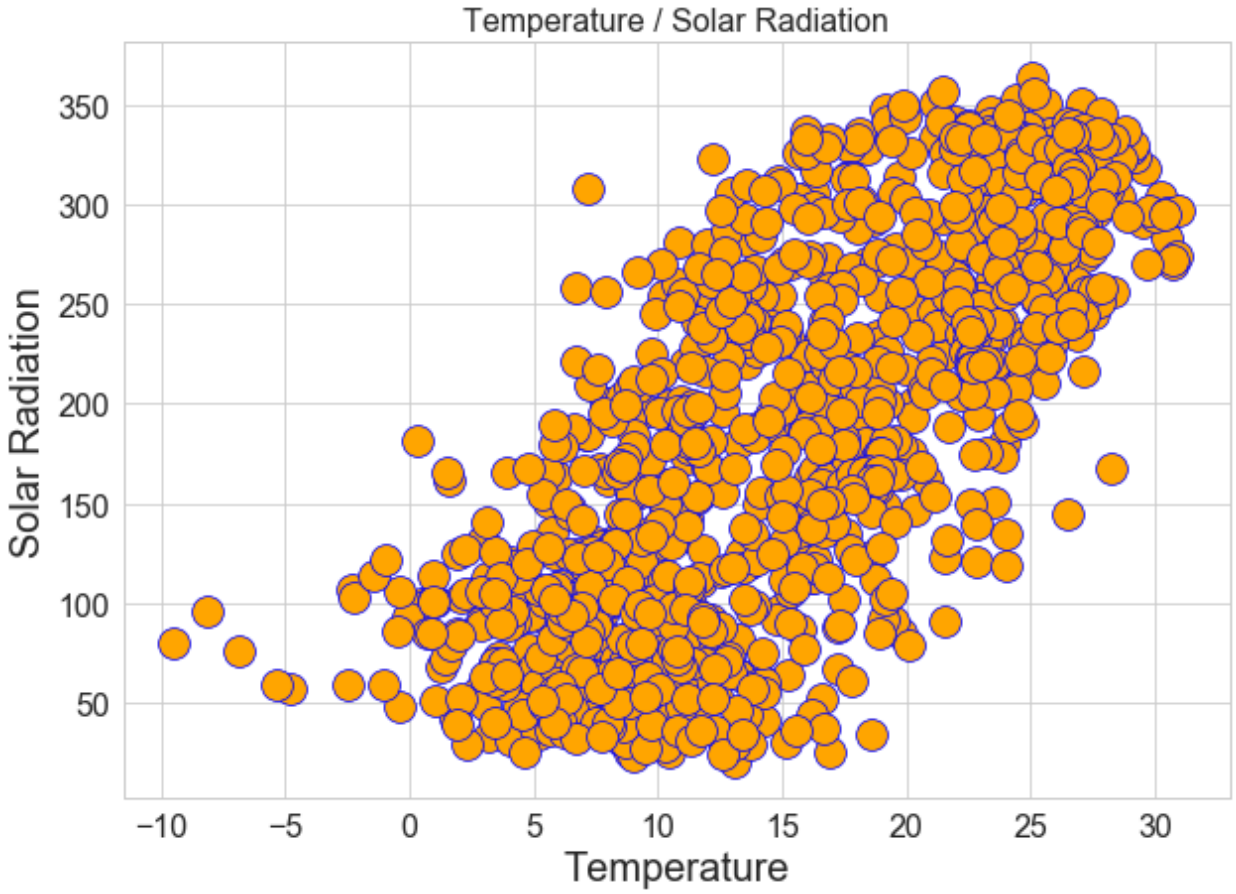

Figure 20 Plot between Temperature and Solar Radiation

Figure 20 shows Temperature relation with solar radiation and it can be seen that there is a positive correlation between them. Higher values of temperature higher values of solar radiation. Although the correlation is not to much since values are a lot spread.

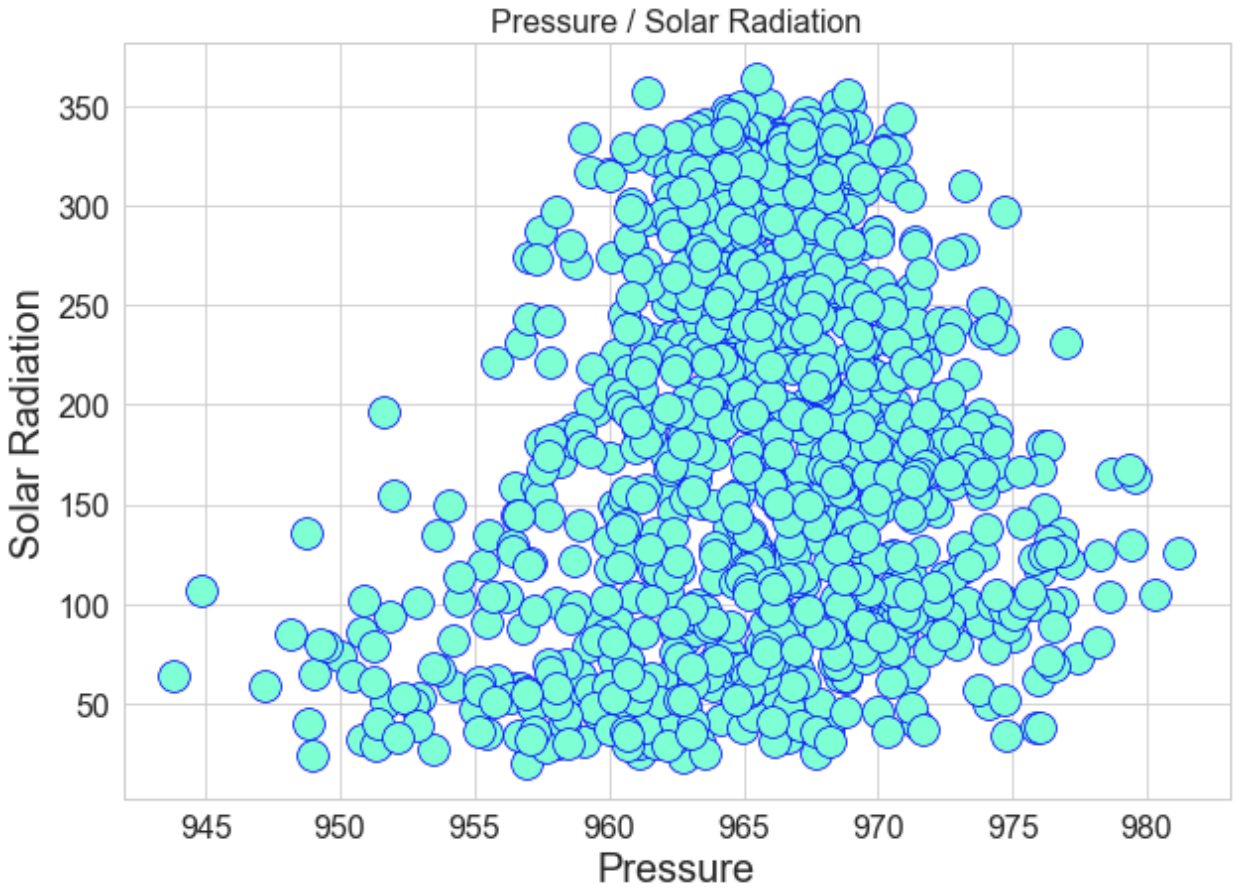

Figure 21 Plot between Pressure and Solar Radiation

The above plot shows pressure that ranges from 950 to 980 hPa, and it can be seen that there is a little relation with solar radiation.

