**PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA**
**Ministry of Higher Education and Scientific Research**

**UNIVERSITY ECHAHID HAMMA LAKHDAR - EL OUED**
**FACULTY OF EXACT SCIENCES**
**Computer Science department**

**End of study memory**
**Presented for the Diploma of**

# ACADEMIC MASTER

Domain: **Mathematics and Computer Science**
Industry: **Computer Science**
Specialty: **Distributed Systems and Artificial Intelligence**

***Theme***

---

# An Automatic Prediction of Solar Radiation for Renewable Energy using Machine Learning models.

**Presented by:**
Bilal Blouzi
Noufle Mohammed Hadi Lalmi

*Sustained on 16-June-2022 from the jury:*

| | | |
|---|---|---|
| Miss. | Nedjoua Houda KHOLLADI MA (...) | Supervisor Univ. El Oued |
| Mr. | Abed Elhamid NDIOUI MA (...) | President Univ. El Oued |
| Mr. | Mohamed Kamal BENBRAIKA MA (...) | Reporter Univ. El Oued |

Academic year
2021/2022

# شكر وتقدير

مع خالص امتناننا نود أن نعبر لمشرفتنا الأستاذة خلادي نجوى بشكر خاص على ثقتها وصبرها ودعمها وعلى تخصيص الوقت اللازم لمتابعة عملنا وعلى إرشادنا لإعداد هذه المذكرة ، نحن ممتنون لها، استفدنا من كفاءتها الكبيرة ، وصرامتها الفكرية ، ونصائحها التي لا تقدر بثمن وأدعو الله أن يرزقها الصحة والسعادة و العافية وطول العمر .

نود أيضًا أن نتقدم بالشكر و التقدير لأعضاء مجلس إدارة شركة الكهرباء والطاقات المتجددة بالحجيرة (SKTM) وشكر خاص للسيد ربوح صالح على ماقدمه لنا من مساعدات وإستقبال ونسأل الله العلي القدير أن يوفقكم الى مايحبه ويرضاه و نتمنى لكم التوفيق و مواصلة النجاح والتميز .

# إهداء

نهدي هذا العمل المتواضع للوالدين لتشجيعهم لنا و لتضحياتهم كشهادة على محبتنا لهم طوال سنوات الدراسية الماضية منذ السنة الدراسية الأولى الى يومنا هذا . الى كل العائلة الكريمة من بنات وذكور الى كل اصدقائنا و لجميع الأساتذة والدكاتره في قسم إعلام الألي خاصة وكلية العلوم الدقيقة بجامعة الوادي عامة.

# Abstract

***As a result of the substantial growth and development of renewable sources of energy, production sources have varied and the network has become more difficult to manage. Therefore, predicting the electricity generated by renewable sources has become critical. From this point of view, machine learning, which is part of AI, seems to be one of the best ways to achieve this goal. Machine learning techniques are capable of controlling the variations in renewable energy output and, therefore, facilitate their integration into the energy mix. Thus, one of the major goals of this research is to perform a comprehensive comparison of three prominent machine learning techniques, including support vector regression, linear regression, and random forest, for the short-term prediction of the solar radiation that causes the power produced by photovoltaic solar panels. The dataset we used in this study represents data from the year 2021 and is related to El Hadjira, which is a semi-desert climate province in Algeria. In the testing phase, the results showed that random forest was the most accurate prediction method, with*** $R^2$***=0.334 and RMSE=238.08*** $W/m^2$.

***Key words: Solar radiation prediction, Forecasting solar power, Machine learning , Linear regression, Random forest, Support vector regression, Artificial intelligence.***

# Résumé

Du fait de la forte croissance et du développement des énergies renouvelables, les sources de production se sont diversifiées et le réseau est devenu plus difficile à gérer. Par conséquent, la prévision de l'électricité produite par des sources renouvelables est devenue critique. De ce point de vue, l'apprentissage automatique, qui fait partie de l'IA, semble être l'un des meilleurs moyens d'atteindre cet objectif. Les techniques de machine learning sont capables de contrôler les variations de production d'énergies renouvelables et donc de faciliter leur intégration dans le mix énergétique. Ainsi, l'un des principaux objectifs de cette recherche est d'effectuer une comparaison complète de trois techniques d'apprentissage automatique de premier plan, y compris la régression vectorielle de support, la régression linéaire et la forêt aléatoire, pour la prédiction à court terme du rayonnement solaire qui provoque l'énergie produite. par des panneaux solaires photovoltaïques. L'ensemble de données que nous avons utilisé dans cette étude représente les données de l'année 2021 et est lié à El Hadjira, qui est une province climatique semi-désertique en Algérie. Lors de la phase de test, les résultats ont montré que la forêt aléatoire était la méthode de prédiction la plus précise, avec $\mathbf{R^2}$ = 0,334 et **RMSE** = 238,08 W/m².

**Mots-clés :** Prédiction du rayonnement solaire, Prévision de l'énergie solaire, Apprentissage automatique, Régression linéaire, Forêt aléatoire, Régression vectorielle de support, Intelligence artificielle.

# الملخص

نتيجة للنمو الكبير وتطوير مصادر الطاقة المتجددة ، تباينت مصادر الإنتاج وأصبحت الشبكة أكثر صعوبة في إدارتها. لذلك ، أصبح التنبؤ بالكهرباء المولدة من المصادر المتجددة أمرًا بالغ الأهمية. من وجهة النظر هذه ، يبدو أن التعلم الآلي ، وهو جزء من الذكاء الاصطناعي ، هو أحد أفضل الطرق لتحقيق هذا الهدف. تقنيات التعلم الآلي قادرة على التحكم في الاختلافات في إنتاج الطاقة المتجددة ، وبالتالي ، تسهل دمجها في مزيج الطاقة. وبالتالي ، فإن أحد الأهداف الرئيسية لهذا البحث هو إجراء مقارنة شاملة لثلاث تقنيات بارزة للتعلم الآلي ، بما في ذلك انحدار ناقل الدعم ، والانحدار الخطي ، والغابات العشوائية ، للتنبؤ قصير المدى بالإشعاع الشمسي الذي يسبب الطاقة المنتجة بواسطة الألواح الشمسية الكهروضوئية. تمثل مجموعة البيانات التي استخدمناها في هذه الدراسة بيانات من عام 2021 وترتبط بالحجيرة ، وهي إقليم مناخ شبه صحراوي في الجزائر. في مرحلة الاختبار ، أظهرت النتائج أن غابة عشوائية كانت الطريقة الأكثر دقة للتنبؤ ، حيث

( RMSE = 238.08 واط/ م$^2$ ) و $R^2 = 0.334$

الكلمات الرئيسية: التنبؤ بالإشعاع الشمسي ، التنبؤ بالطاقة الشمسية ، التعلم الآلي ، الانحدار الخطي ، الغابة العشوائية ، دعم الانحدار المتجه ، الذكاء الاصطناعي.

# Contents

# List of Figures

# List of Tables

# *Introduction*

# General introduction

Renewable Energy is becoming a technology and an ever more viable alternative to traditional nonrenewable energy sources. Therefore, the relentless danger of climate change forces humanity to strive for new and more effective ways to produce energy. Particularly when we base all our basic needs on this form of energy.

Global renewable electricity capacity is projected to rise by over 1 TW, a 46 percent increase over the period 2018 to 2023 Solar photovoltaic (PV) represents more than half of this expansion and dominates the growth of renewable ability. However, because the energy output of PV panels depends on weather conditions such as cloud cover and solar irradiance, the PV panels' energy output is unstable.

In today's world, the use of technologies that utilize artificial intelligence is expanding at an ever-increasing rate owing to its capacity to resolve issues that are exceptionally difficult. Machine learning is a computer science subfield, and it is categorized as a form of artificial intelligence. It can be used in many domains, and the advantage of this approach is that a model can solve problems that clear algorithms cannot represent. The problem of solar radiation The variability and unpredictability which reach the surface of the Earth are well known.

Because of this, having an accurate forecast of this variable will allow for better planning and operation of power delivery at either the economic level or at the level of energy output. This can be accomplished by either making alternative arrangements for traditional power and overall timetables or by investing the appropriate amount of energy resources and reserves in order to minimize the operating costs of the power system.

This topic will present an approach for predicting solar radiation based on machine learning techniques. SVR, LR, and RF have been used in studies. The relevance of the studied models was evaluated for short-term solar radiation forecasting to ensure optimal management and security requirements in this field while using an integral solution based on a single tool and an appropriate predictive model.

The dataset that will be used in this study represent data over a period of time and are related to the arid (desert, semi-desert) climate province in Algeria. Also we used a USA dataset Related to Nevada region.

**This report will contain four chapters**

- **Chapter 1:** Will contain a background of the subject and Solar power systems.
- **Chapter 2:** Will contain other similar works and their evaluation.
- **Chapter 3:** Contain a description of the dataset and the tools that were used in this project.
- **Chapter 4:**The last chapter will describe the implementation of the project and the obtained results.

**Finally the general conclusion.**

# Chapter 1

# An Overview for Solar Power Systems

## 1.1 Introduction

As the world progresses, energy needs grow. We have more electronic devices that consume electricity, we travel to more places and more distant, more sophisticated goods are manufactured, requiring more energy in the factories that produce them.

Devices that have traditionally worked with other energy sources, such as vehicles (that worked with oil), bicycles (that traditionally had pedals), kitchens (traditionally powered with gas), and so on, are now electrifying themselves.

## 1.2 Energy

**Energy** is a strategic resource and countries seek to be energy independent since an energy dependent country has to pay a large bill for the energy consumed and therefore its development is slowed considerably . Integrating energy generation in factories and urban environments is appealing because it avoids wasting energy in transportation infrastructure while also saving money. Renewable energies are often the simplest to integrate into these environments because they require less infrastructure and produce less noise, dirt, pollution, and other issues.

Traditional energy sources have serious drawbacks: they are finite, pollute the environment, and will eventually run out (fossil fuels), forcing humanity to seek out clean and renewable energy sources. At the same time as scientific and technological advancements progress, society is becoming more aware of its environmental responsibilities. This is one of the main reasons why renewable energy is booming right now, posing numerous challenges to the scientific and technical community.

Renewable energy is clean, endless, and its technology is rapidly evolving thanks to initiatives such as the Kyoto Protocol, a global commitment to combat global climate change. Wind power, hydro-power, solar power, biomass, bio-fuel, and other renewable energy sources are examples. Solar energy, specifically photovoltaic solar energy, is becoming increasingly important in the energy mix in this type of situation.[1]

## 1.3 Renewable Energy

Renewable technologies are considered clean energy sources, and their optimal use minimizes environmental impacts, produces minimal secondary waste, and is sustainable in light of current and future economic and social societal needs. The sun is the source of all energy. Solar energy is divided into two types: heat and light. Sunlight and heat are transformed and absorbed in a number of ways by the environment.

Some of these processes result in renewable energy flows like biomass and wind energy. By replacing renewable energy sources with traditional energy sources, renewable energy technologies provide a great chance to reduce greenhouse gas emissions and global warming [2]. In Fig 1.1, we can see the different types of renewable energy sources in the nature.

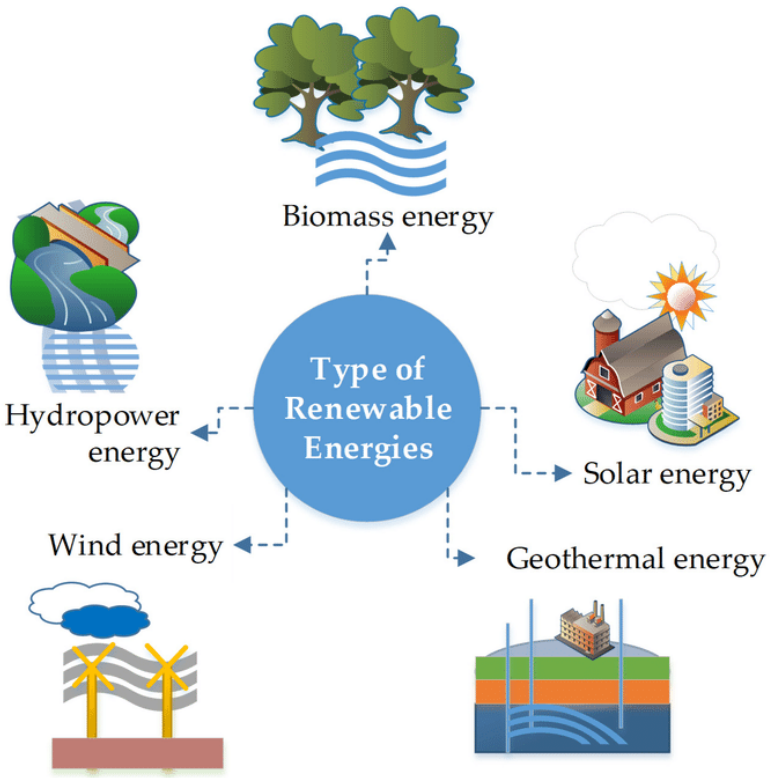



Figure 1.1: *Types of Renewable Energy [3]*

## 1.4 Solar Power

Our sun is 333,000 times the size of our planet when viewed from 93 million miles away. It has an 865,000-mile diameter, a surface temperature of 5,600 ° C, and a core temperature of 15,000,000 ° C. It is a huge mass of constant nuclear activity.Our sun gives us all the energy we need to live and supports all life forms, whether directly or indirectly. Our climate and weather are dictated by the sun. Our world would be a frozen wasteland of ice-covered rock if it were not for it. Solar power is a fantastic concept.

It's a fantastic idea to integrate the sun's energy and use it to power electrical equipment. There are no ongoing electricity bills, no reliance on a power outlet: a completely free and infinite source of energy that does not harm the environment!.Of course, the truth is a little more complicated. However, generating electricity purely from sunlight is a valuable resource with applications and benefits all over the world.[4]

### 1.4.1 The Source of Solar Power

Intense nuclear activity deep in the sun's core produces massive quantities of radiation. As a result of this radiation, photons, or light energy, are produced. These photons have no physical mass of their own, yet they are extremely energetic and have a lot of velocities.

Light wavelengths are carried by distinct photons. Non-visible light (infrared and ultra-violet) will be carried by some photons, while visible light will be carried by others (white light). These photons push outwards from the sun's core throughout time. A photon can take a million years to get from the core to the surface. These photons travel at a speed of 670 million miles per hour once they reach the sun's surface. It takes them around eight minutes to reach Earth. Photons can collide with and be deflected by other particles on their way from the sun to the earth, and they are destroyed when they come into touch with anything that can absorb radiation, causing heat. On a bright day, this is why you feel warm: your body is absorbing photons from the sun. Many of these photons are absorbed by our atmosphere before they reach the earth's surface. One of the two reasons the sun feels so much hotter in the middle of the day is because of this.When the sun is directly above, photons must travel through a thinner layer of atmosphere to reach us, as opposed to when the sun is sinking when photons must travel through a much wider layer of atmosphere to reach us.

This is also one of the two reasons why a sunny day in winter is so much colder than a sunny day in summer. In winter, when your location on the earth is tilted away from the sun, the photons have to travel through a much thicker layer of atmosphere to reach us.[4]

### 1.4.2 The Nature of Light Energy

Light is a form of energy. Since the sun's light is made up of many various colors that when combined generate white light, it appears white. Each of the sun's visible and invisible radiations has a different energy level. Red is the low-energy end of the visible spectrum (red to violet), whereas violet is the high-energy end. The energy of light in the infrared region is lower than that of visible region. The ultraviolet region of light has more energy than the visible zone.

Visible light is merely a small part of the wide spectrum of radiation. The interaction of one light ray with another or other physical things may frequently be interpreted as if light were traveling as a wave, according to studies of light and comparable radiation. There is a set space between the tops of all waves (called the wavelength). A frequency can be used to represent this wavelength. The wavelength and frequency have an inverse relationship. As the frequency of light waves rises, the energy associated with the wave rises as well (wavelength decreases). Red light has a wavelength of about $3\times10-24\ kW-h$ per photon and violet has $4.5 \times 10 - 24\ kW - h$ [5].

## 1.5 Photovoltaic System

The photovoltaic system consists of a field of modules and a set of components that adapt the electricity produced by the modules to the specifications of receptors [6].

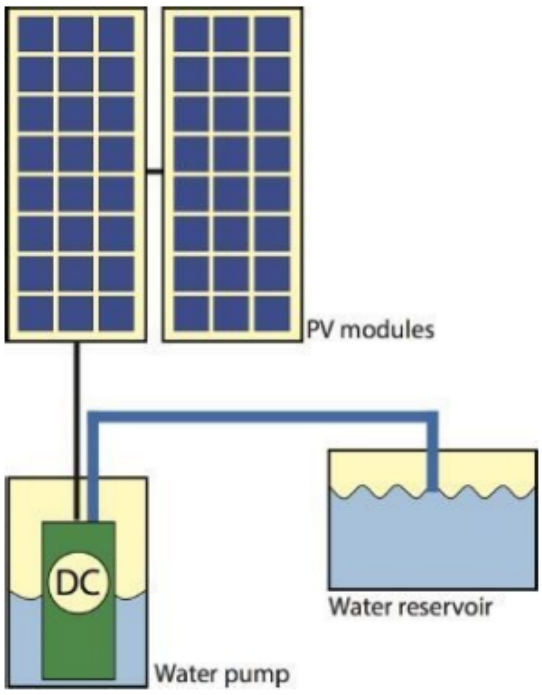



Figure 1.2: *Schematic of a simple DC PV system to power a water pump.*

### 1.5.1 Photovoltaic Cell

A PV cell is a non-mechanical device that turns sunlight directly into electricity. There are no mechanical moving parts in the system, unlike other power plants such as hydroelectric power plants, steam power plants, thermal power plants, and nuclear power plants. A PV cell is a semiconductor diode with a sun-exposed p-n junction. Photons of solar energy make up sunlight. The different wavelengths of the sun spectrum correlate to different quantities of energy in these photons. When photons reach a PV cell, they have three options: they can reflect off of it, travel through it, or be absorbed by the semiconductor material. Photons that have been absorbed are responsible for generating energy. Electrons are released from the atoms of semiconductor materials when they absorb enough sunlight.

Only photons with energies greater than the PV cell's bandgap are effective for generating electricity; the rest of the energy is lost as heat in the PV cell.[7]

Figure 1.3: *Photovoltaic Panel [***wiki***]*

### 1.5.2 Storage Battery

The fact that solar energy is not available for operation of the powered system requires the use of batteries in the installations autonomous to store energy. In autonomous solar systems:

- **Lead-acid batteries:** They constitute the overwhelming majority of the accumulators. Its good technological mastery, its low cost, its good load/discharge efficiency. Its operating conditions are not difficult to satisfy militate in favor of its wide use.
- **Nickel Cadmium Batteries:** They are the most expensive, but also very resistant to overloads and discharges, and resists well to low temperatures [8].

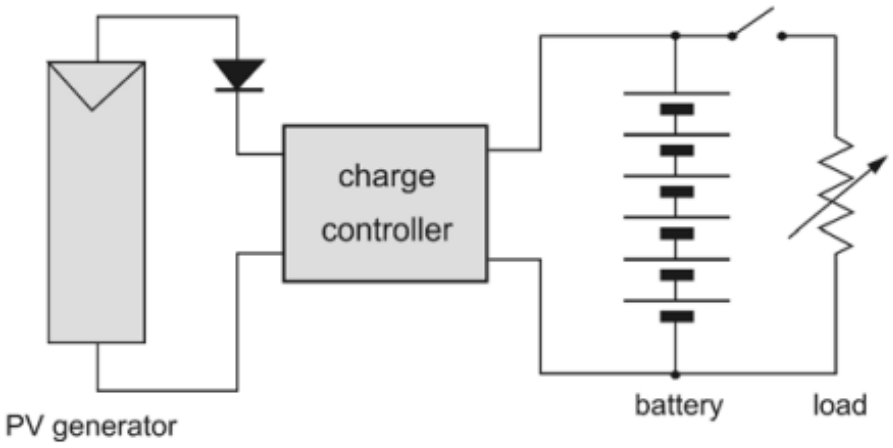



Figure 1.4: *Scheme of a PV system with a chemical battery storage.*

### 1.5.3 Regulators

In any autonomous photovoltaic system, a so-called regulation system, which is used to control the current going through the accumulators, protecting them against overloads

and deep discharges, to maximize its service life. The regulator also allows an optimal transfer of energy from the field photovoltaic to use [8].

### 1.5.4 Inverters

To supply AC equipment, a device Static conversion electronics or DC/AC converter is used for the transformation AC direct current [8]. In Fig 1.5, it simplify that process.

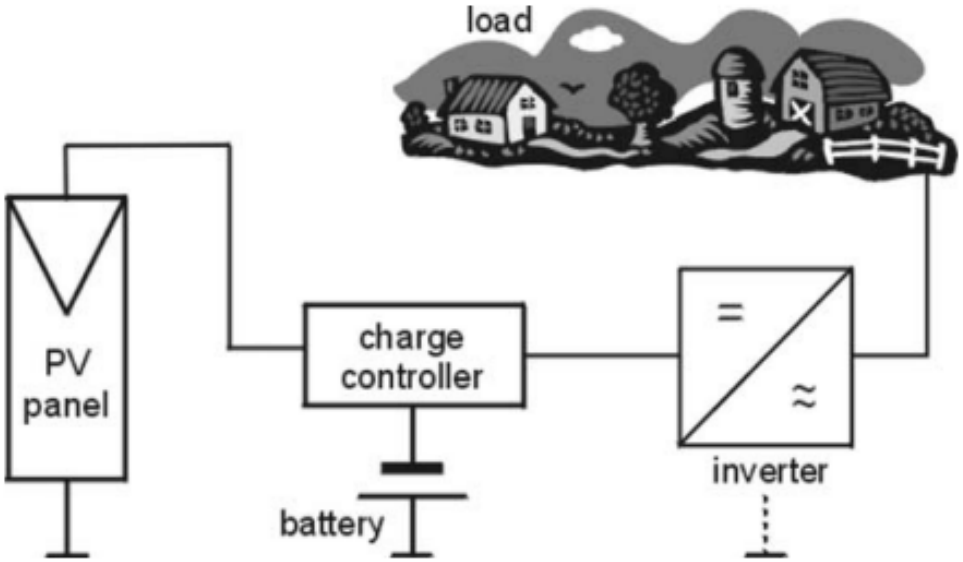



Figure 1.5: *Block diagram of an autonomous PV system with storage and an inverter for AC loads.*

### 1.5.5 Charge(Users)

There are two types of devices powered by the system, the direct current such as telecommunications equipment, water pumping, and AC in the case of domestic use. This case requires an inverter. The use of photovoltaic energy must be thought in terms of power. It is,therefore, more advantageous to look for consumers running continuously rather than adding an inverter and a 220 $V_ac$ consumer [8].

## 1.6 Solar Radiation

### 1.6.1 Definition

Solar radiation is the energy per unit vicinity obtained from the sun in the shape of electromagnetic radiation. The SI unit of photovoltaic irradiance is Watt per rectangular meter $W/m^2$. The find out about and size of photovoltaic irradiance is fascinating for the prediction of the power era of photovoltaic energy plants [9].

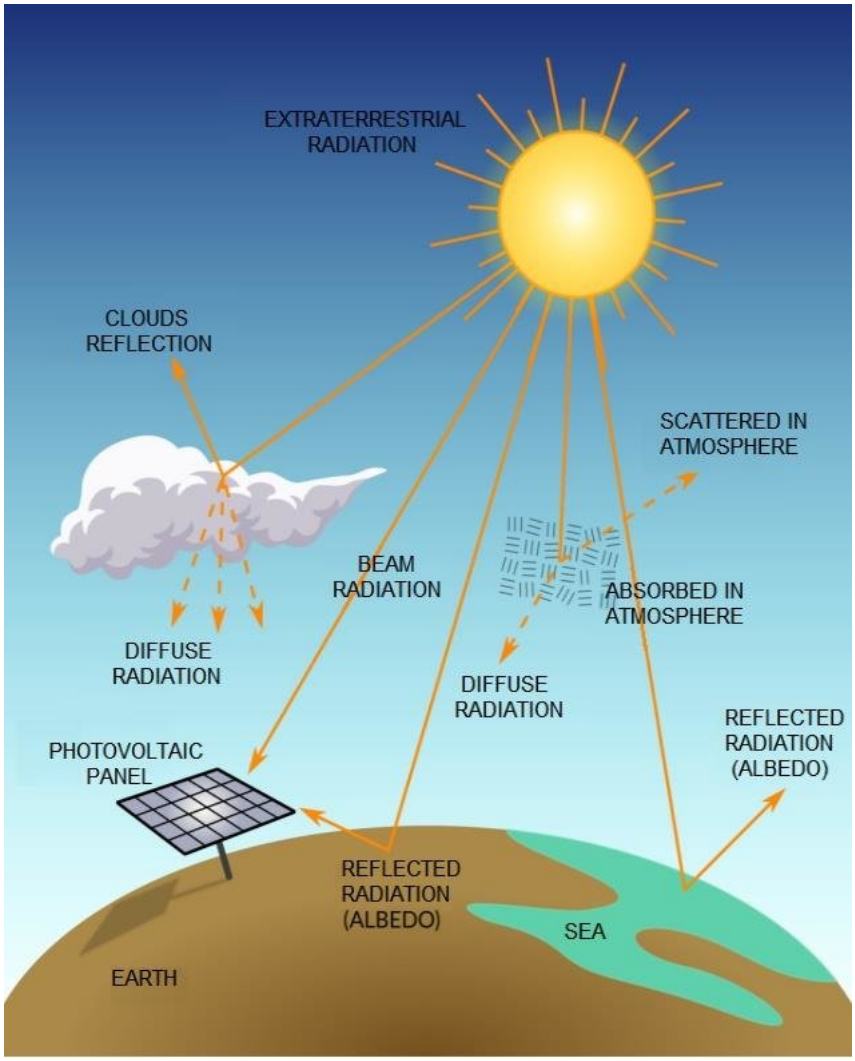



Figure 1.6: *Representation of Solar Radiation Components.*

### 1.6.2 Type of Solar Radiation

#### Direct Normal Irradiance (DNI)

Additionally acknowledged as beam irradiance is the photovoltaic radiation measured at a floor of the earth perpendicular to the Sun. It solely measures the direct radiation from the solar disk and excludes the diffuse radiation [9]. In Fig 1.6 from [10], a representation of Solar Radiation Components.

#### Diffuse Horizontal Irradiance (DHI)

Is the radiation measured on a horizontal floor on Earth, coming from mild scattered by way of the atmosphere. It measures radiation from all points in the sky except for radiation from the solar disk. In the absence of atmosphere, there has to be nearly no diffuse sky radiation [9].

### Reflected Radiation

Is the radiation mirrored by using non-atmospherical factors such as the ground. However, photovoltaic panels tend to be tilted away from the mirrored radiation trajectory, so it hardly ever has relevance in the whole radiation acquired by using their surface. An exception is in stipulations the place the floor is surrounded with the aid of snow, which can make more significant extensively the mirrored radiation received [9].

### Global Horizontal Irradiance (GHI)

Global horizontal irradiance is the total irradiance from the solar on a horizontal floor on Earth [9]. GHI that hits the surface of the earth consists of two components: direct normal irradiance and diffuse horizontal irradiance.

The geometric relationship between GHI, DNI, and DHI can be expressed as:

$$GHI = DNI.\cos\theta_z + DHI \tag{1.1}$$

where, $\theta_z$ the zenith angle. Zenith angle is the angle between zenith and center of the Sun's disk. The unit of global horizontal irradiance is watts per square meter. The relationship between zenith angle ($\theta_z$) azimuthal angle ($A_z$) and altitude ($\alpha$)is shown in Figure 1.7.[11]

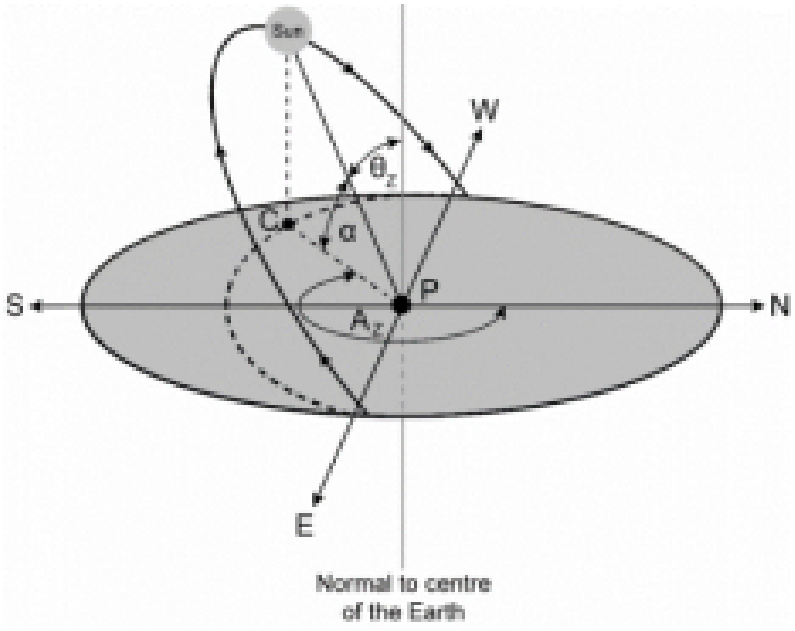



Figure 1.7: *Locus of the sun*

## 1.7 Conclusion

This chapter talked about energy and its importance to humankind, especially Renewable energies. And that by exploring the main types of renewable energies focusing on solar power. Then we mentioned the Photovoltaic cells and how they work in the solar power system. In the end, we briefly explained solar radiation and its different types.

# Chapter 2

# Literature Review

## 2.1 Introduction

PV power generation is dependent on many factors, such as weather conditions and PV module temperature. In this chapter we will explain more about the different methods used to forecast solar radiation from weather conditions and review some research articles related to solar radiation forecasting, trying to discuss their significant results.

## 2.2 Numerical Weather Prediction

The majority of NWP models are based on numerical integration equations, which necessitate domain knowledge to describe the radiation process and atmospheric fluctuations. The primary advantage of NWP, according to Cornaro et al. in [18], is that it is a deterministic physical model. The NWP model, however, is constrained by the non-linearity of the domain equations as well as the poor spatial resolution of the integration grid, which ranges from 100km to a few km, which is too wide in comparison to the size of a PV plant, according to the authors. The spatial resolution of NWP models is explored in [19, 20, 21]. There are two types of NWP models: global and mesoscale models. NWP models do not allow detailed mapping of small-scale features due to their coarse resolutions.Although the resolution of the NWP has increased in recent years, the range of resolutions still varies between 16 and 50 km depending on the model, undermining predicting accuracy.In terms of temporal scales, Lorenz and Heinemann [22] reported that NWP models are commonly used to anticipate atmospheric states up to 15 days in advance, demonstrating the limitations of using NWP models for longer-term forecasts. In conclusion, the accuracy of NWP models is determined by the availability of meteorological information, and NWP models perform better when used for short-term forecasting.

## 2.3 Statistical Models

Time series provides statistical information to foresee the nature of the quantified element. Depending on the variable's change with time, these observations are often collected through time at regular intervals such as quarterly, monthly, weekly, daily, or even hourly and minutes [12]. The goal of time series analysis is to forecast future values by analyzing the pattern of previous data.