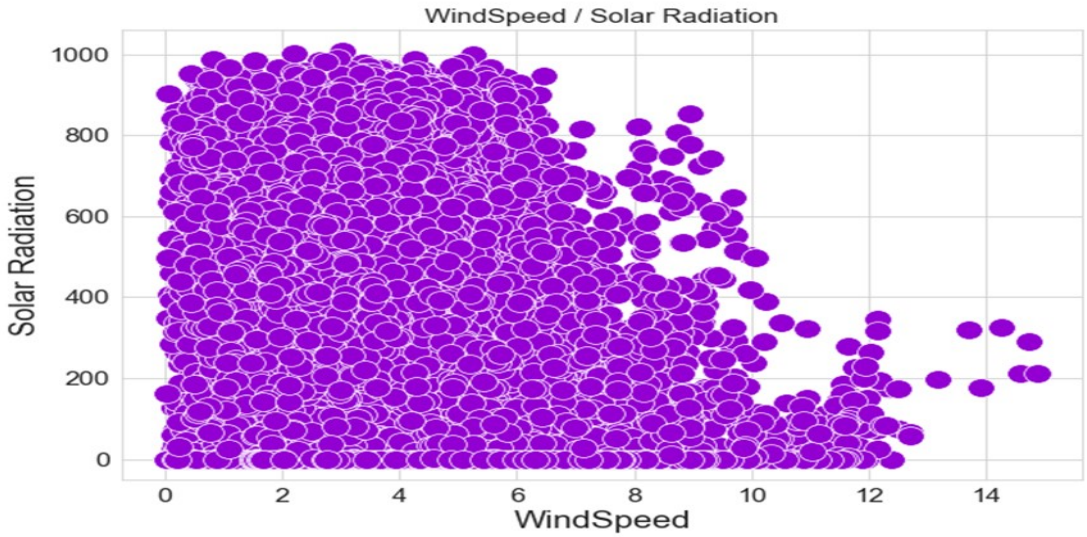

Figure 22 Plot between Wind Speed and Solar Radiation

From wind figure 11 we can say that the lower wind speed is the higher solar radiation we have.

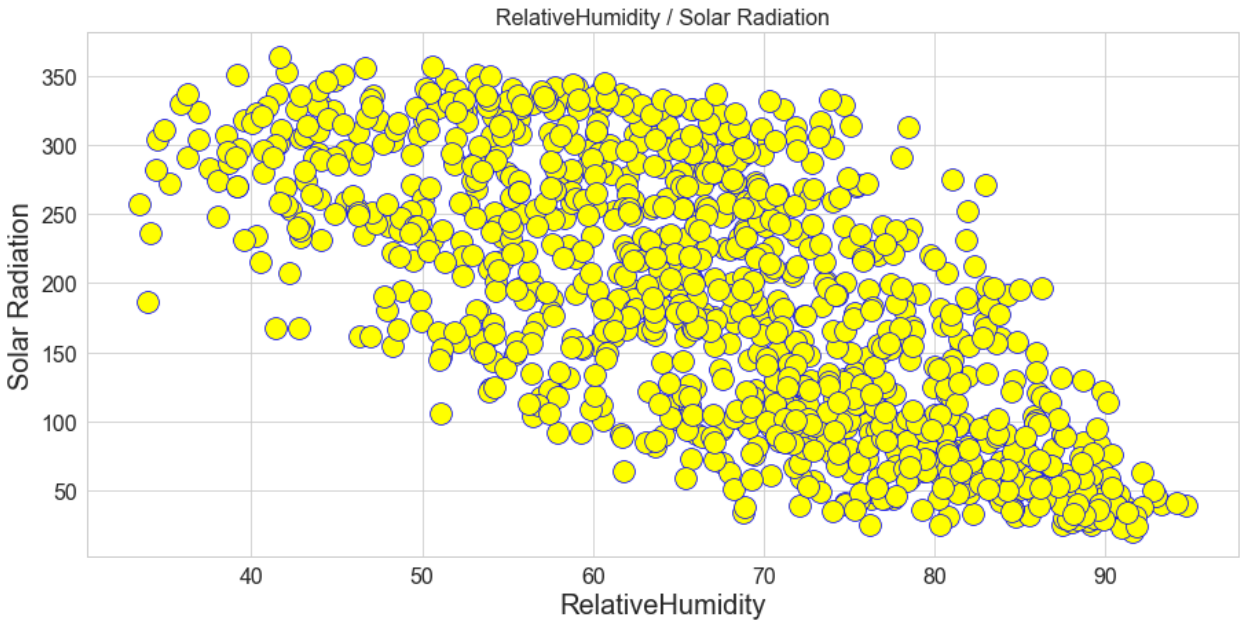

Figure 23 Plot between Relative Humidity and Solar Radiation

Based at this plot we can say that there is a negative correlation between the two variables tested.

3.5 Econometric Model and statistical considerations

In order to measure the impact of the above-mentioned factors on solar radiation and thereafter perform the forecast, we will be evaluation a multi variable regression. We will proceed by evaluation the regression in terms of statistical significance and estimation functional form of the model will be logarithmic of the *log – log* type. The dependent variable on this model will be the solar radiation indicator expressed in Wh/m^2. Meanwhile, temperature, wind speed, pressure and relative humidity will be used as independent variables in this model. The model in algebraic form is presented as follows:

$$\log(SolarRadiation) = \beta_1 + \beta_2 \log(\text{Temp}) + \beta_3 \log(WS) + \beta_4 \log(Pres) + \beta_5 \log(RH) + \varepsilon_i$$

Equation 18 Log - log type model

when: ε_i represents the error term (residuals) in this model.