### 2.3.1 Exponential smooth

The exponential smoothing method, also known as the exponentially weighted moving average (EWMA), is a unique methodology for statistical analysis and prediction of historical time series data that uses an exponential window function. In general, it gives historical observations an uneven set of weights over equal weights, lowering the data exponentially from the most current to the most distant data points. It can, however, quickly learn and make decisions based on assumptions. The technique was first formulated by Brown [13] and has since seen many applications. Overtime, it was extended by

Holt in 1957 and by Winter in 1960. It is thus called Holt-Winter's method [14]. The governing equation is as follow:

$$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t + \alpha(Y_t - \hat{Y}_t) \tag{2.1}$$

where, current observation is $Y_t$ ; predicted value is $\hat{Y}_t$ ; and smoothing constant is $\alpha$, which remains between 0 and 1. Therefore, the forecasting equation outputs the predicted value at t + 1 which is equal to the sum of the last predicted value $\hat{Y}_t$ and the forecasted adjustment factor $\alpha(Y_t - \hat{Y}_t)$.

### 2.3.2 The autoregressive moving average model (ARMA)

ARMA is a statistical time series analysis that is commonly used in forecasting. Many academics have tested the model in many forecasting applications (solar and wind forecasting), and it has consistently delivered high prediction accuracy. For forecasting PV production from historical data, the model contains two polynomials: AR and MA.[15]. The mathematical expression is as follows:

$$X(t) = \sum_{i=1}^{p} \alpha_i X(t-i) + \sum_{j=1}^{q} \beta_j e(t-j) \tag{2.2}$$

where, predicted PV output is represented through function X(t) which is a combination of AR and MA functions. p and q indicate the number of processes or the order, while $\alpha_i$ and $\beta_j$ are the coefficients of AR and MA models, respectively. e(t) is randomly generated white noise; it is not correlated with a model's prediction.

ARMA models are extremely versatile, and they may be used to represent a variety of time series simply varying the order. In addition, this model may detect the presence of an underlying linear auto-correlation structure. [16].

In order to provide very short-term forecasts of the solar irradiance, Mathieu David et al [15] , The ARMA-GARCH approach was utilized in their research, which is an effective combination of models for generating extremely short-term point predictions of solar irradiance with confidence intervals. They also employed a recursive ARMA model to construct point predictions, which is a simple and practical method. Also in , [11],The ARMA approach fared equally for the 15-minute, 30-minute, 1-hour, and 2-hour forecasting horizons, but outperformed the LASSO and RR strategies for the 3-hour and 4-hour forecasting horizons. This approach surpasses previous statistical models by using simply the actual value of solar irradiation.

### 2.3.3 Autoregressive integrated moving average (ARIMA)

ARIMA is also known as Box-Jenkins model and was developed by George Box and Gwilym Jenkins in 1976 [17]. The ARIMA model is a prominent time series analysis

approach that offers the usual level of forecast accuracy for short-term horizons. It is an expanded version of ARMA. Furthermore, this model may remove non-stationary values from the data being studied. To analyze and forecast time series features, it has a framework that includes autoregression (AR) and moving average (MA). [18] .The general form of ARIMA (p, d, q) model of time series $X_1$, $X_2$, $X_3$ is as follows:

$$\Phi_p(B)\Delta^d X_t = \Theta_q(B)\alpha_t \tag{2.3}$$

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 \dots \phi_p B^p \tag{2.4}$$

$$\Theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 \dots \theta_q B^q \tag{2.5}$$

where, the backward shift operator is B; backward difference is $\Delta$ = 1 - B and $BX_y = X_{y-1}$ ; $\phi_p$ and $\phi_q$ are polynomial numbers of order p and q, respectively. effectively. As a result, the ARIMA (p,d,q) model is a composite sum of autoregressive part (p), an integrating part $I(d) = \Delta^{-d}$, and a moving average part (q). The variables in $\phi$ and $\Theta$ are precisely selected so that the zeros of all polynomials fall out from the unit circle to evade the creation of interminable processes. To consider the arbitrary disturbance taken from a fixed distribution with zero mean and $\sigma_a$ variance $a_t$, $a_{t-1}$, $a_{t-2}$ are introduced as white noise process. Therefore, the intrinsic characteristics of the time series can be comprehend by the white noise process and backshift operator.

In this paper, [19], ARIMA model has been trained on GHI data of March 2016 collected from The Petroleum Institute, Abu Dhabi, UAE and a prediction has been done for 31 March 2016, based on the previous 30 days, and the hourly forecast has a $R^2$ value of 88.63%.

In this paper,[20], ARMA and ARIMA models are compared in terms of prediction accuracy for the multi-period prediction of GHI. From this analysis, it is clear that the ARIMA model has better accuracy in the multi-period prediction approach.

## 2.4 Machine Learning

### 2.4.1 Definition

Machine learning is one of the domains of artificial intelligence (AI). A machine learning algorithm has to aims to analyze the information available from a large number of statistical data to learn to carry out a study without having explicitly been programmed for this upstream [8].Machine learning has been a real success in recent years. With the growth exponential number of digital data available, we need to use new analysis methods, and so-called machine learning methods correspond to this need.

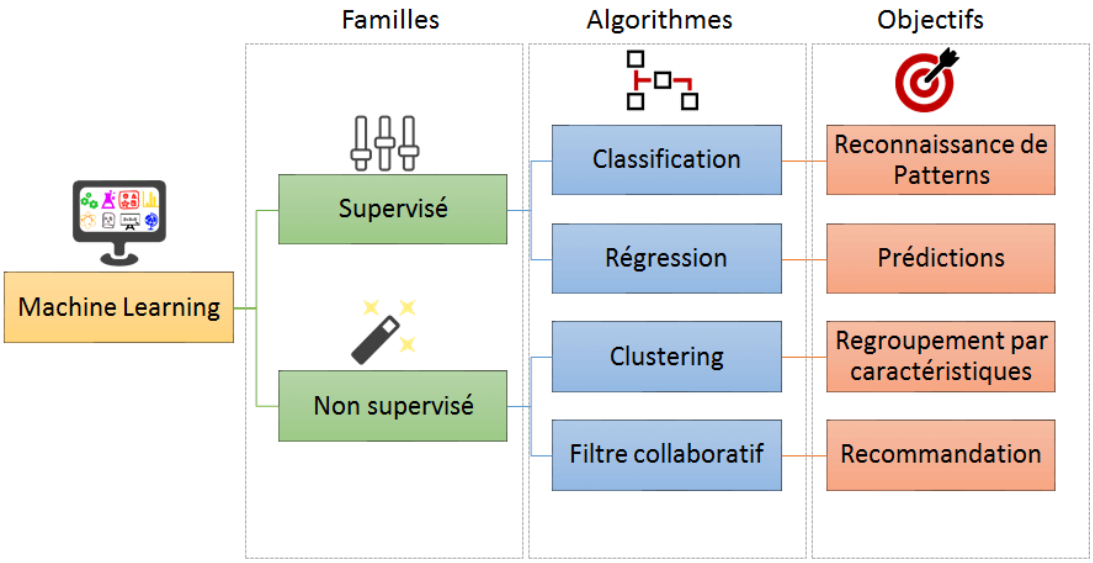



Figure 2.1: *Familly of Machine Learning*

### 2.4.2 Types of Learning in Machine Learning

#### Supervised Learning

Supervised learning is intended to create a predictive function for one of the variables Of our database as a function of others. The variable we want to predict is called "variable to be explained," the other variables used to guide the prediction are the "variables Explanatory". The algorithm tries to learn, by browsing the database available, The different causal links between the explanatory variables and the variable to be explained. Once our model is created, it matches each possible combination of explanatory variables, a prediction of the corresponding variable to be explained. For this, he must group the individuals of the base into subgroups by maximizing the homogeneity of the variable to be defined. Still, the groups must be separated only according to their explanatory variables.

There are two categories of supervised learning:

- Regression algorithms, when the variable to be explained, is quantitative. In this case,the prediction is a value.

- Classification algorithms, when the variable to be explained, is qualitative. In this case,the prediction is often a probability of belonging to the different modalities of the variable in question [21].

**Unsupervised Learning**

Non-served learning occurs when there is no response variable to predict. These algorithms are also used on a database, but in this case, its purpose is to determine the structure present in that database.

To do this, it must, like the supervised algorithms, group individuals into the most homogeneous subgroups possible, however here we no longer have variables to explain, so homogeneity must be done on all variables.

What interests us here is not the prediction of new data, but instead, how groups are determined and what commonalities exist between the individuals of each subgroup [21].

Of the two learning families discussed above, we will only use supervised learning algorithms in this brief. We aim to estimate a variable to be explained, the S/P ratio of contracts, based on the information available about these contracts, which are the explanatory variables. Supervised learning methods, therefore, appear to be adapted to our problems, such as the CART algorithm (Classification And Regression Trees), bagging, or random forest. The purpose of this section is to present the operation and the main characteristics common to these methods .

**Reinforcement Learning**

In this type of study, each marked and unlabeled statistic may want to be utilized to shape the indispensable knowledge. The framework receives a reward for each good or incorrect forecast. Depending on the reward, the subsequent forecast should be generated. At the factor when new information is given to the framework, the framework will pastime to stumble on the excellent execution way or be part of a couple of execution pathways for forecasting and pause for the reward. When the received reward takes place to be greatest with recognition to the previous rewards for equal input, at that point, this pathway flips out to be agreeable Reinforcement gaining knowledge of is utilized in net-primarily based games, for example, Chess [21].

### 2.4.3 Classification And Regression

#### Classification

The classification consists of determining the instructions of belonging of new objects from recognized previous examples, the variable to consider can therefore take discrete values called classes. We will see that our problem is multi-class, that is, the exchange of opinions can be represented using several courses, as an antagonist to the binary classification represented using a two-class output variable [8].

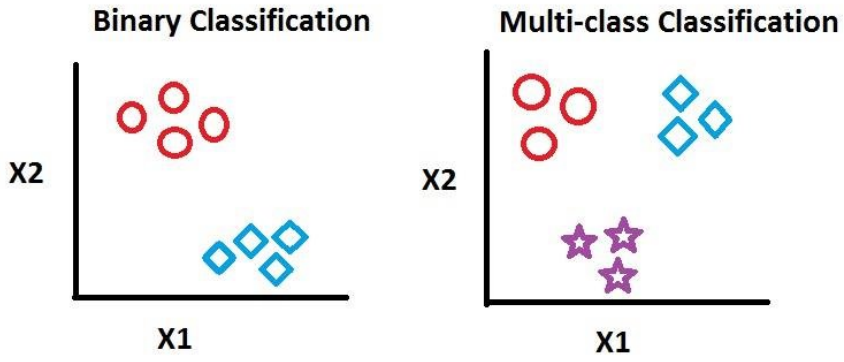



Figure 2.2: *1 Binary vs Multiclass classification*

#### Regression

Regression is used when predicting a continuous variable, which can therefore take any value. The classes representing the note change have therefore were considered constant and the regression algorithm allowed to obtain decimal values that we finally rounded to get a vector of predicted levels [8].

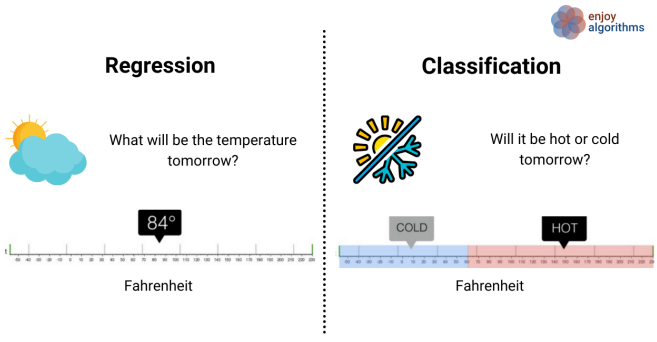



Figure 2.3: *Classification and Regression in Machine Learning*

## 2.5 Algorithms of Machine Learning

### 2.5.1 Support Vector Regression SVR

**Definition**

Support vector regression is a supervised machine learning algorithm purposed by Vapnik and improved by Smola, and Scholkopf. Our construction of SVMs for the regression problem is based on the $\varepsilon$-insensitive loss function. This loss function has the same structure as robust loss functions: It combines two functions one of which is $f(u) = |u|$ and the constant function[4] : f(u) = const (we considered case const = 0).[22]

The $\varepsilon$-insensitive implies some new properties of the SVM solutions, namely the sparsity of solutions. By changing (increasing) the value of $\varepsilon$ one controls (increases) the sparsity of the SVM solutions.

However, the difference between the robust approach and SVM approach reflects also the fact that the loss function for the SVM regression is more Formally it does not belong to the family of Huber's robust estimators, since the uniform distribution function does not possess a smooth derivative.

complicated than the loss function for robust regression. For linear functions it has the form:

$$L(a) = \frac{1}{C}(w, w) + \sum_{i=1}^{\epsilon} |y_i - (w, x)|_{\varepsilon} \tag{2.6}$$

where (w, w) is the regularization functional and l/C is the regularization parameter. The addition of the regularization term into the functional dramatically changes the situation: On one hand it connected SVM regression to regularization techniques introduced for solving ill-posed problems, and on the other hand it increases the number of free parameters.

Now, in order to estimate the regression function we have to specify three free parameters: the value of c-insensitivity, the regularization parameter C, and the kernel parameter (the order of the polynomial for polynomial kernels, the width parameter for radial basis kernels, the order of the spline for spline generating kernels, and so on). [22]

**Related works**

In [23], Jebli et al studied the prediction accuracy of four different ML models for real-time and short-time solar energy forecasts. The models were experimented with data from a Moroccan region (semi-desert). The SVR Model show poor results according to other models.

In [11], The SVR outperformed all other techniques, with the lowest R2_score being 0.80 for the 4-hours ahead forecasting.

### 2.5.2 Linear Regression (LR)

The word "regression" and the methods for examining the correlations between two variables are said to have originated around 100 years ago. Francis Galton, a distinguished British scientist who studied heredity, was the first to suggest it in 1908. One of his findings was that offspring of tall parents are taller than the norm, but not as tall as their parents. The term "regression toward mediocrity" was used to describe these statistical procedures. The term regression, as well as its history, refers to statistical relationships that exist between variables. Simple regression, in particular, is a regression approach for examining the connection between one dependent variable (y) and one independent variable I (x).

To statisticians, the regression line is significant in and of itself, and it will aid us in multiple regression. A scatter plot depicting the heights of 1078 dads and their sons is one example. On the diagram, each pair of fathers and sons becomes a dot. The father's height is displayed on the x-axis, while his son's height is plotted on the y-axis. The families whose fathers are 64 inches tall to the closest inch are shown on the left-hand vertical strip (within the chimney), whereas those whose fathers are 72 inches tall are shown on the right-hand vertical strip. There are plenty alternative comics that might be drawn. Given the heights of their dads, the regression line approximates the average height of the sons. This line connects the vertical strips in the middle. The regression line is smoother than the dashed SD line. The abbreviation "SD" stands for "standard deviation."[24]. In Fig 2.4 we see an example of a LR model.

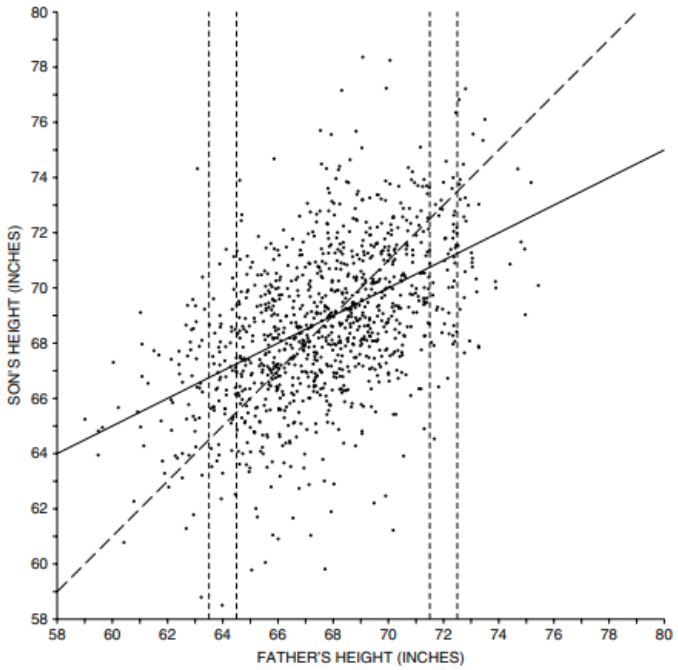



Figure 2.4: *Linear Regression exemple*

#### LR Model Representation

Because of its easy representation, linear regression is an appealing model. The representation is a linear equation that combines a specific set of input values ($x$) and yields the expected output for that set of input values ($y$). As a result, both the input and output values ($x$) are numerical.

Each input value or column is assigned one scale factor, known as a coefficient, which is often symbolized by the Greek letter Beta $\beta$. One more coefficient is added, which gives the line an extra degree of freedom (for example, going up and down on a two-dimensional plot) and is known as the intercept or bias coefficient. In a simple regression situation (a single $x$ and a single $x$, for example,, the form of the model would be:

$$y = B0 + B1 \times x \tag{2.7}$$

When there are several inputs (x) in higher dimensions, the line is termed a plane or a hyper-plane. As a result, the representation is the equation's form as well as the coefficients' precise values (e.g. B0 and B1 in the above example). The complexity of a regression model, such as linear regression, is frequently discussed. The number of coefficients utilized in the model is referred to as this..

When a coefficient hits zero, it effectively removes the input variable's influence on the model and, as a result, the model's prediction. This is important to consider when considering regularization strategies, which alter the learning process to minimize the complexity of regression models by exerting pressure on the absolute size of the coefficients and pushing some to zero. Let's take a look at some of the methods we can learn this representation from data now that we know what it is for a linear regression model. [25].

### 2.5.3 Random Forest (RF)

#### Definition Of The Model

The method of machine learning that we will see in this part is the "random forest," better known as RF, introduced in 2001 by L. Breiman in his publication bearing the name of the method in question.

As a reminder, the purpose of these so-called established methods is to create several slightly different trees from a single initial base to group these trees and thus reduce the variance of our model. To make a large number of other trees, we can either modify the learning base of each tree or change the algorithm used to create it [25].

### How It's Work

In the bagging section, the trees were created from the CART algorithm. It consists in choosing at each node the separation maximizing the intergroup variance among all possible, a separation representing a test on one of the p explanatory variables of the model. However, in this part, this algorithm will be modified so that before each separation, which is called a node, we randomly pull $m$ explanatory variables among the $p$'s of our model. The separation chosen for the node in question will therefore be the optimal one only among the tests involving the selected m variables. Thus, at each node, the information carried per variable p-m drawn randomly is ignored. Apart from the choice of separation, the rest of the algorithm takes place Then, for each bootstrap sample, we create the maximum tree using our new modified algorithm. As for the bagging method, we do not prune the trees because we want to keep the bias as low as possible. Finally, we aggregate the maximum trees together, so that the prediction of our random forest model is the average of the predictions of the maximum trees. It is interesting to note that bagging could be considered as a method of Random forest, in the particular case where $m = p$. Indeed, if $m = p$, then at each node, we choose optimal separation among the tests for one of the selected p variables, that is, all Variables. We find the CART algorithm used in the bagging method. To remain in line with the ratings used previously:

- B number of bootstrap samples created .
- $n_{learning}$ The size of the sample of data used for learning
- estimation of the output variable of individual I by our random forest model.
- the estimation of the output variable of individual I buy the maximum tree created from of the bootstrap k sample with a random forest algorithm with the parameter $mtry = m$.

The formula is, therefore, the same as for a bootstrap, the difference in how to build the tree after creating the bootstrap sample [25]. In Fig 2.5 the RF psuedocode.

### Advantages of using Random Forest

- The random forest algorithm is not biased, since, there are multiple trees and each tree is trained on a subset of data. Basically, the random forest algorithm relies on the power of "the crowd"; therefore the overall biasedness of the algorithm is reduced.
- This algorithm is very stable. Even if a new data point is introduced in the dataset the overall algorithm is not affected much since new data may impact one tree, but it is very hard for it to impact all the trees.
- The random forest algorithm works well when you have both categorical and numerical features.
- The random forest algorithm also works well when data has missing values or it has not been scaled well (although we have performed feature scaling in this article just for the purpose of demonstration).

**Random Forest pseudocode:**

1. Randomly select **"k"** features from total **"m"** features.
    0. Where **k << m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split.**
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees.**

Figure 2.5: *Random Forest pseudocode*

**Disadvantages of using Random Forest**

- A major disadvantage of random forests lies in their complexity. They required much more computational resources, owing to the large number of decision trees joined together.
- Due to their complexity, they require much more time to train than other comparable algorithms.

## 2.6 Related Works

In the table that follows, we have outlined many study publications that are associated with solar power forecasting together with their most impressive achievements. We make an effort to categorize them in accordance with the year of publication, technique used, dataset, and performance metrics.

| Ref. | Year | Method | Data Used | Performance Metrics |
|---|---|---|---|---|
| [11] | 2020 | SVR<br>KRR<br>LASSO<br>RR<br>ARMA | National Renewable Energy Laboratory (NREL), Wichita, Kansas, USA | RMSE: 0.0943%<br>RMSE: 0.1106%<br>RMSE: 0.1107%<br>RMSE: 0.1106%<br>RMSE: 0.0984% |
| [8] | 2020 | RF | HI-SEAS Weather Station | MAE: 9.64 |
| [26] | 2021 | MLR<br>SVR<br>RF | | MAE: 0.58 /$R^2$: 0.87<br>MAE: 0.36 /$R^2$: 0.92<br>MAE: 0.25 /$R^2$: 0.96 |

## 2.7 Conclusion

In this chapter, we tried to explain the current Solar Irradiance forecasting methods giving their principles and basic techniques. Then we introduced Machine Learning, its main types, and some of its models, which we will use some of them in our project to forecast solar radiation. Finally, we reviewed some research articles and present their significant results by classifying them in a table for a better vision.

In the next chapter, we will present the Data set and the main tools that will be used for this project.

# Chapter 3

# Data and Tools

## 3.1 Introduction

In this chapter, we are going to provide the data set that we utilized for our project, along with several data visualization charts that will highlight the most important aspects of the data. In addition, we will have a quick discussion of the environment and libraries that assist us in the building of our code.

## 3.2 Nevada Dataset

### 3.2.1 General Information

The NSRDB [27] dataset includes both observed weather data (temperature, pressure, cloud cover, solar zenith angle,etc.) and solar intensity data measured in watts per square meter. The dataset includes several solar radiation measures such as Diffused Normal Irradiance (DNI), Diffused Horizontal Irradiance (DHI) and Global Horizontal Irradiance (GHI). We choose to include GHI measurements since it incorporates DHI, DNI and ambient solar radiation reflected from nearby surfaces. This makes it a good indicator for solar panel readings.

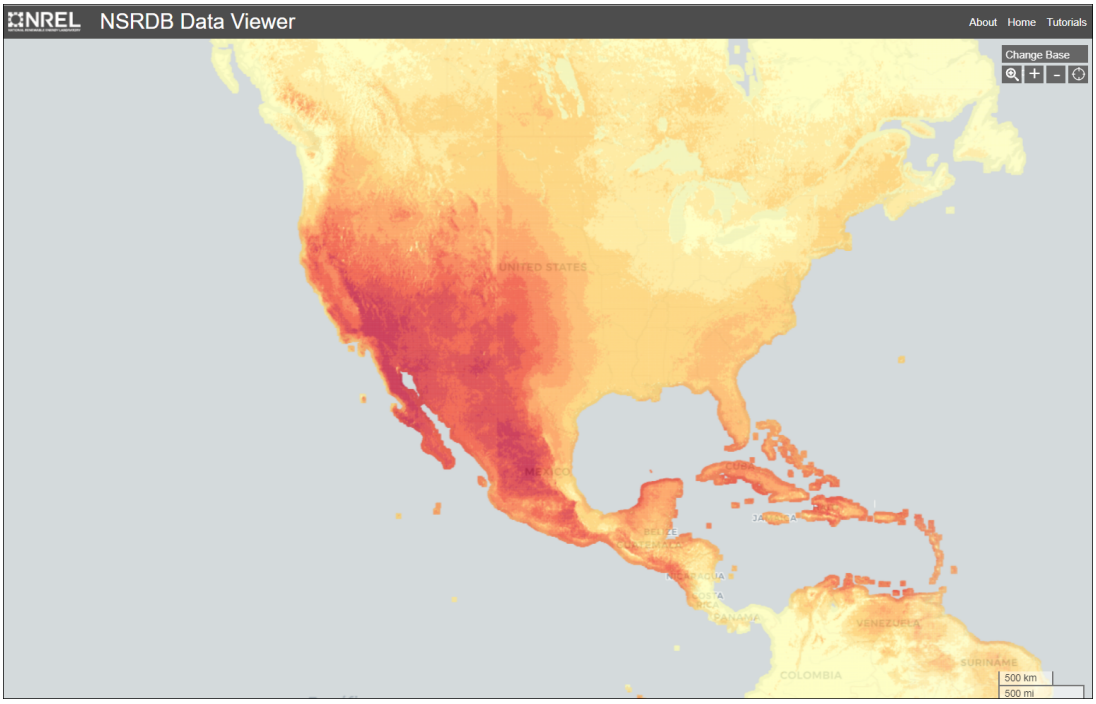



Figure 3.1: *Solar Power Intensity over North America [NSRDB]*

The NSRDB data is measured once every 10 minutes. We investigate a single location - Las Vegas, Nevada, USA. We limit our analysis to the year 2019 only. Our dataset contains more than 50,000 distinct observations, each with 16 features (including Time values such as Month, Day, Hour and Minute) shown in table 3.1 with corresponding measure of GHI.

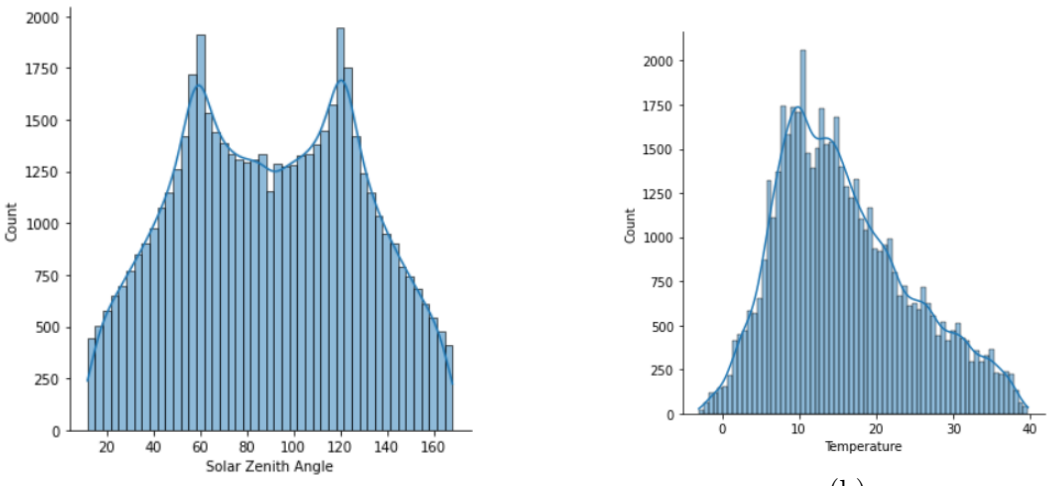
| | | | |
|---|---|---|---|
| **Month** | **Day** | **Hour** | **Minute** |
| **Temperature** | **Cloud Type** | **Fill Flag** | **Surface Albedo** |
| **Ozone** | **Pressure** | **Dew Point** | **Precipitable Water** |
| **Wind Direction** | **Wind Speed** | **Relative Humidity** | **Solar Zenith Angle** |

Table 3.1: Variables of Nevada dataset

### 3.2.2 Data Visualization Charts

We plot distribution of each feature using histograms, pie-charts and box-plots. We also study pairwise correlation between features using correlation-matrix & skew index table.

#### Solar Zenith And Temperature




(a) (b)

Figure 3.2: Plots: (a) Solar Zenith (b) Temperature

### Wind Direction

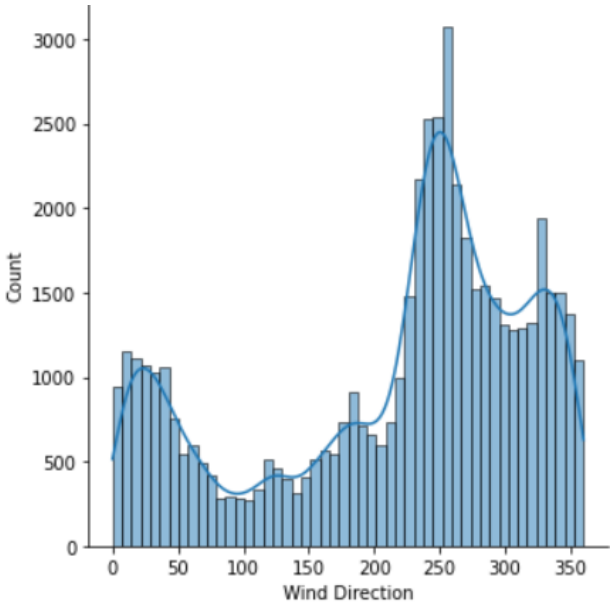
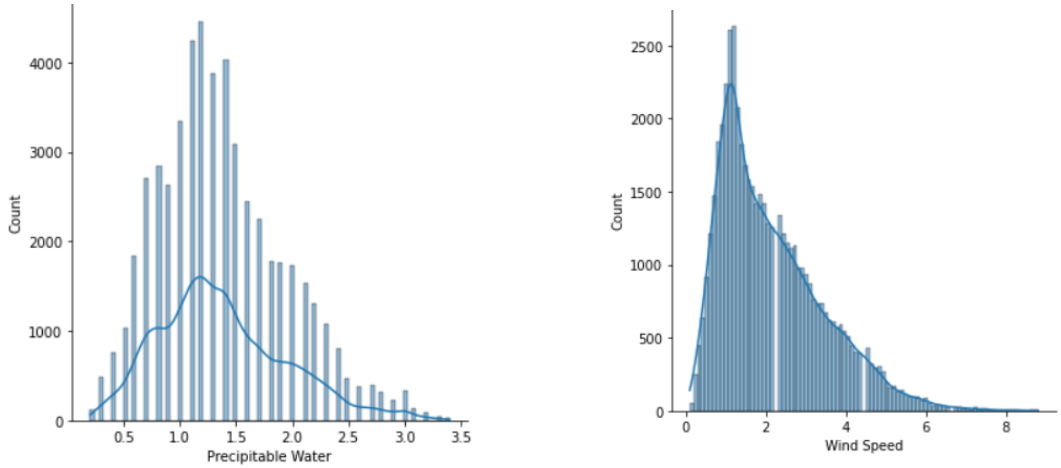



Figure 3.3: *Wind Direction*

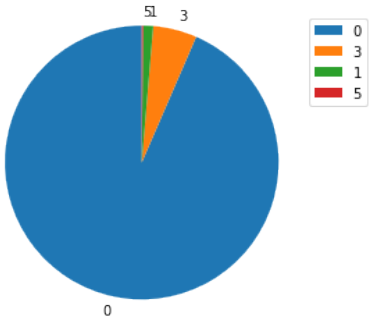
### Precipitable Water And Wind Speed



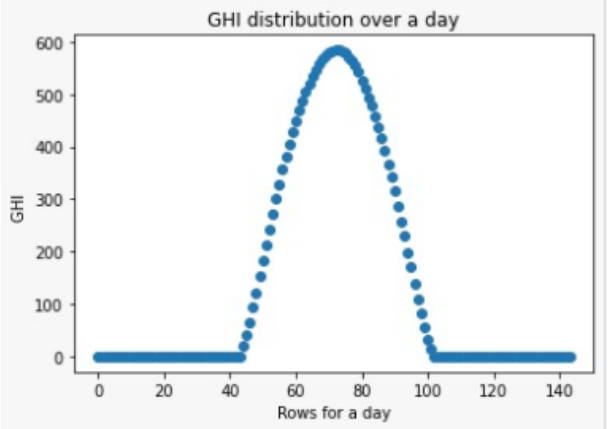

(a) (b)

Figure 3.4: Plots: (a) Precipitable Water (b) Wind Speed

### Fill Flag And GHI




(a)




(b)

.