To check the results, we use the above model and test it in EViews:

Dependent Variable: LOG(SOLARRADIATION)
Method: Least Squares (Gauss-Newton / Marquardt steps)
Date: 02/15/21 Time: 23:58
Sample: 1/01/2017 1/01/2020
Included observations: 1096
LOG(SOLARRADIATION)=C(1)+C(2)*LOG(TEMPERATURE)+C(3)*LOG(WINDSPEED)+C(4)*LOG(PRESSURE)+C(5)*LOG(RELATIVEHUMIDITY)

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	-81.79585	16.16403	-5.060364	0.0000
C(2)	5.447428	0.589497	9.240811	0.0000
C(3)	-0.255373	0.027615	-9.247773	0.0000
C(4)	8.791890	2.139627	4.109077	0.0000
C(5)	-0.877523	0.062154	-14.11853	0.0000

R-squared	0.561904	Mean dependent var	5.656485
Adjusted R-squared	0.560298	S.D. dependent var	0.537022
S.E. of regression	0.356100	Akaike info criterion	0.777339
Sum squared resid	138.3464	Schwarz criterion	0.800147
Log likelihood	-420.9818	Hannan-Quinn criter.	0.785969
F-statistic	349.8309	Durbin-Watson stat	0.867202
Prob(F-statistic)	0.000000		

Table 4 The preliminary results of the model

Therefore, our selected equation can be expressed as following:

$$\log(SolarRadiation) = -81.9 + 5.44\log(\text{Temp}) - 0.26\log(WS) + 8.79\log(Pres) - 0.88\log(RH)\,\varepsilon_i$$

As set out above, the selected model is based on the multiple regression model. For the evaluation of the model, I will use the Ordinary Least Squares Method (OLS). For this estimate to serve to draw conclusions with high statistical reliability, we will rely on all the basic assumptions of the Gauss-Markov Theorem [37](Gujarati, D.).According to this theorem for a regression model to be statistically significant and useful for economic analysis must meet several conditions [38], [39]:

1. First condition: the model must be linear or logarithmic to apply the ordinary least squares method. Linearity should be according to the parameters b_i.
2. Second condition: the model must have a single parameter b0 (called the ordinate at the origin), i.e. the mathematical expectation of the residuals is E (ε) = 0.
3. Third condition: the model must have the residual variance ε Constant, i.e. there must be no heteroskedasticity because the b_i parameters are irregular.
4. Fourth condition: the model must not have residual autocorrelation, i.e. covariance Cov (εi; εj) = 0 for each i ≠ j.
5. Fifth condition: the model must not have multicollinearity, i.e. statistically significant correlation between the independent variables x_i between them or the variables and residues ε of the model, for each i = 1, 2, ..., n.

By meeting all these conditions, the model is free of all its invaluable deviations and coincidences, making the ordinary least squares method the best explanatory and interpretive technique, i.e. minimizing any error term:

$$min\left\{\sum_i^n (y_i - \hat{y}_i)^2\right\} = min\left\{\sum_i^n (\varepsilon_i)^2\right\}$$

Equation 19 Error term

for i = 1, 2, …, n

Y denotes the actually observed value of the dependent variable and y_i denotes the value obtained of the variable *y* by regression.

In all regression models with data that form time series, some econometric tests should be considered for the validity and usefulness of these models. The most important tests and their meaning are summarized in the following section.

3.5.1 Model's Assumptions and respective testing

Model Significance Hypothesis

The hypotheses that will be tested in the following section are related to a series of criteria that must be met in order for the model to be statistically significant and valid for further analysis.

To test the overall (global) significance of the model, the following hypotheses will be tested:

Hypothesis 0 (H_0): The model is statistically insignificant ($\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$)

Alternative Hypothesis (H_a): The model is statistically significant (at least one of the coefficients is different from 0).

According to the results obtained from the program E-Views 11, the indicator of the observed Fisher is higher than that of the critical Fisher and its probability is less than 5%. This indicates that H_0 falls, so the model for Tirana with these parameters is statistically significant. In addition, the indicator of the adjusted R – squared is equal to 56.2%, so the explanatory nature of the estimated model is moderately high. The R – square value indicates that 56.2% of the "y" are explained by the "x" variables.

Coefficients' Significance

Despite the global significance of the model, we should also evaluate the coefficients of each indicator. These coefficients indicate the level of impact that each one of them has on solar radiation when the other indicators remain unchanged.

The β_0 coefficient is estimated at -81.79585. This coefficient indicates that if all other variables remain the same over time, the solar radiation level will decrease by around 81%. However, in practice, these conditions cannot be implemented.

Temperature is considered an important determinant of solar radiation. To test the impact of this variable on determining the level of solar radiation the following hypotheses will be tested:

a) Hypothesis 0 (H0): Temperature does not affect the determination of Solar radiation ($\beta_2 = 0$)

b) Alternative Hypothesis (Ha): Temperature affects the determination of Solar Radiation ($\beta_2 \neq 0$)

Temperature coefficient is positive and in the value of 5.447428 and the probability is less than 5%, so this coefficient is statistically significant. This means that an increase by 1% in the temperature level will bring an increase in solar radiation by 5.45% in Tirana.

Wind speed is considered an important determinant of solar radiation. To test the impact of this variable on determining the level of solar radiation the following hypotheses will be tested:

a) Hypothesis 0 (H_0): Wind Speed does not affect the determination of Solar Radiation ($\beta_3 = 0$)

b) Alternative Hypothesis (H_a): Wind Speed affects the determination of Solar radiation ($\beta_3 \neq 0$)

The Wind speed coefficient is negative and in the value of -0.255373 and the probability is less than 5%, so this coefficient is statistically significant. This means that an increase by 1% in the level of wind speed will bring a decrease in solar radiation values by 0.26% in Tirana.

Pressure is considered an important determinant of solar radiation. To test the impact of this variable on determining the level of solar radiation the following hypotheses will be tested:

a) Hypothesis 0 (H_0): Pressure does not affect the determination of Solar Radiation ($\beta_4 = 0$)

b) Alternative Hypothesis (H_a): Pressure affects the determination of Solar radiation ($\beta_4 \neq 0$)

Pressure coefficient is positive and in the value of 8.791890 and the probability is less than 5%, so this coefficient is statistically significant. This means that an increase by 1% in pressure level will bring an increase in solar radiation by 8.79% in Tirana.

Relative Humidity is considered an important determinant of solar radiation. To test the impact of this variable on determining the level of solar radiation the following hypotheses will be tested:

a) Hypothesis 0 (H_0): Relative Humidity does not affect the determination of Solar Radiation ($\beta_5 = 0$)

b) Alternative Hypothesis (H_a): Relative Humidity affects the determination of Solar radiation ($\beta_5 \neq 0$)

The relative humidity coefficient is negative and in the value of -0.877523 and the probability is less than 5%, so this coefficient is statistically significant. This means that an increase by 1% in the level of relative humidity will bring an decrease in solar radiation by 0.88% in Tirana.