Figure 3.5: Plots: (a) Fill Flag (b) GHI over a day

### Mean GHI by Hour

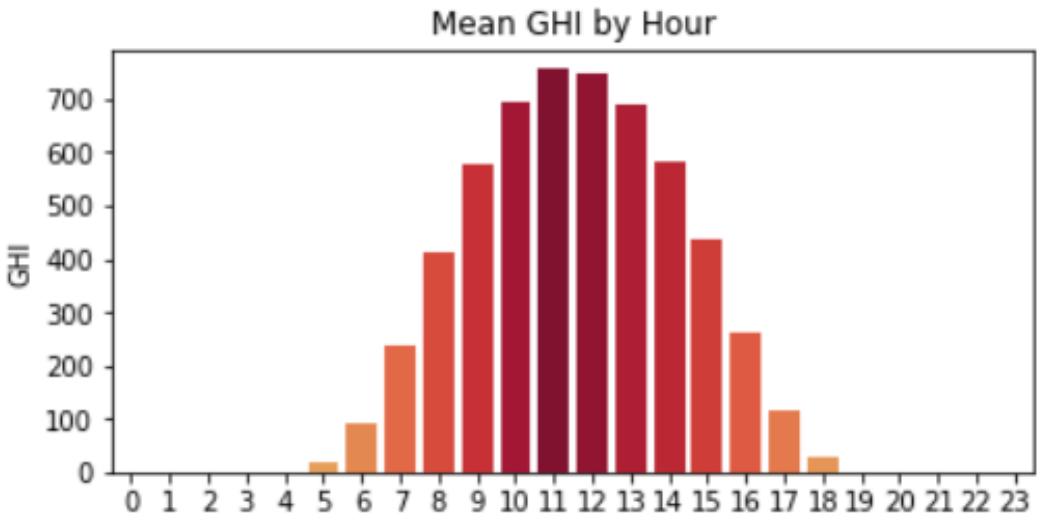



Figure 3.6: *Mean GHI by Hour*

**Mean GHI by Month**

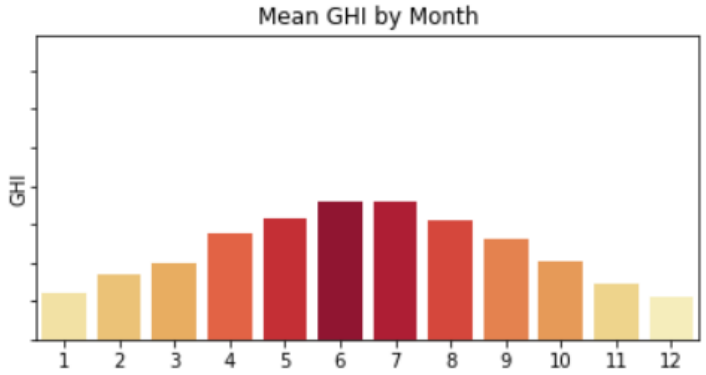



Figure 3.7: Mean GHI by Month

### 3.2.3 Preprocessing Dataset

We spotted outliers in the distribution curves. However, we assumed that they are the natural part of the weather observations we are studying, hence we didn't remove them. Also we didn't find any missing/NaN values in our dataset, which would result in imbalance observations.

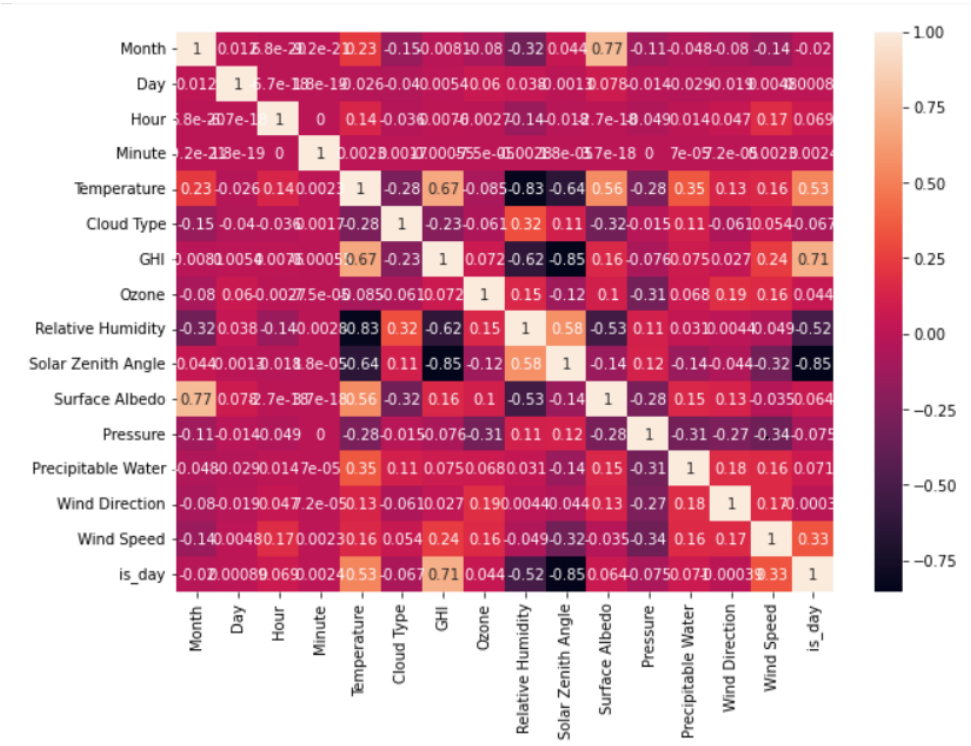



Figure 3.8: *Correlation Matrix*

The Correlation Matrix (Fig: 3.8) shows that features 'Dew Point' and 'Precipitable Water' are highly correlated. Since highly correlated features have almost the same effect on the dependent variable, we drop one of them. We test the performance of our baseline model by dropping both of the features one by one. Based on the results, we choose to drop 'Dew Point'. The Pie-Chart for 'Fill Flag' feature shows that more than 90% observations correspond to '0'. It indicates that the value is not available. Therefore 'Fill Flag' adds negligible information to the dataset. Hence, we drop 'Fill Flag'.

**Splitting The Data**

We split the data into 20% for Testing and 80% for Training.

## 3.3 El Hadjira Dataset

During the one-week period, we did a practical training at the SKTM (Sharikate Kahraba Wa Taket Moutadjadida) renewable energy station in El Hadjira. They introduce to us the different technologies used in the station to generate solar energy, which is then transported to the distribution station and finally to the final consumers.

At the end of the training they provide us with some data collected in the station to use it in our project. Below you will find an overview of the company as well as a summary of the most significant features of the area.

### 3.3.1 Photovoltaic Power Center

Due to its geographical location, Algeria has one of the highest solar deposits in the world (05 billion GWH/year), with a period of sunshine over the desert and highlands that can exceed 3000 h/y, according to specialists.In the case of El Hadjira/Ouargla, 2800 h/y. El Hadjira Solar PV Station 30MW is part of the National Renewable Energy and Energy Efficiency Program.

Figure 3.9: *Some solar panels from Photovoltaic power center (SKTM)*

The expectations for the realization of this station are:

- Diversifying the sources of electricity production and developing the means of production from renewable sources.

- The annual capacity produced by the photovoltaic plant is 52,000 megawatt-hours/year.

- Preserving primary resources: saving fossil fuels;(approximately 9200 Tons/year of gas).

- Protecting the environment by reducing greenhouse gas emissions (30,000 Tons/year of CO2 emissions reduction).

**General presentation of the station**

- Country: Algeria – W: Ouargla – El Hadjira

- Geometric coordinates: 32.35° N and 05.50° E.

- Area: Sixty (60) Hectares,

- Power Crete: 30,000 KWc

- Injection voltage: 30 kV

- Project manager and contracting authority: Sharikate Kahraba Wa Taket Moutadjadida - SKTM

El Hadjira Station consists of thirty (30) sub-fields. from 1 megawatt. Their main equipment is:

- Photovoltaic generator (PV modules, supports, junction box, electrical panels, wires)

- DC / AC conversion and conversion station (inverters, transformers, cells, electrical panels, wires)

- Auxiliary systems (emergency unit, battery charger rectifiers, lighting, remote monitoring and anti-intrusion, detection and firefighting.)

**The Weather Station of Photovoltaic power center (SKTM)**

Figure 3.10: *Different components of weather station (SKTM)*

### 3.3.2 Description of The Dataset

As we mentioned before, the data obtained from the meteorological station of the SKTM's station. which were collected over the year of 2021 (12 month). The data comes in the form of Excel files divided into daily files and monthly folders.

Here are the most important points and characteristics of the collected data :

- Information is collected on a daily basis, every half hour (30 minutes).

- The collection process starts from 6 am to 8 pm, which is the estimated time from sunrise to sunset.
- The dataset contains the most important weather factors affecting the production of solar energy (Temperature, Barometric pressure..etc) as well as the amount of received solar radiation. In addition to the amount of power extracted every 30 minutes.

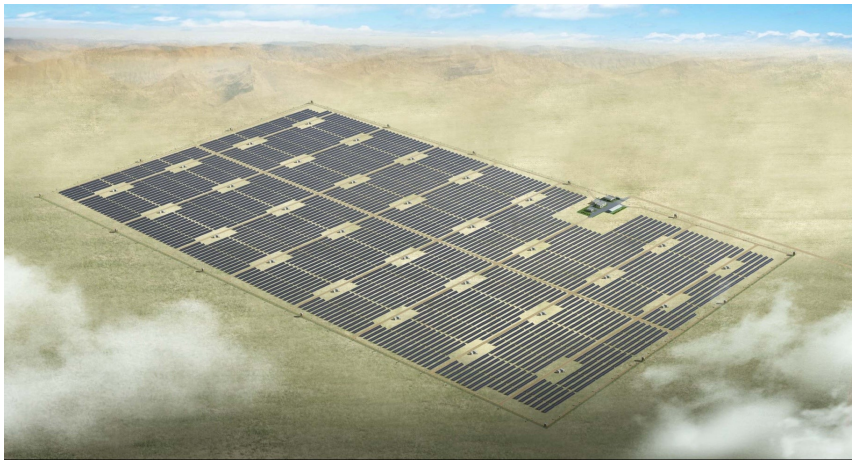


Figure 3.11: *Sky Image of the station*

### Variables of the dataset

The following (Tab: 3.2) is a table that shows these factors and their units of measurement as they are found in the data:

| | |
|---|---|
| **Solar radiation:** Watts per $meter^2$ ($W/m^2$ ) | **Temperature:** degrees Celsius (°C) |
| **Humidity:** percentage (%) | **Barometric pressure:** (Hpa) |
| **Wind speed:** meter per second (m/s) | **Total Power:** Kilo Watt (KW) |

Table 3.2: Variables of the dataset

## 3.3.3 Preprocessing Dataset

Data preprocessing is a data mining technique in which raw data is converted into an understandable format. The real data is often incomplete: missing attribute values, missing specific attributes of interest, data preprocessing is a proven method of solving such problems.

### Step One: Merge the Data collected

Due to the fact that the data acquired is scattered over multiple files, we were obliged to combine them all. Therefore, we will end up with a single large file that contains all of the variables and the data associated with them. After the merge we convert the Excel file to a '.csv' file, that will help us manipulate the data later on the coding phase.

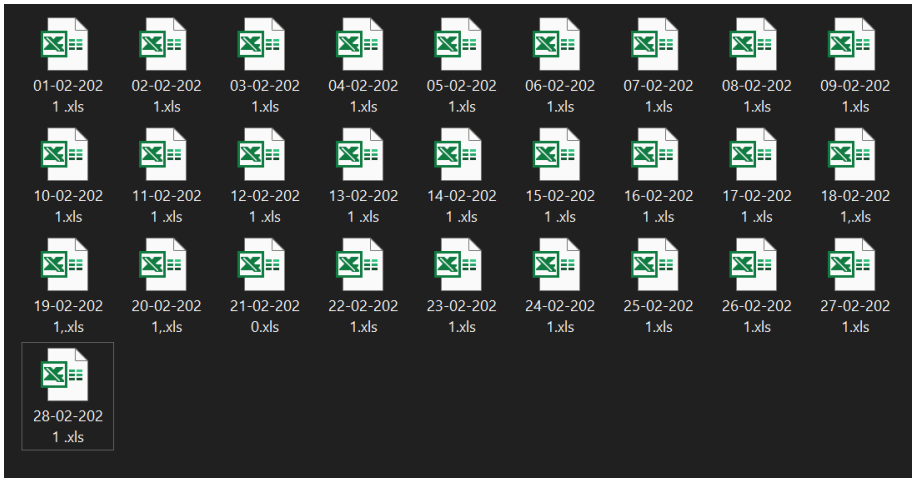



Figure 3.12: *Example of one month data separated in many docs.*

### Step Two: Handling the Missing Data

After doing an analysis of the data, we discovered that certain hour's values were missing.This is due to technical malfunctions affecting the meteorological station or the station's receivers. Consequently, we are required to address this problem and clean up the data. In order to ensure that we only retain the essential and useful information from which the model will learn. At the end of this step, we will have a total of 9919 rows.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 06:00 | -130 | 6.0000 | 0.0000 | 6.7000 | 1.2000 | 45.8000 | 1005.2000 |
| 06:30 | -130 | 8.0000 | 0.0000 | 6.2000 | 0.3000 | 46.5000 | 1005.1000 |
| 07:00 | 0 | 0.0000 | 0.0000 | 6.0000 | 0.4000 | 47.7000 | 1005.5000 |
| 07:30 | 0 | 0.0000 | 4.0000 | 6.1000 | 0.7000 | 47.3000 | 1005.9000 |
| 08:00 | 0 | 0.0000 | 23.4000 | 5.8000 | 1.1000 | 47.2000 | 1005.8000 |
| 08:30 | 0 | 0.0000 | 67.3000 | 6.8000 | 1.7000 | 46.7000 | 1006.4000 |
| 09:00 | 0 | 0.0000 | 112.2000 | 7.5000 | 1.0000 | 44.9000 | 1006.3000 |
| 09:30 | 0 | 0.0000 | 217.2000 | 8.7000 | 0.8000 | 43.5000 | 1006.8000 |
| 10:00 | 0 | 0.0000 | 231.5000 | 10.2000 | 2.4000 | 38.8000 | 1007.8000 |
| 10:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 11:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 11:30 | 0 | 0.0000 | 655.2000 | 15.7000 | 0.8000 | 33.0000 | 1007.8000 |
| 12:00 | 0 | 0.0000 | 603.0000 | 13.3000 | 0.3000 | 26.0000 | 1007.3000 |
| 12:30 | 0 | 0.0000 | 498.3000 | 21.0000 | 0.2000 | 23.8000 | 1006.7000 |
| 13:00 | 0 | 0.0000 | 511.8000 | 21.2000 | 1.8000 | 23.1000 | 1006.2000 |
| 13:30 | 0 | 0.0000 | 371.1000 | 20.7000 | 2.2000 | 22.7000 | 1005.3000 |
| 14:00 | 0 | 0.0000 | 352.1000 | 19.6000 | 1.0000 | 23.8000 | 1005.4000 |
| 14:30 | 0 | 0.0000 | 560.7000 | 19.2000 | 2.7000 | 23.9000 | 1005.3000 |
| 15:00 | 0 | 0.0000 | 530.1000 | 20.6000 | 2.7000 | 22.6000 | 1005.2000 |
| 15:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 16:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 16:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 17:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 17:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 18:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 18:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 19:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 19:30 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |
| 20:00 | 1000 | ---- | ---- | ---- | ---- | ---- | ---- |

Figure 3.13: *Example of some missing values in dataset.*

### Step Three: Correlation Matrix

We applied a Pearson correlation to the dataset. After analysing the results we found that Temperature, pressure, wind speed and humidity do not have a large direct effect on each other, yet they do not vary independently because they are all parameters that describe the local atmosphere. Similarly, the GHI of has a big influence on the power. This gives credence to the idea that the quantity of power generated by photovoltaic cells is mostly determined by the intensity of the incoming solar radiation.

| | Power | GHI | Temperature | Wind_Speed | Humidity | Pressure |
|---|---|---|---|---|---|---|
| Power | 1.00 | 0.92 | 0.22 | 0.16 | -0.20 | 0.04 |
| GHI | 0.92 | 1.00 | 0.45 | 0.19 | -0.37 | -0.02 |
| Temperature | 0.22 | 0.45 | 1.00 | 0.13 | -0.80 | -0.25 |
| Wind_Speed | 0.16 | 0.19 | 0.13 | 1.00 | -0.05 | -0.07 |
| Humidity | -0.20 | -0.37 | -0.80 | -0.05 | 1.00 | 0.22 |
| Pressure | 0.04 | -0.02 | -0.25 | -0.07 | 0.22 | 1.00 |

Figure 3.14: *Correlation Matrix.*

### Step Four: Splitting the Dataset

At this stage, we will split the data into two distinct parts: one will be used for learning, and the other will be used for testing. Following the presentation of the learning part for the model to learn from, it will precede to the evaluation phase, which is where the model's capacity for learning will be tested and rated based on the results obtained. Within the scope of our project, the data was divided into the following:

- Learning phase: from the $1^{st}$ month to the $11^{th}$ month in the dataset.
- Testing phase: the entire $12^{th}$ month of in data.