Residuals' Normal Distribution

One of the econometric problems associated with the above model is the distribution of the model's residuals. In order for the model to show stability and be valid for further analysis, we need to study the residuals' distribution, and more specifically the residuals must be normally distributed. To test the normality of the model's residuals distribution the following hypotheses are tested:

a) Hypothesis 0 (H0): The model's residuals are normally distributed

b) Alternative Hypothesis (H1): The model's residuals are not normally distributed.

Using EViews commands we get the following table:

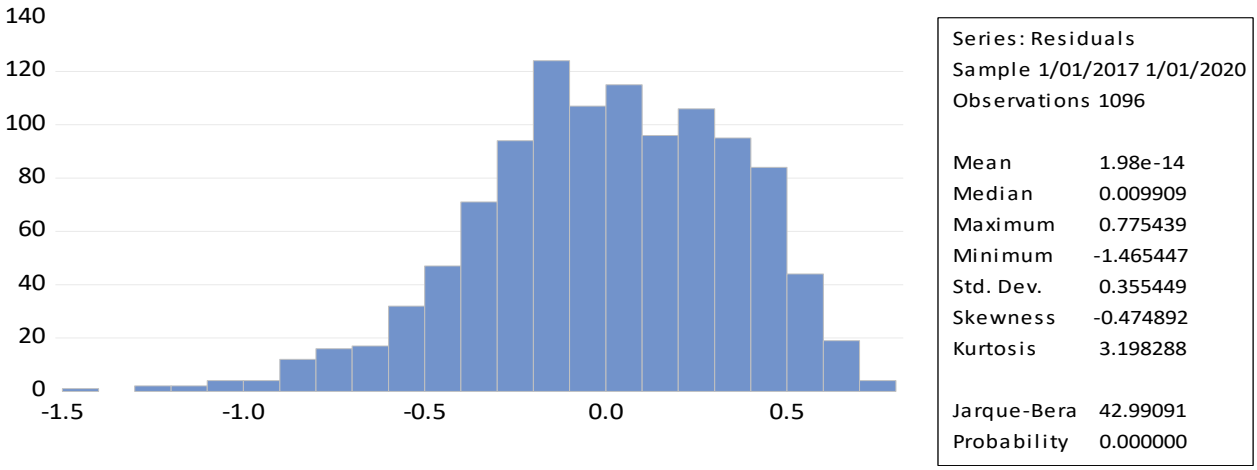

Table 5 Residuals distribution

Since the Jarque – Bera indicator is smaller than the critical value of $\chi 2$ and the probability less than 5% then H_0 stands, so the model has a normal distribution of residuals. Furthermore the

value of the indicator Kurtosis is close to the number 3 and the Skewness indicators takes value on the(-0.5;0.5) interval, which testifies to the normal distribution of residuals.

Heteroscedasticity

To assess whether the model suffers from the problem of Heteroscedasticity, the following equation for μ_i^2 is evaluated:

$$\mu_i^2 = \alpha_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 + \alpha_5 x_5$$

In addition, I will test the following hypotheses:

a) Hypothesis 0 (H0): $\alpha_2 = \alpha_3 = \alpha_4 = 0$ (The model suffers from heteroskedasticity)

b) Alternative Hypothesis (Ha): exists $\alpha_i \neq 0$ (Model does not suffer from heteroskedasticity)

To test the above hypotheses the White test on heteroskedasticity was performed.

Heteroskedasticity Test: White
Null hypothesis: Homoskedasticity

F-statistic	18.98114	Prob. F(11,1084)	0.0000
Obs*R-squared	177.0095	Prob. Chi-Square(11)	0.0000
Scaled explained SS	192.7878	Prob. Chi-Square(11)	0.0000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 02/16/21 Time: 00:02
Sample: 1/01/2017 1/01/2020
Included observations: 1096
Collinear test regressors dropped from specification

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	147.1986	72.32628	2.035203	0.0421
(LOG(TEMPERATURE))^2	0.291748	1.763326	0.165453	0.8686
(LOG(TEMPERATURE))*(LOG(WINDSPEED))	1.228585	0.611097	2.010458	0.0446
(LOG(TEMPERATURE))*(LOG(PRESSURE))	-3.701624	3.192151	-1.159602	0.2465
(LOG(TEMPERATURE))*(LOG(RELATIVEHUMIDITY))	5.171438	1.437404	3.597763	0.0003
(LOG(WINDSPEED))^2	0.064211	0.023917	2.684708	0.0074
(LOG(WINDSPEED))*(LOG(PRESSURE))	6.887457	2.320148	2.968542	0.0031
(LOG(WINDSPEED))*(LOG(RELATIVEHUMIDITY))	0.266511	0.072474	3.677330	0.0002
LOG(WINDSPEED)	-55.50620	17.71591	-3.133127	0.0018
(LOG(PRESSURE))*(LOG(RELATIVEHUMIDITY))	3.293977	4.297417	0.766502	0.4435
(LOG(RELATIVEHUMIDITY))^2	0.726141	0.111685	6.501710	0.0000
LOG(RELATIVEHUMIDITY)	-57.90949	34.50754	-1.678169	0.0936

R-squared	0.161505	Mean dependent var	0.126228

Adjusted R-squared	0.152996	S.D. dependent var	0.187240
S.E. of regression	0.172322	Akaike info criterion	-0.668017
Sum squared resid	32.18918	Schwarz criterion	-0.613279
Log likelihood	378.0734	Hannan-Quinn criter.	-0.647306
F-statistic	18.98114	Durbin-Watson stat	1.672014
Prob(F-statistic)	0.000000		

Table 6 Heteroskedasticity test

From the results of the White test, it is concluded that the probability of Fisher and $\chi 2$ is greater than 5%. This means that H_0 falls, so our model does not suffer from heteroskedasticity.

Autocorrelation

To test whether the model suffers from the problem of Serial Autocorrelation of residuals, the following hypotheses are tested:

a) Hypothesis 0 (H_0): cov (μi: μj) = 0 for i ≠ j (Model has no autocorrelation up to 2 lags)

b) Alternative Hypothesis (H_a): cov (μi: μj) ≠ 0 for i ≠ j (Model suffers from autocorrelation)

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 2 lags

F-statistic	301.9497	Prob. F(2,1089)	0.0000
Obs*R-squared	390.9705	Prob. Chi-Square(2)	0.0000

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 02/16/21 Time: 00:06
Sample: 1/01/2017 1/01/2020
Included observations: 1096
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	12.65951	12.98674	0.974803	0.3299
C(2)	-0.900308	0.475770	-1.892317	0.0587
C(3)	0.027433	0.022199	1.235805	0.2168
C(4)	-1.046200	1.718243	-0.608878	0.5427
C(5)	-0.097071	0.050152	-1.935525	0.0532
RESID(-1)	0.444814	0.029545	15.05527	0.0000
RESID(-2)	0.224260	0.029651	7.563418	0.0000

R-squared	0.356725	Mean dependent var	1.98E-14
Adjusted R-squared	0.353181	S.D. dependent var	0.355449
S.E. of regression	0.285870	Akaike info criterion	0.339806
Sum squared resid	88.99476	Schwarz criterion	0.371736
Log likelihood	-179.2136	Hannan-Quinn criter.	0.351888

F-statistic	100.6499	Durbin-Watson stat	2.059662
Prob(F-statistic)	0.000000		

Table 7 Breusch- Godfrey Correlation test

To test the above hypotheses we performed the BG (Breusch Godfrey) test on autocorrelation. From the conclusions of the BG test, it is stated that the probability of Fisher and $\chi 2$ is lower than 5%. This means that H_0 stands, so our model does not suffer from autocorrelation up to 2 model lags (time delays).