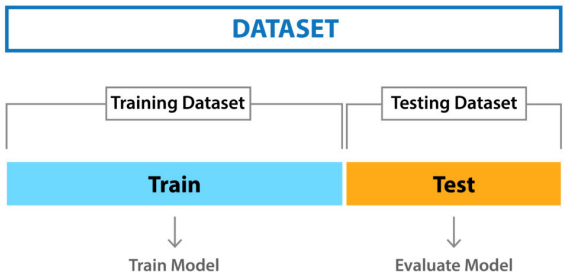



Figure 3.15: *Splitting the Dataset Diagram.*

## 3.4 Tools

The following tools and libraries have been used during implementation of this project:

### 3.4.1 Visual Studio Code (VS Code)

In a surprisingly short period of time, Visual Studio Code has become very popular among web developers. Part of that is because it's fast, lightweight, and is available on the three main platforms (Windows, Mac, Linux). But it also boasts several features that set it apart from the competition. This includes support for Intelligence, refactoring capabilities, and an extensive ecosystem of extensions. But even beyond the features that help developers craft code, there is also debugging support. It's possible to open a .NET Core project from within Visual Studio Code and get end-to-end execution and debugging functionality[28].

The success of Visual Studio Code speaks volumes for its features and functionality. Although it has been officially released for just three years (it left public preview in April 2016), it has quickly become one of the top editors in terms of popularity, competing with Sublime Text, Atom, and UltraEdit for the top spot[28].

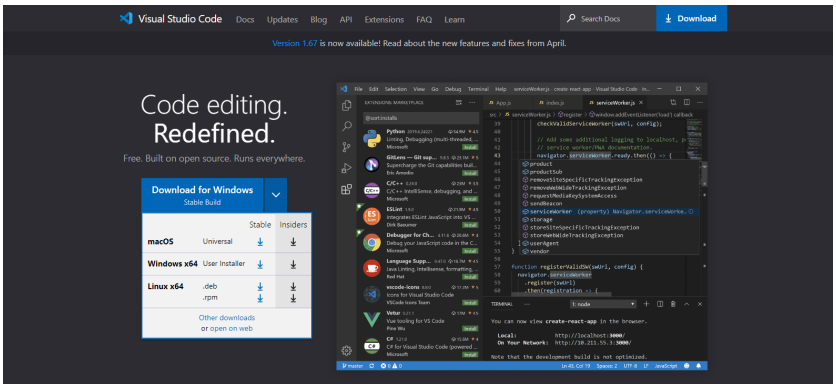



Figure 3.16: *Home page for Visual Studio Code*

### 3.4.2 Python

Python is a strong, procedural, object-oriented, functional language crafted in the late 1980s by **Guido Van Rossum**. The language is named after Monty Python, a comedy group. The language is currently being used in diverse application domains. These include software development, web development, Desktop GUI development, education, and scientific applications.

So, it spans almost all the facets of development. Its popularity is primarily owing to its simplicity and robustness, though there are many other factors too which are discussed in the chapters that follow[29].

There are many third party modules for accomplishing the above tasks. For example Django, an immensely popular Web framework dedicated to clean and fast development, is developed on Python. This, along with the support for HTML, E-mails, FTP, etc., makes it a good choice for web development[29].

Third party libraries are also available for software development. One of the most common examples is Scions, which is used for build controls. When joined with the inbuilt features and support, Python also works miracles for GUI development and for developing mobile applications, e.g., **Kivy** is used for developing multi-touch applications[29].

Python also finds its applications in scientific analysis. **SciPy** is used for Engineering and Mathematics, and **IPython** is used for parallel computing. Those of you working in statistics and machine learning would find some of these libraries extremely useful and easy to use. **SciPy** provides **MATLABMATLABMATLAB** like features and can be used for processing multidimensional arrays. Figure 1.1 summarizes the above discussion[29].

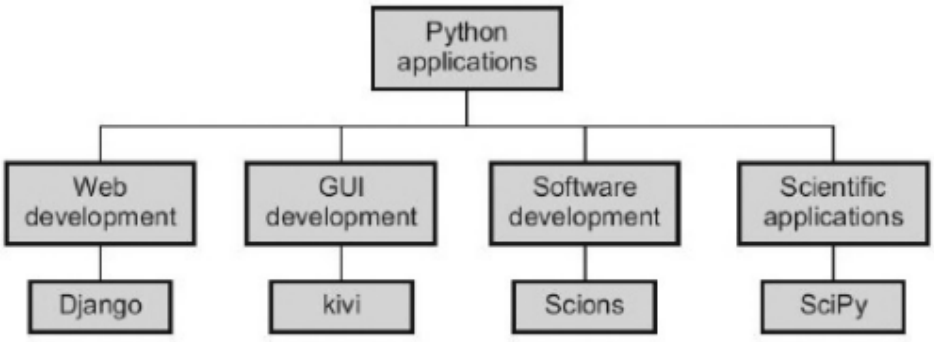



Figure 3.17: *Some of the applications of Python*

### 3.4.3 Jupyter Notebook

Jupyter is a tool that allows data scientists to record their complete analysis process, much in the same way other scientists use a lab notebook to record tests, progress, results, and conclusions[30].

The Jupyter product was originally developed as part of the IPython project. The IPython project was used to provide interactive online access to Python. Over time it became useful to interact with other data analysis tools, such as R, in the same manner. With this split from Python, the tool grew into its current manifestation of Jupyter. IPython is still an active tool that's available for use. The name Jupyter itself is derived from the combination of Julia, Python, and R. Jupyter is available as a web application from a number of places. It can also be used locally over a wide variety of installations[30].

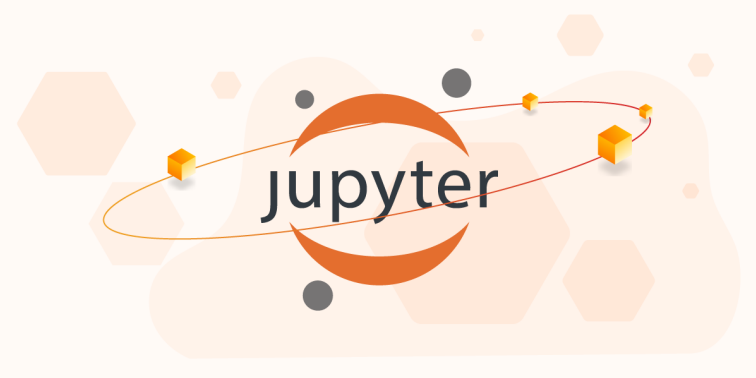



Figure 3.18: *The Jupyter Notebook logo*

Big data is the topic on everyone's mind. I thought it would be good to see what can be done with big data in Jupyter. An up-and-coming language for dealing with large datasets is Spark. Spark is an open source big data processing framework. Spark can run over Hadoop, in the cloud, or standalone. We can use Spark coding in Jupyter much like the other languages we have seen[30].

### 3.4.4 Tensorflow

TensorFlow is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications[31].

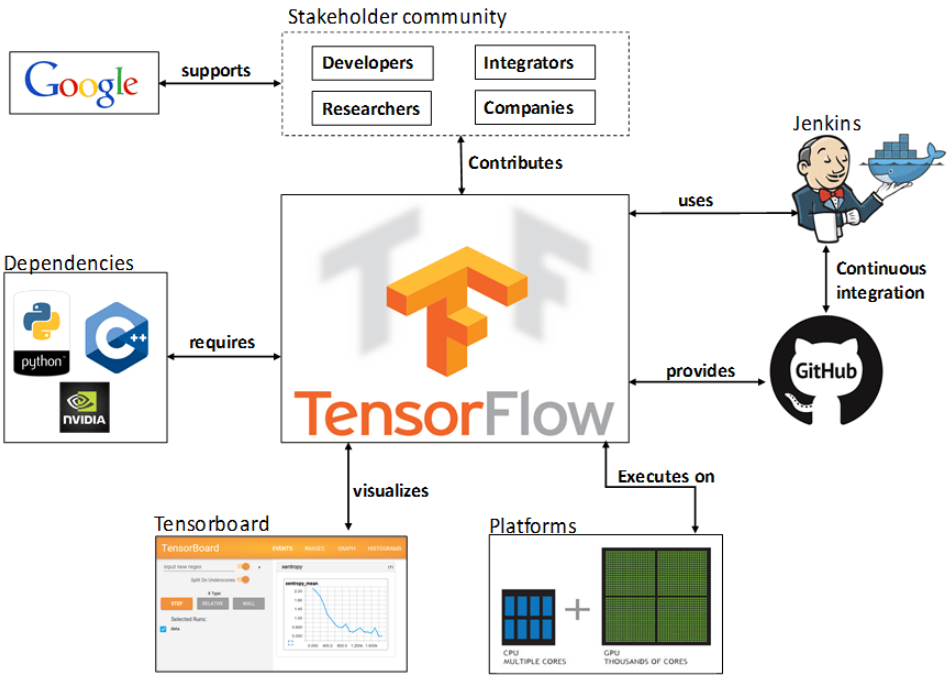



Figure 3.19: *TensorFlow*

- **Advantages :**
  - Supported by Google.
  - A very large community.
  - Multi-GPU support.
- **Disadvantages :**
  - Slower than other frameworks in many benchmarks.
  - Although Tensorflow is catching up.
  - Theano still outperforms RNN support.

### 3.4.5 Libraries

For those who are less familiar with the Python data ecosystem and the libraries, I will give a brief overview of some of them :

#### Numpy

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more [32].

#### Pandas

is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language , pandas provide high-level data structures and functions designed to make working with structured or tabular data fast, easy, and expressive. Since its emergence in 2010, it has helped enable Python to be a powerful and productive data analysis environment. [33].

#### sklearn

Scikit-learn (sklearn) is a free machine learning library for Python that may be used for both unsupervised and supervised learning. It also includes a number of utilities for model fitting, data preparation, model selection, and model assessment.[34] Scikit-learn features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

## 3.5 Conclusion

In this chapter, we reviewed the data and how we obtained it, and we touched on the method of collecting it and processing it. Then prepare it for the learning stage and finally evaluation. To facilitate a deeper level of comprehension, we have included several visualization charts directly connected to the data. In the end, we showed the working environment, as well as the most significant technologies and libraries that were used in the construction of the project code.

In the next chapter, we will aim to show the results and evaluations obtained from each model, as well as compare these models to one another in an attempt to get the best one of them.

# Chapter 4

# Implementation and Results

## 4.1 Introduction

The simulation experiment is the primary topic of this last chapter. The major purpose of this study is to assess the performance of the methods that have been provided as a potential solution for the deployment of a solar radiation prediction system. During the period of learning and testing, it is vital to examine and evaluate how well each of the proposed approaches (SVR - LR - RF) is performing.

The primary criteria for efficiency are developed based on two fundamental points: specification tests, which verify that the program is doing the work for which it was built, and performance tests, which will be used to determine how efficiently this work is performed. Specification tests are developed to ensure that the program is doing the work for which it was built.

## 4.2 Estimation Of Error

In order to guarantee that the analysis is as accurate as is practically possible, all of the tests should be carried out under the exact same conditions. Each experiment was carried out on the same system so that there would be no variation in the results of the performance research. A total of three times went through each and every test.

This is due to the fact that the findings of several simulation runs were contradictory with one another. It was important to apply mathematical tools in both directions in order to acquire the performance error of each prediction and to enhance the assessment of each parameter. This was accomplished by using the tools in both ways simultaneously. forms of study that are common in this types of research:

### 4.2.1 Mean Absolute Error (MAE)

The absolute mean error is a measure of variance between expectations and observations that is often employed

$$\mathbf{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y - y_i| \tag{4.1}$$

- n = The number of errors
- y = Measured solar irradiation
- $y_i$ = Solar irradiation estimated by a neural network.

### 4.2.2 Mean Square Error (MSE)

When comparing various estimators, the mean square error is useful, especially when one of them is biased.

$$\mathbf{MSE} = \frac{1}{n}\sum_{j=1}^{n}(y - y_i)^2 \tag{4.2}$$

- n = The number of errors
- y = Measured solar irradiation
- $y_i$ = Solar irradiation estimated by a neural network.

### 4.2.3 Root Mean Square Error (RMSE)

The Normalized Root Mean Square Error (NRMSE) The RMSE is a measure of the anticipated values fluctuation around the measured values. The better the model, the lower the value.

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y - y_i)^2} \tag{4.3}$$

- n = The number of errors
- y = Measured solar irradiation
- $y_i$ = Solar irradiation estimated by a neural network.

### 4.2.4 Normalized Root Mean Square Error (NRMSE)

The RMSE makes it easier to compare models with different scales. the normalised RMSE (NRMSE), which connects the RMSE to the variable's observed range.

$$\mathbf{NRMSE} = \frac{\mathbf{RMSE}}{\bar{y}} = \tag{4.4}$$

$$\mathbf{NRMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\log\frac{\hat{y}+1}{y+1})^2} \tag{4.5}$$

### 4.2.5 Coefficient Of Determination ($R^2$)

The correlation coefficient indicates how closely the expected and actual values are related. Clearly, a correlation coefficient with a value that is more equivalent to the unit signifies a better forecast.

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{j=1}^{n}(y - y_i)^2}{\frac{1}{n}\sum_{j=1}^{n} y} \tag{4.6}$$

## 4.3 Proposed System

### 4.3.1 Introduction

All algorithms based on machine learning follow a predictive model that estimates a certain type of data with high accuracy. A large data set is essential for the learning algorithm to understand the behavior of the system. The first step for machine learning is data acquisition. The collected data were shared by various interested parties and summarized in useful information. The steps included in this process are data purification and data delimitation. The data were separated into two disjoint sets, training, testing . The training dataset was used for model training and testing. The dataset was used for model optimization and evaluation [8].

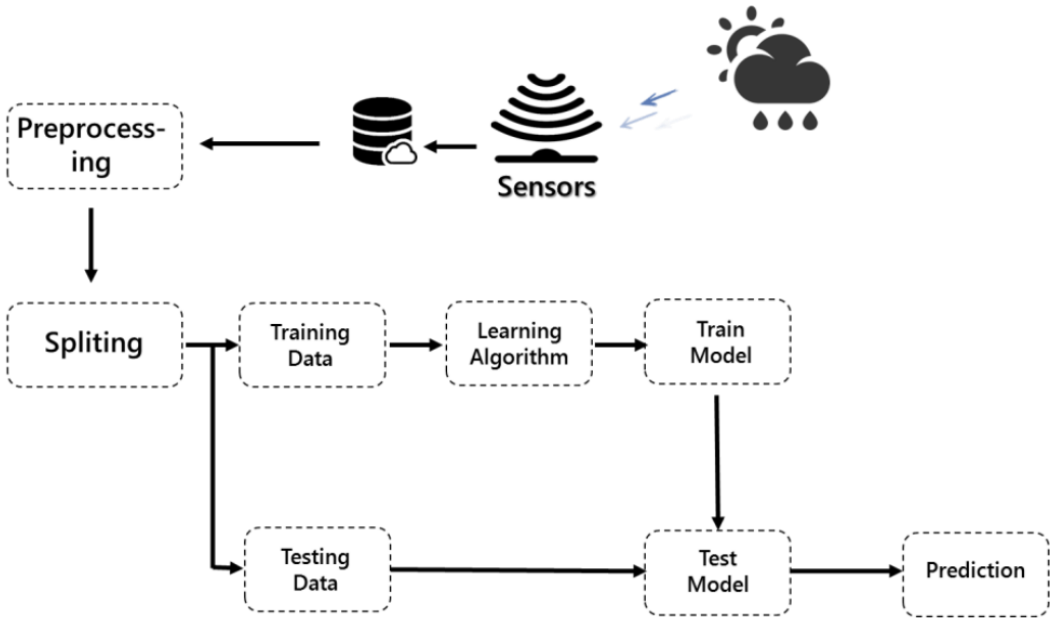



Figure 4.1: *From weather to a prediction model diagram.*

**Training Dataset:** The sample of data used to fit the model.

**Testing Dataset:** The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

**Dataset:** A dataset consists of about two components, the two components are rows and columns. In addition, a main feature of a record is that it is organized in such a way that each row contains an observation.

**Machine learning Algorithm:** An "algorithm" in machine learning is a procedure that is run on data to create a machine learning "model".

**Machine Learning Model:** A "model" in machine learning is the output of a machine learning algorithm run on data.

## 4.4 Simulation Results

### 4.4.1 Part One

#### Support Vector Regression SVR

We applied SVR in 4 different Kernels:

| Cases | Kernel |
|---|---|
| C1 | linear |
| C2 | poly |
| C3 | rbf |
| C4 | sigmoid |

Table 4.1: The SVR Kernels

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| C1 | 123.16 | 23601 | 153.62 | 0.77 | 0.14 | 123.58 | 23708 | 153.97 | 0.76 | 0.14 |
| C2 | 118.86 | 29423 | 171.53 | 0.71 | 0.16 | 117.45 | 28868 | 169.90 | 0.71 | 0.16 |
| **C3** | 73.70 | 14098 | 118.73 | 0.86 | 0.11 | 74.03 | 14281 | 119.50 | 0.86 | 0.11 |
| C4 | 174.35 | 49158 | 221.71 | 0.52 | 0.20 | 174.15 | 48863 | 221.05 | 0.52 | 0.20 |

Table 4.2: The SVR results for the testing and learning phase

Then we compared the real outputs with the outputs predicted by the SVR throughout the learning and testing phase. As far as the assessment criteria were concerned, we had the most successful outcome in the $3^{rd}$ case ($R^2$ =0.86, NRMSE=0.11, MAE = 74.03). The plots of the case are provided down below (Fig : 4.2).

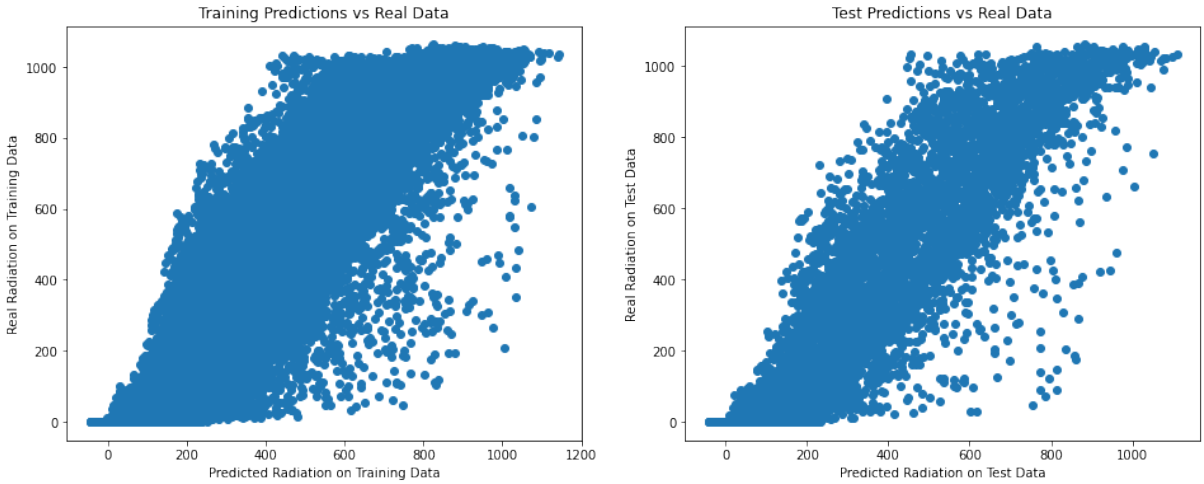



Figure 4.2: *SVR Training and Testing Predictions vs Real Data.*

### Linear Regression LR

We applied Baseline LR:

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| **LR** | 124.55 | 22968 | 151.55 | 0.77 | 0.14 | 124.67 | 23109 | 152.01 | 0.77 | 0.14 |

Table 4.3: The LR results for the testing and learning phase

As shown in Table(4.3) the LR model achieved ($R^2$ =0.77, NRMSE=0.14, MAE = 124.67 ). The plots of the results are provided down below (4.3 .

### Random Forest RF

We applied RF model:

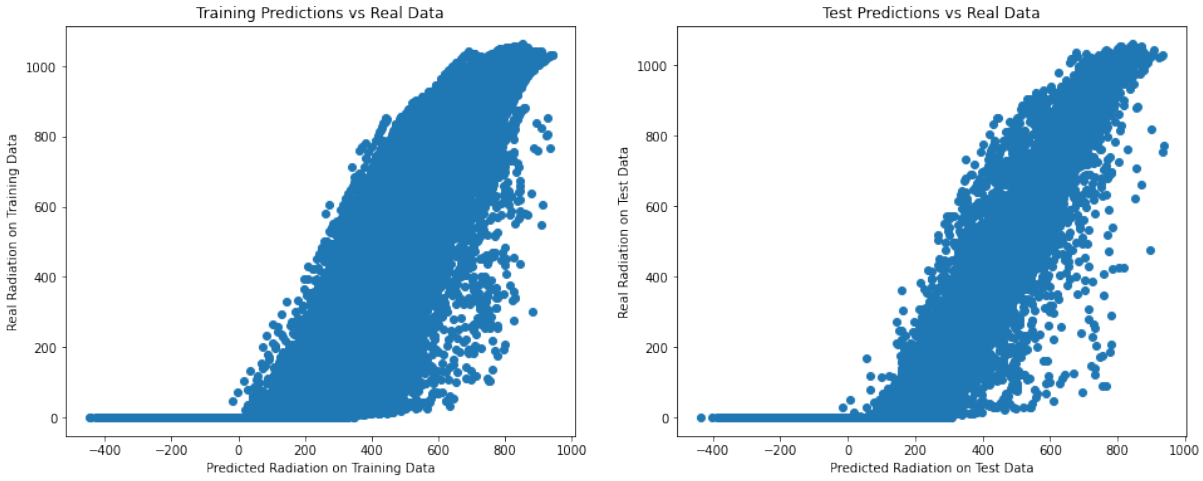



Figure 4.3: *LR Training and Testing Predictions vs Real Data.*

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| **RF** | 7.5 | 380.53 | 19.50 | 0.99 | 0.01 | 19.03 | 2194 | 46.85 | 0.97 | 0.04 |

Table 4.4: The LR results for the testing and learning phase

In Table(4.4) the RF model achieved ($R^2$ =0.97, NRMSE=0.04, MAE = 19.03 ). The plots of the results are provided down below (4.4 .

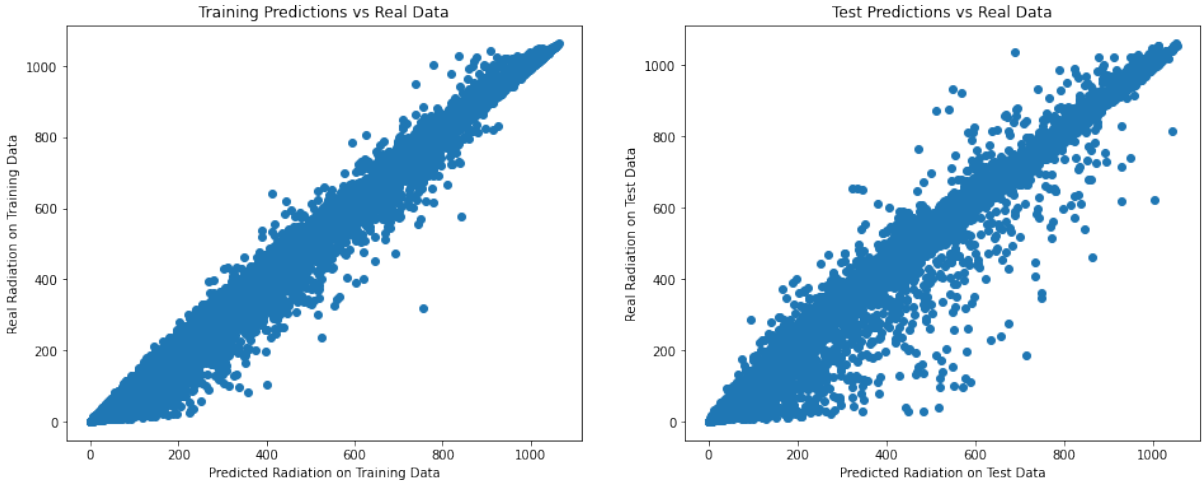



Figure 4.4: *RF Training and Testing Predictions vs Real Data.*

**Comparison of Results**

The table that follows (4.5) presents a comparison between the accuracy of the Training set and the accuracy of the Testing set using the Nevada Dataset.

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| SVR | 73.70 | 14098 | 118.73 | 0.86 | 0.11 | 74.03 | 14281 | 119.50 | 0.86 | 0.11 |
| LR | 124.55 | 22968 | 151.55 | 0.77 | 0.14 | 124.67 | 23109 | 152.01 | 0.77 | 0.14 |
| **RF** | 7.5 | 380.53 | 19.50 | 0.99 | 0.01 | 19.03 | 2194 | 46.85 | 0.97 | 0.04 |

Table 4.5: the performance evaluation for different techniques.

Through the results obtained and shown in the previous Table (4.5), we find that:

- The SVR model had similar results in both phases learning and testing. The results were good thanks to its ability to deal with non-linearity.
- On the contrary the LR model had less performance than SVR and RF because the model could not handle the zeros and negative values, for that the results were not that much satisfying.
- The RF model achieved the best results between all the models, in both phases. And that because of its ability to learn and avoid overfitting.

So in this part we can say that the RF model is the most satisfying model in terms of performance.

### 4.4.2 Part Two

**Support Vector Regression SVR**

We applied Support Vector Regression because of its ability to handle non-linearity in the data. The performance of the SVR depends on the selection of an appropriate kernel function and parameters. In our work, we used four distinct SVR kernel functions: Linear Kernel, Sigmoid Kernel, Polynomial Kernel, and the Radial Basis Function (RBF). The SVM uses the kernel function to transform the data from the input space to the high-dimensional feature space.The hyperparameters of the SVR technique in 4 learning and testing cases:

| Cases | Kernel | Degree | Gamma | Verbose | C |
|---|---|---|---|---|---|
| C1 | linear | Ignored | scale | True | 1 |
| C2 | poly | 1 | scale | True | 1 |
| C3 | rbf | Ignored | scale | True | 3 |
| C4 | sigmoid | Ignored | scale | True | 1 |

Table 4.6: The SVR hyperparameters

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| C1 | 244.64 | 118994.71 | 344.95 | -0.027 | 0.267 | 213.69 | 67928.79 | 260.63 | 0.202 | 0.219 |
| C2 | 250.14 | 94894.96 | 308.05 | 0.180 | 0.238 | 206.73 | 63940.11 | 252.86 | 0.249 | 0.213 |
| C3 | 237.46 | 82764.31 | 287.68 | 0.28 | 0.223 | 208.06 | 68092.06 | 260.94 | 0.02 | 0.201 |
| C4 | 253.19 | 91287.92 | 302.13 | 0.211 | 0.234 | 205.24 | 60091.35 | 245.13 | 0.294 | 0.206 |

Table 4.7: The SVR results for the testing and learning phase

After that, we compared the real outputs with the outputs predicted by the SVR throughout the learning and testing phase. As far as the assessment criteria were concerned, we had the most successful outcome in the $4^{th}$ case ($R^2$ =0.294, NRMSE=0.206, MAE = 205.24). The plots of the case are provided down below.

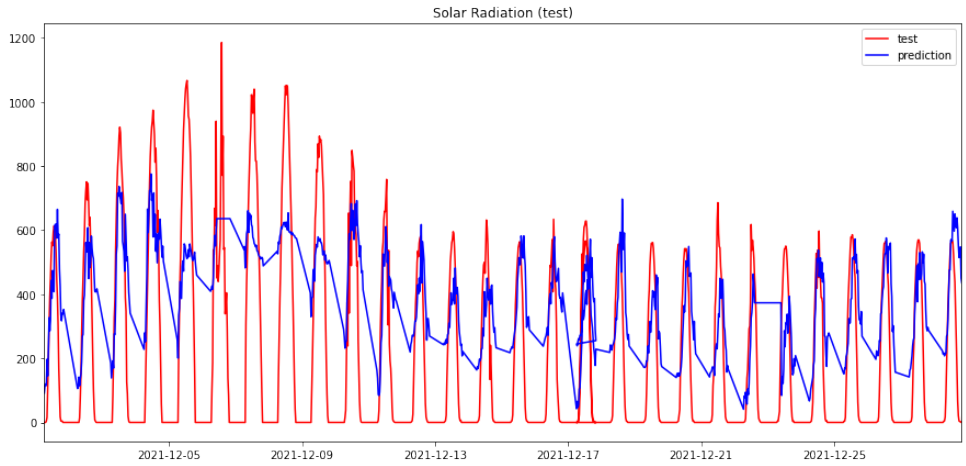



Figure 4.5: *Actual and calculated outputs for the SVR test.*

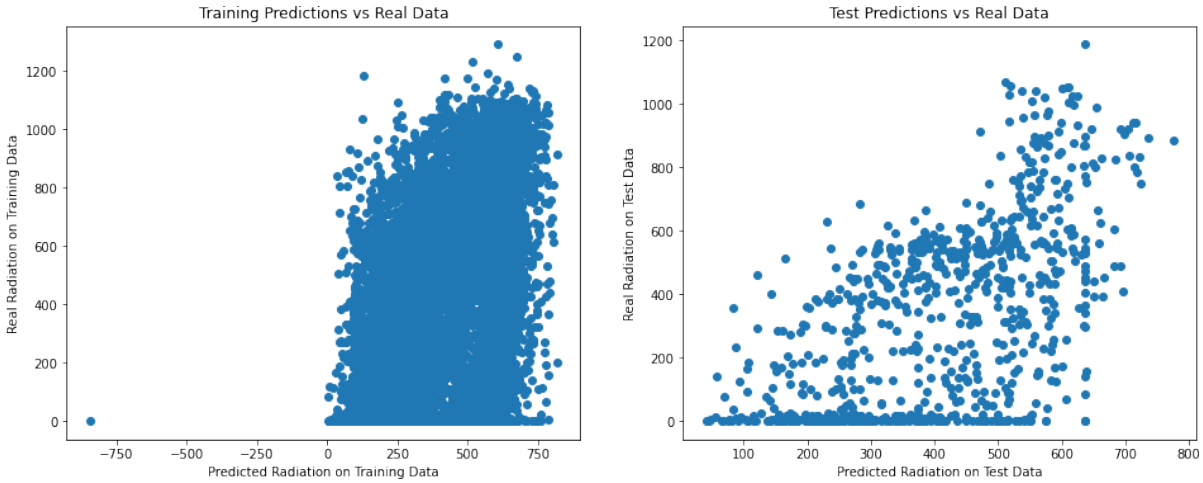



Figure 4.6: *SVR Training and Testing Predictions vs Real Data.*

### Linear Regression LR

The hyperparameters of the LR technique in 4 learning and testing cases:

| Cases | Fit_intercept | copy_X | n_jobs | Positive |
|---|---|---|---|---|
| C1 | True | True | none | False |
| C2 | False | True | none | False |
| C3 | True | True | 2 | True |
| C4 | True | False | 1 | False |

Table 4.8: The LR hyperparameters

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| C1 | 255.99 | 90330.91 | 344.95 | -0.027 | 0.267 | 213.69 | 67928.79 | 260.63 | 0.202 | 0.219 |
| C2 | 430.40 | 261698.44 | 511.56 | -1.259 | 0.396 | 326.82 | 152696.30 | 390.76 | -0.791 | 0.329 |
| C3 | 256.46 | 90483.04 | 300.80 | 0.218 | 0.233 | 217.13 | 66931.97 | 258.71 | 0.214 | 0.218 |
| C4 | 255.99 | 90330.91 | 300.55 | 0.219 | 0.233 | 207.76 | 59618.87 | 244.16 | 0.300 | 0.205 |

Table 4.9: The LR results for the testing and learning phase

Then we compared the real outputs with the outputs predicted by the LR throughout the learning and testing phase. As far as the assessment criteria were concerned, we had the most successful outcome in the $4^{th}$ case ($R^2$ =0.30, NRMSE=0.205, MAE = 207.76). The plots of the case are provided down below.