Multicollinearity

To test whether the model suffers from the problem of Multicollinearity, the following hypotheses are tested:

a) Hypothesis 0 (H0): cov (xi: xj) = 0 for i ≠ j (The model has no multicollinearity, so it has no relation through explanatory variables)

b) Alternative Hypothesis (Ha): cov (xi: xj) ≠ 0 for i ≠ j (The model has multicollinearity, so it is related through explanatory variables).

Variance Inflation Factors
Date: 02/16/21 Time: 00:03
Sample: 1/01/2017 1/01/2020
Included observations: 1096

Variable	Coefficient Variance	Uncentered VIF	Centered VIF
C(1)	261.2758	2258222.	NA
C(2)	0.347506	96562.03	2.229032
C(3)	0.000763	9.450582	1.162336
C(4)	4.578002	1868958.	1.269775
C(5)	0.003863	555.6278	2.301999

Table 8 VIF multicollinearity

To test the above hypotheses we performed the Variance Inflation Factors (VIF) test on multicollinearity. From the conduction of the VIF test, it is concluded that the VIF values for our three variables are less than 10, so we can state that the evaluated model does not suffer from multicollinearity.

On the previous sub – chapter I conducted the econometric tests that are suggested to be done while testing the Gauss – Markow Hypotheses. All the five hypotheses have been met, therefore the estimated econometric model is statistically significant and also valid for further analysis.

3.6 Comparison of algorithms

3.6.1 ARIMA test

As mentioned above the regression model resulted to be statistically significant and can be used for further analysis. Considering this fact, we may proceed by performing forecasting methods in order to compare the prediction of the radiation.

So before using our algorithm we have to split our data into train data and test data so to do this we use the following code in python:

```
y = dataset["SolarRadiation"]

dataset = dataset.drop(columns="SolarRadiation")

X_train,X_test,y_train,y_test=train_test_split(dataset,y,test_si
ze = 0.2,random_state=0)
```

```
X_train samples :  (876, 4)
y_train samples :  (876,)
X_test samples :  (220, 4)
y_test samples :  (220,)
```

Figure 24 Dataset split for training

From the figure above we can see that our dataset is separated and ready to be used for algorithms.

The first forecasting algorithm will be a static forecasting model based on the ARIMA approach. The overall ARIMA equation structure is as following:

$$\Delta^D y_t = c + \phi_1 \Delta^D y_{t-1} + \ldots + \phi_p \Delta^D y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}.$$

Equation 20 ARIMA equation

Where $\Delta^D y_t$ denotes a D-th differenced series, and ε_t is an uncorrelated process with mean zero. In lag operator notation, $Liyt=yt-i$. The ARIMA (p,D,q) model can be written as:

$$\phi^*(L) y_t = \phi(L)(1-L)^D y_t = c + \theta(L)\varepsilon_t$$

ARIMA stands for Auto – Regressive Integrated Moving Average. The model is defined and analyzed in three different parts which are complementary to each – other: the autoregression which allows us to predict by taking into the consideration the historical values of the series,

the integration which determines if the series shows any trend feature over time and the moving average which is a used to determine the degree of forecasting values by taking into account even the seasonal effect.

To test this in python I used source code from the Kaggle.

So to in order to get the q (number of auto regressor AR) and p(number of moving average MA) values we use following code:

```
lag_acf = acf(ts_dif,nlags=20) #ACF

lag_pacf = pacf(ts_dif, nlags=20, method='ols') #PACF

fig,ax = plt.subplots(1,2, figsize=(14,4))

ax1, ax2 = ax.flatten()

ax1.plot(lag_acf)

ax1.axhline(y=0,linestyle='--',color= 'black')

ax1.axhline(y=-1.96/np.sqrt(len(ts_)),linestyle='--',color=
'gray')

ax1.axhline(y=1.96/np.sqrt(len(ts_)),linestyle='--',color=
'gray')

ax2.plot(lag_pacf,)

ax2.axhline(y=0,linestyle = '--', color = 'gray')

ax2.axhline(y=-1.96/np.sqrt(len(ts_)),linestyle='--
',color='gray')

ax2.axhline(y=1.96/np.sqrt(len(ts_)),linestyle='--',    color    =
'gray')
```

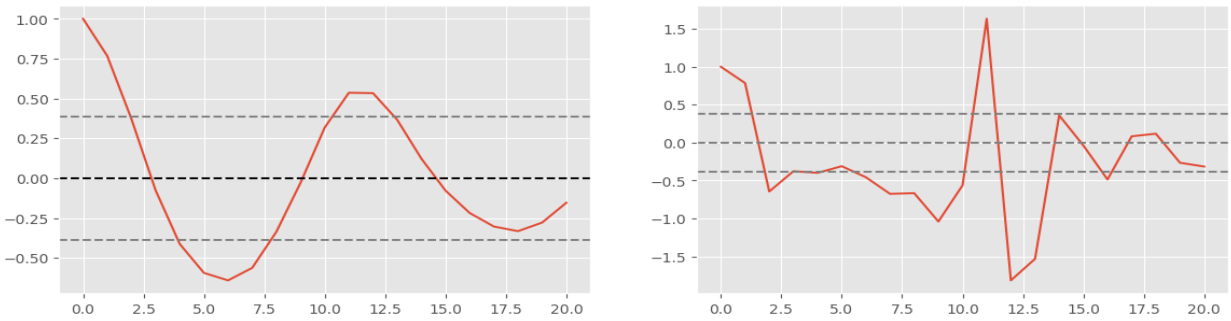

Figure 25. Plots of ACF and PACF

From these plots we can get the values of q and p. were number of lags used is 20.

P value is gathered from the lag value PACF chart when red line crosses upper chart for first time and Q value is gathered similarly like P value it is gathered from the lag value of ACF chart when red line crosses upper chart for first time. Here we can see that p = 3 and q = 1,

As for the results on the above plot, the suggested series is the ARIMA (3,1,1). With this into account, we may proceed by forecasting the Solar Radiation values for the period of 1/1/2020 – 1/31/2020 in Tirana.

```
print(model_fit.summary())
```

ARIMA Model Results

Dep. Variable:	D.SolarRadiation	No. Observations:	1084
Model:	ARIMA(3, 1, 1)	Log Likelihood	-328.038
Method:	css-mle	S.D. of innovations	0.327
Date:	Wed, 17 Feb 2021	AIC	668.075
Time:	20:40:04	BIC	698.006
Sample:	01-13-2017 - 01-01-2020	HQIC	679.406

	coef	std err	z	P>|z|	[0.025	0.975]
const	3.823e-05	5.07e-05	0.754	0.451	-6.12e-05	0.000
ar.L1.D.SolarRadiation	0.3522	0.030	11.602	0.000	0.293	0.412
ar.L2.D.SolarRadiation	-0.0165	0.032	-0.514	0.608	-0.080	0.047
ar.L3.D.SolarRadiation	0.0417	0.030	1.370	0.171	-0.018	0.101
ma.L1.D.SolarRadiation	-0.9999	0.003	-379.891	0.000	-1.005	-0.995

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	2.0358	-0.0000j	2.0358	-0.0000
AR.2	-0.8195	-3.3341j	3.4333	-0.2884
AR.3	-0.8195	+3.3341j	3.4333	0.2884
MA.1	1.0001	+0.0000j	1.0001	0.0000

Table 9 Arima Model Results

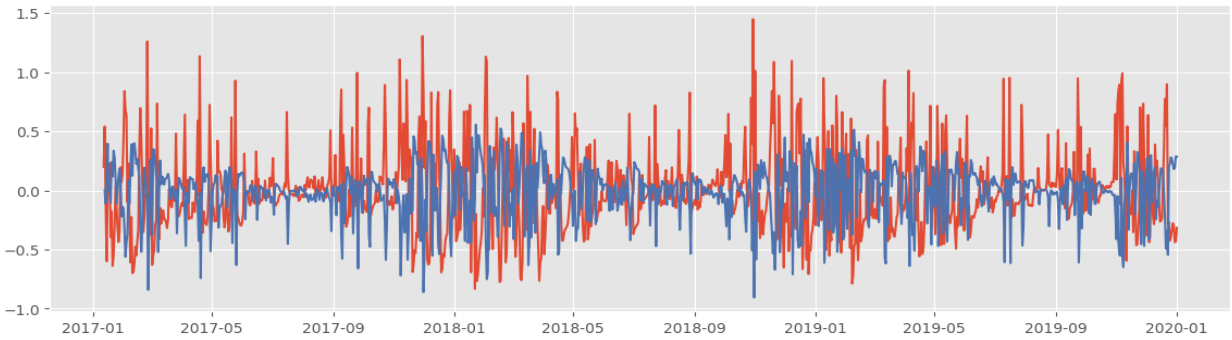

Table 10 Plot ARIMA model prediction

Red line is difference of mean of log(input values) minus log (input values), meanwhile blue line is the mean of predicted values.

I order to check the accuracy of the model I performed the following commands:

```
r2 = model.score(X_test, y_test)

lin_mae = mean_absolute_error(y_pred, y_test)

lin_mse = mean_squared_error(y_test, y_pred )

lin_rmse = np.sqrt(lin_mse)
```

	Model	R2	Mae	Mse	Rmse
0	ARIMA(3,1,1)	0.92379	39.163258	5137.345976	71.675281

Table 11 ARIMA stats.

The results of the above commands are as above from here we can see R^2= 92.38%, MAE = 39.163258(Wh/m^2), MSE= 5137.3459(Wh/m^2) and RMSE = 71.67528 (Wh/m^2). From these results we can say that ARIMA performed very well having an accuary of 92% and RMSE of 71.67(Wh/m^2).

For further information lets have a look to the predicted values of this algorithm.

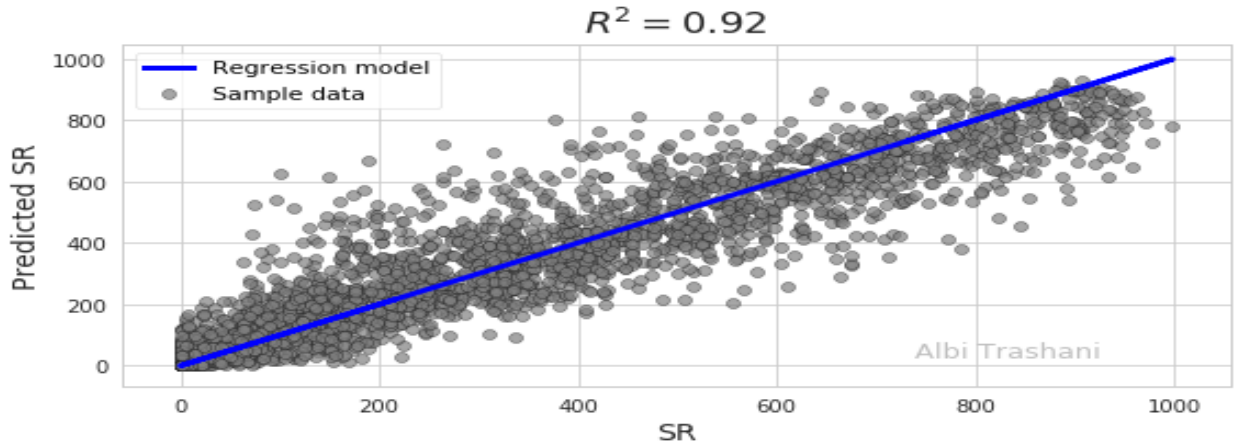

Table 12 Plot between actual and predicted SR

From this plot it can be easily seen how was the test accuracy. Predicted values are very near regression line.

Similarly, all procedures are followed for other models.

Other algorithms that I used are Linear Regression Elastic Net, Lasso & Ridge, Random Forest Regression, Gradient Boosting, Decision Tree and K-Neighbors Regressor. The obtained results can be seen below sorted by R^2 value.

I just used the premade commands with the respective libraries and tested them:

```
models =models.append( KNeighborsRegressor();LinearRegression();
DecisionTreeRegressor(); RandomForestRegressor();ElasticNet(alpha
= 0.01); Ridge(alpha = 5); Lasso(alpha = 0.5))
```

Model	R2	Mae	Mse	Rmse
Random Forest	0.929613	36.456472	4725.568316	68.742769
ARIMA(3,1,1)	0.923790	39.163258	5137.345976	71.675281
Gradient Boosting	0.904407	49.209308	6417.832718	80.111377
Decision Tree	0.859656	48.231043	9422.294746	97.068505
KNN	0.830719	61.086442	11365.053021	106.607003
Linear Regression	0.556109	130.581872	29801.624379	172.631470
Elastic Net	0.555686	130.701332	29830.053625	172.713791
Ridge	0.555686	130.617843	29809.121321	172.653182
Lasso	0.555686	130.610507	29807.347244	172.648044

Table 13 Models Accuracy sorted

After being tested all models that we choose they show that they all have good statistical values. The best model from these is the Random Forest with an R^2 value of 93% and RMSE=68.74 (Wh/m^2).

Prediction of future SR values

As a last step of this thesis will be the prediction of Solar Radiation for next 30 days. We need to keep in mind that the later the forecasted values are, the highest is the standard error for the forecasted values.

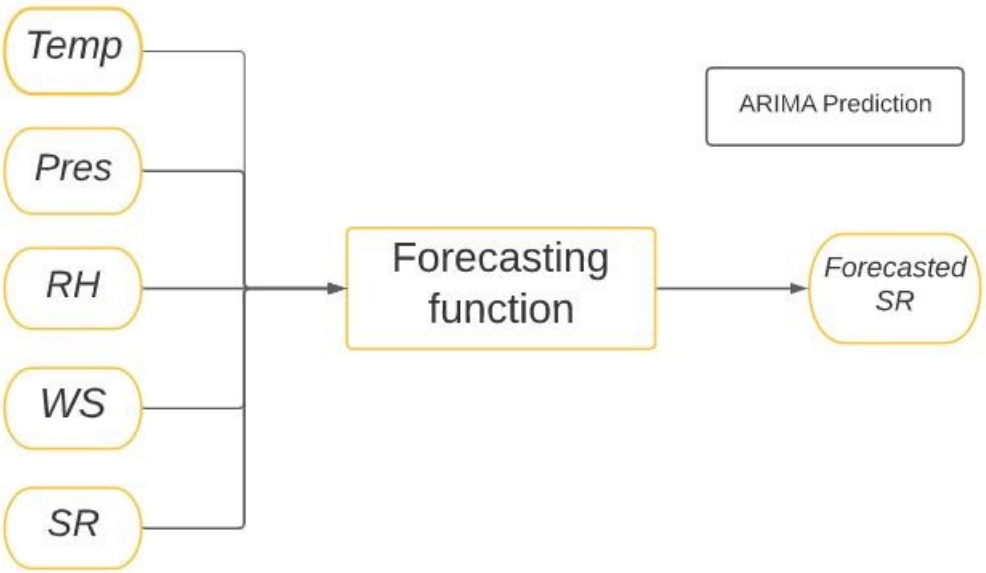

Figure 26 Network for the prediction of values

The code takes as input only past values of Solar radiation and predict the future values.

The network shown in figure 44 is the next step value prediction for the desired period of time. Firstly, I take as input values previous data. And it outputs the forecasted values for "h" time. Which in my case 'h' is 30 since we want prediction for the next 30 following days.

After compiling the code, I got the following results:

Index	Date	SR (Forecasted)
0	1/1/2020	278.37
1	1/2/2020	187.75
2	1/3/2020	282.42
3	1/4/2020	281.99
4	1/5/2020	282.04
5	1/6/2020	281.87
6	1/7/2020	266.82
7	1/8/2020	283.15
8	1/9/2020	267.11
9	1/10/2020	283.41
10	1/11/2020	282.63
11	1/12/2020	279.68
12	1/13/2020	282.18
13	1/14/2020	276.99
14	1/15/2020	281.93
15	1/16/2020	278.37
16	1/17/2020	265.09
17	1/18/2020	281.21
18	1/19/2020	280.2
19	1/20/2020	281.39
20	1/21/2020	264.65
21	1/22/2020	276.15
22	1/23/2020	281.05
23	1/24/2020	276.3
24	1/25/2020	281.94
25	1/26/2020	261.34
26	1/27/2020	281.52
27	1/28/2020	276.75
28	1/29/2020	279.49
29	1/30/2020	281.84
30	1/31/2020	274.06

Table 14 SR Forecasted

After getting these values lets plot them comparing with last year same date value.

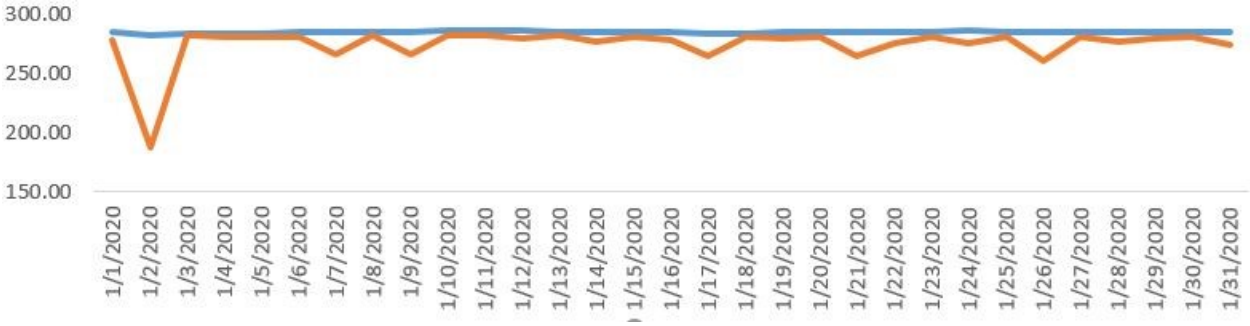

Figure 27 Last year vs predicted

From this plot we can see with orange color the previous year's values and with blue the forecasted values.

Judging on a retrospective analysis, the forecasted values have the same trend compared to the same period on the previous years. This effect is attributed to the selection of the ARIMA model with the seasonal effects (also called S – ARIMA) which reflects the seasonal factors and coefficients even in the forecasted values.

In order to check if the technique has provided an accurate result we should compare them with the actual data of solar radiation and calculate the deviations. The lower the level of deviation is more accurate is the data.

Chapter 4

Conclusions

4.1 Results

Electricity can be produced in different ways and methods, one of those ways is by using Solar power. It produces renewable energy, and which does not discharge any type of harmful gases. In this study as I have repeatedly mentioned I have taken a detailed look of solar radiation in Tirana. For the first time a prediction like this one is actually performed for Tirana. Even though with very little information and data, I managed to produce valuable data for this city. If more information was available, if any other study would have been conducted for solar radiation of Tirana the more accurate and meaningful our study would be. The conclusion below will give you a clearer view of what this paper achieved. I have used machine learning algorithms to measure and predict the solar radiation daily in this city. The study I have conducted first checks the model using the Ordinary Least Squares Method (OLS). The model contains the measured average daily values of wind, pressure, humidity and temperature. After testing it using E – views I got the following assumptions. All the dependent variables are statistically significant. The coefficients of temperature and pressure are both positive which means that an increase on those coefficients will result on an increase of the solar radiation level. Meanwhile, the coefficients of wind speed and relative humidity are both negative, which means that an increase on those coefficients will result on an decrease of the solar radiation level. By using Jarque – Bera we saw the model has a normal distribution of residuals. Furthermore, the value of the indicator Kurtosis is close to the number 3 and the Skewness indicators takes value on the (-0.5;0.5) interval, which testifies to the normal distribution of residuals. To check if the model suffers from heteroscedasticity the White test was performed. From the results of the White test, it is concluded that the probability of F-statistic and $\chi 2$ is less than 5%. This means that our model does not suffer from heteroscedasticity. In order to see if there exist an autocorrelation, I performed the BG test and I saw that probability of F-statistic and $\chi 2$ is lower than 5%. Proving that our model does not suffer from autocorrelation up to 2 model lags (time delays). After performing these econometric tests. I we start with the second part of the conclusions for this the thesis.

As I mentioned in above sections there are many ways and methods to predict solar radiation some and there it will be always better or worse methods. After testing our model for this thesis it can be said that ARIMA along with Random Forest and gradient boosting method performed very well with highest score of R^2 and RMSA equal to 92% and 68.74(Wh/m^2)

Lastly using the ARIMA for forecasting future values there can be seen a stability of the trend during the forecasted period (1/1/2020 – 1/31/2020) with minor deviation from day to day.

4.2 Limitations and Future Work

There are a few limitations for this thesis but, I would mention one of them which is the issue of the Solar Radiation values. The reason behind this is that solar radiation measurement sensors are very expensive and for Albania Solar radiation values are gathered by addition of the BHI (Beam irradiation on horizontal plane at ground level) and DHI. (Diffuse irradiation on horizontal plane at ground level). Even though this measurement is still correct it would be better to have the original measurement of values as with a better measured value the prediction would be much more accurate.

Since there are so many other algorithms, I would suggest for future work to try and work on other algorithms and compare the result in order to see which one would perform better. Also, another interesting study it would be if we add factors to models.

References

[1] F. Beshart, A. A. Dehghan and A. R. Faghih, "Empirical models for estimating global solar radiation: A review and case study," *Renewable and Sustainable Energy Reviews,* pp. 798-821, 2013.

[2] T. R. Ayodele, A. S. Ogunjuyigbe and C. G. Monyei, "On the global solar radiation prediction methods," *JOURNAL OF RENEWABLE AND SUSTAINABLE ENERGY 8,* pp. 1-21, 2016.

[3] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solara Energy,* pp. 342-349, 2009.

[4] S. Ferrari, M. Lazzaroni, V. Piuri, L. Cristaldi and M. Faifer, "Statistical Models approach for Solar Radiation Prediction," in *International Instrumentation and Measurement Technology Conference*, 2013.

[5] M. Paulescu, O. Mares, E. Paulescu, N. Stefu, A. Pacurar and D. Calinoiu, "Nowcasting solar irradiance using the sunshine number," *Energy Conversion and Management,* pp. 690-697, 2013.

[6] A. Ogunjuyingbe and T. R. Ayodele, "Prediction of monthly average global solar radiation based on statistical distribution of clearness index ," *Energy,* pp. 1733-1742, 2015.

[7] D. Yang, V. Sharma, Z. Ye, I. Lihong, L. Zhao and A. W. Aryaputera, "Forecasting of global horizontal irradiance by exponential smoothing, using decompositions," *Energy,* pp. 111-119, 2015.

[8] G. E. Hinton, N. Srivastava and K. Swersky , "Neural Networks from Machine Learning".

[9] M. I. Jordan and T. M. Mitchel, "Trends, Prespectives and prospects," *Machine Learning ,* pp. 345,255, 2015.

[10] M. Cocea and H. Liu, "Studies in Big Data," *Granular Computing Based Machine Learning,* p. 113, 2018.

[11] A. Vogelsang and M. Borg, "Requirements Engineering for Machine Learning," 2019.

[12] T. M. Mitchell, "The Discipline of Machine Learning," 2006.

[13] M. Sugiyama, " Introduction to Statistical Machine Learning," 2016.

[14] A. Wilson, "A Brief Introduction to Supervised Learning," 2019.

[15] M. R. M. Talabis, "Information Security Analytics," 2015.

[16] D. C. MONTGOMERY, E. A. PECK and G. G. VINING, Introduction to Linear Regression Analysis, 2012.

[17] S. R. Gunn, Support Vector Machines for Classification and Regression, 1998.

[18] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3 (1972), pp..*

[19] G. K. Smyth, Nonlinear regression, John Wiley & Sons, Ltd, Chichester, 2002.

[20] W. S. Noble, "Support vector machines," in *NATURE BIOTECHNOLOGY*, 2006.

[21] Elsevier and S. Kalogirou, "Machine Learning methods for solar radiation forecasting," *Renewable Energy, an international journal ,* 1991.

[22] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification," 2007.

[23] F. Livingston, "Implementation of Breiman's Random Forest Machine Learning," *Machine Learning Journal Paper,* 2005.

[24] O. Harrison, "Data & Machine Learning," 2018.

[25] A. Tamir, Applications of Markov Chains in Chemical Engineering, 1998.

[26] J. A. H. Assoc., "State-of-the-Art Machine Learning Techniques," 2020.

[27] S. Kalogirou, "Renewable Energy," *Machine learning methods for solar radiation forecasting: A review,* 2017.

[28] M. S. T. Hassani, "Using internal evaluation measures to validate the quality of diverse stream clustering algorithms.," *Vietnam J Comput Sci 4,* pp. 171-183, 2017.

[29] A. Dasgupta and A. Nath, "Classification of Machine Learning Algorithms," *International Journal of Innovative Research in Advanced Engineering (IJIRAE),* 2016.

[30] R. Polikar, "Ensemble Learning," 2009.

[31] G. Brown, "Diversity in Neural Network Ensembles," 2004.

[32] Z.-H. Zhou, Ensemble Methods: Foundations and Algorithms, 2012.

[33] R. E. Schapire, "A Brief Introduction to Boosting," *Proceedings of the Sixteenth International Joint,* 1999.

[34] Z. Zhang, "Diversity in Neural Network Ensembles, Theory, Implementation, and Visualization," 2019.

[35] A. Buja and W. Stuetzle, "OBSERVATIONS ON BAGGING," *Statistica Sinica 16,* pp. 323-351, 2006.

[36] M. Sewell, "Ensemble Learning," 2011.

[37] A. &. S. K. Ludwig, "Calendar year rreserves in the Multivariate Additive Model", 2010.

[38] C. Dougherty, "Introduction to Econometrics," (3th ed.), Oxford Press, 2007.

[39] C. Brooks, "Introductory Econometrics for Finance," (2nd ed.) Cambridge University Press, Cambridge, 2008.

[40] D. Gujarati, "Basic Econometrics" (4th ed.)., New York: The McGray - Hill Companies, 2004.