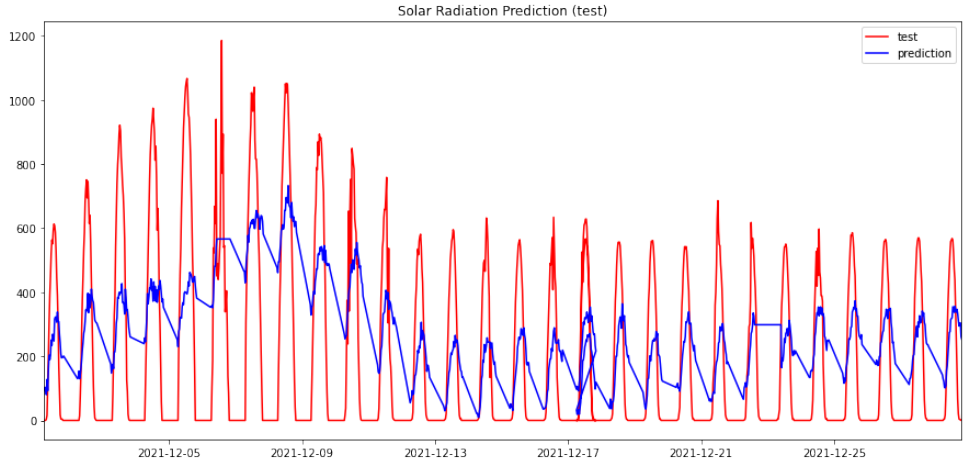



Figure 4.7: *.Actual and calculated outputs for the LR test.*

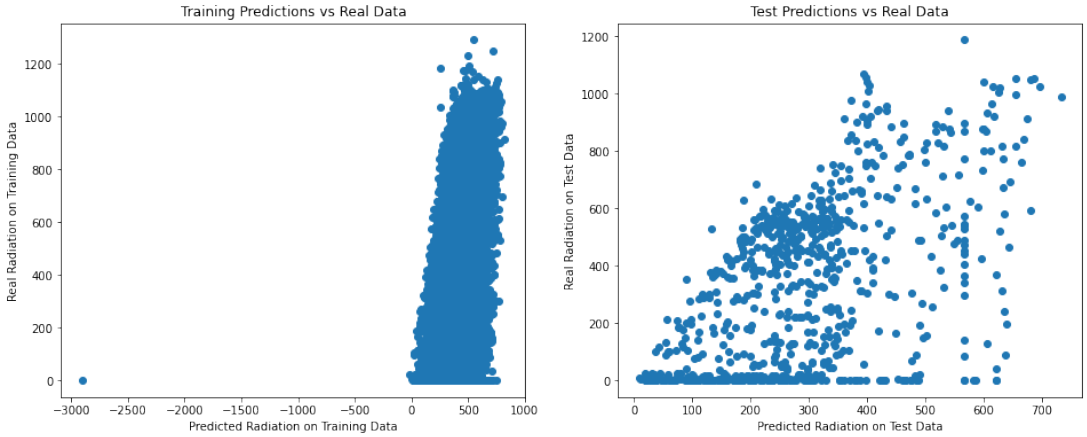



Figure 4.8: *LR Training and Testing Predictions vs Real Data.*

### Random Forest RF

The hyperparameters of the RF technique in 4 learning and testing cases:

| Cases | n_estimators | random_state | N_jobs | warm_start |
|---|---|---|---|---|
| C1 | 50 | 2 | 2 | True |
| C2 | 20 | 2 | 2 | True |
| C3 | 20 | 0 | 2 | False |
| C4 | 20 | None | None | False |

Table 4.10: The RF hyperparameters

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| C1 | 72.54 | 9464.94 | 97.28 | 0.918 | 0.075 | 187.56 | 56108.96 | 236.87 | 0.341 | 0.199 |
| C2 | 74.41 | 10484.71 | 102.39 | 0.909 | 0.079 | 186.99 | 55639.38 | 235.88 | 0.347 | 0.198 |
| C3 | 74.07 | 10505.27 | 102.49 | 0.909 | 0.079 | 188.42 | 56733.68 | 238.18 | 0.334 | 0.2 |
| C4 | 73.70 | 10261.79 | 101.30 | 0.911 | 0.078 | 184.46 | 56686.43 | 238.089 | 0.334 | 0.2 |

Table 4.11: The RF results for the testing and learning phase

Then we compared the real outputs with the outputs predicted by the RF throughout the learning and testing phase. As far as the assessment criteria were concerned, we had the most successful outcome in the $4^{th}$ case ($R^2$ =0.334, NRMSE=0.2, MAE = 184.46). The plots of the case are provided down below.

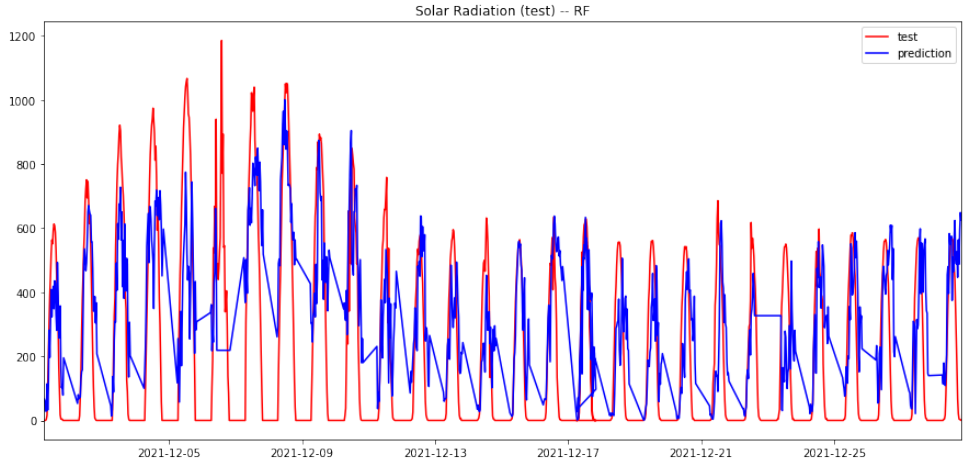



Figure 4.9: *.Actual and calculated outputs for the RF test.*

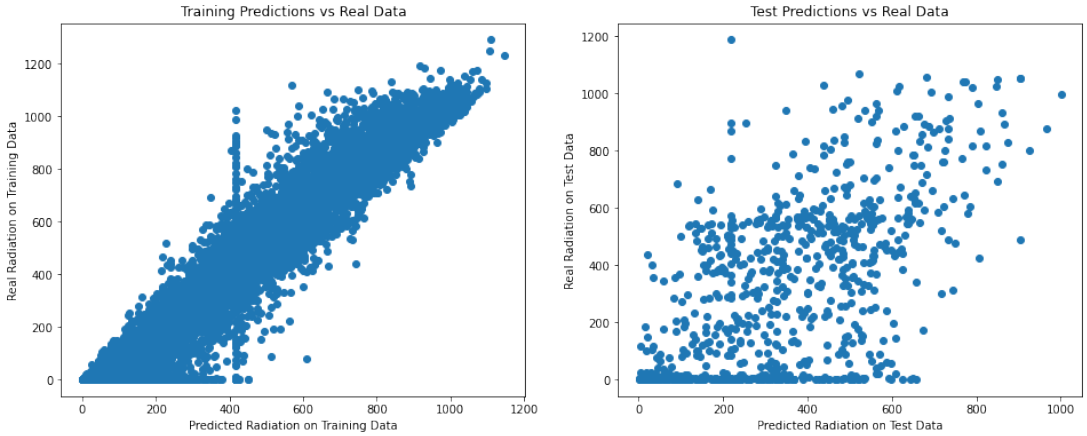



Figure 4.10: *RF Training and Testing Predictions vs Real Data.*

## 4.5 Comparison of Results

The table that follows (4.12) presents a comparison between the accuracy of the Training set and the accuracy of the Testing set using the same Dataset.

| | Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| SVR | 253.19 | 91287.92 | 302.13 | 0.211 | 0.234 | 205.24 | 60091.35 | 245.13 | 0.294 | 0.206 |
| LR | 255.99 | 90330.91 | 300.55 | 0.219 | 0.233 | 207.76 | 59618.87 | 244.16 | 0.300 | 0.205 |
| **RF** | 73.70 | 10261.79 | 101.30 | 0.911 | 0.078 | 184.46 | 56686.43 | 238.089 | 0.334 | 0.2 |

Table 4.12: the performance evaluation for different techniques.

Through the results obtained and shown in the previous Table (4.12), we find that:

- In the learning phase, the RF model performed much better than the other models. Where it received the best R2 score and MAE values. On the other hand, the SVR and LR models had a very low R2 score even the MAE value is really high comparing to RF.
- In the testing phase, we see that the values of all metrics have become nearly equivalent, with a slight outperformance of the RF model.

Since the RF model is the most outperforming model, we give it another test. And in this test we changed the learning set and it becomes the whole dataset instead of just part of it, and we specify the testing set to be just one month. The results obtained are shown in Table (4.13). And we see that the performance of the RF model raised after we rise the size of learning set.

| Learning | | | | | Testing | | | | |
|---|---|---|---|---|---|---|---|---|---|
| MAE | MSE | RMSE | R2 | NRMSE | MAE | MSE | RMSE | R2 | NRMSE |
| 73.06 | 10199 | 100 | 0.91 | 0.07 | 60.56 | 6994 | 83.63 | 0.92 | 0.07 |

Table 4.13: RF model performance evaluation for the second test.

From Part One & Two we conclude that:

- RF model outperformed the SVR and LR models.
- The performance of the RF model depends on how well it is trained and on the quality of the data that is used. With additional historical data, the model performance will improve.
- The ability of using different Machine Learning techniques to predict solar radiation and achieve very satisfactory results.

## 4.6 Conclusion

In this chapter, we provide detailed results for each model, as well as comparisons between the current models. Finally, some conclusions drawn from our analysis of the results obtained.

# *Conclusion*

# General Conclusion

In this thesis, we tried to provide a solution to the problem of predicting solar radiation using several methods belonging to the field of machine learning. The main objective of this project is to compare different machine learning techniques in terms of learning and performance in order to solve the problem of solar radiation prediction. This is done by collecting weather data a (Solar radiation, Temperature, Humidity and Barometric pressure) about a semi-desert region (El Hadjira & Nevada in our case) and processing this information and then preparing it as a dataset for different technologies.

In this regard, we have presented three different techniques, which are SVR, LR, and RF to apply them on a short term solar radiation prediction . After teaching and evaluating them, we get generally weak results for all models, especially during the testing phase, and this is due to the weak learning of models because of the small dataset which they learned from.In the first dataset (Nevada), the RF model the best performance and achieved the highest R2 score, which was 0.97. With the second Dataset (El Hadjira) and during the testing phase, the R2 scores obtained by all of the models were quite comparable, with values hovering around 0.2 to 0.3. During the learning phase, the RF model produced satisfactory results and achieved the highest R2 score, which was 0.9.

In addition, different learning samples were used, the results showed that the proposed model required concrete training beforehand in order to give an acceptable predictive accuracy.

Throughout the time working on this project, we learned so much about artificial intelligence techniques specially machine learning models and how to apply these models to solar radiation prediction.

# Notation And Abbreviated Terms

**RE :** Renewable Energy
**DNI :** Direct Normal Irradiance
**DHI :** Diffuse Horizontal Irradiance
**GHI :** Global Horizontal Irradiance
**PV :** Photovoltaic
**DC :** direct current
**AC :** alternating current
**SR :** Solar radiance
**SI :** solar irradiance
**ML :** Machine Learning
**AI :** artificial intelligence
**VS Code :** Visual Studio Code
**CART :** Classification And Regression Trees
**SKTM :** Sharikate Kahraba Wa Taket Moutadjadida
**SVR :** Support Vector Regression
**RF :** Random forest
**LR :** Linear Regression
**ARMA :** The autoregressive moving average model
**ARIMA :**The autoregressive integrated moving average
**MLP :** Multilayer Perceptron
**CNN :** Convolutional Neural Network
**DNN :** Deep neural network
**LSTM :** Long Short-Term Memory
**MAE :** Mean Absolute Error
**MSE :** Mean Square Error
**RMSE :** Root Mean Square Error
**NRMSE :** Normalized Root Mean Square Error
$R^2$ **:** Coefficient Of Determination
**GPU:** graphics processing unit

# *Bibliography*

# Bibliography


[1] Pedro Francisco Jiménez Pérez and Llanos Mora López. "Data mining models for short-term solar radiation prediction and forecast-based assessment of photovoltaic facilities". PhD thesis. Universidad de Málaga, 2016.

[2] NL Panwar, SC Kaushik, and Surendra Kothari. "Role of renewable energy sources in environmental protection: A review". In: *Renewable and sustainable energy reviews* 15.3 (2011), pp. 1513–1524.

[3] Ram Avtar et al. "Exploring Renewable Energy Resources Using Remote Sensing and GIS—A Review". In: *Resources* (Aug. 2019).

[4] Michael Boxwell. *The Solar Electricity Handbook-2017 Edition: A simple, practical guide to solar energy–designing and installing solar photovoltaic systems.* Greenstream Publishing, 2017.

[5] Paul Hersch and Kenneth Zweibel. *Basic photovoltaic principles and methods.* Tech. rep. Solar Energy Research Inst., Golden, CO (USA), 1982.

[6] Lila Croci. "Gestion de l'énergie dans un systeme multi-sources photovoltaique et éolien avec stockage hybride batteries/supercondensateurs". PhD thesis. Université de Poitiers, 2013.

[7] Marcelo Gradella Villalva, Jonas Rafael Gazoli, and Ernesto Ruppert Filho. "Comprehensive approach to modeling and simulation of photovoltaic arrays". In: *IEEE Transactions on power electronics* 24.5 (2009), pp. 1198–1208.

[8] Oussama Bouguerra and Oussama Benslimane. "SOLAR RADIATION PREDICTION USING MACHINE LEARNING". PhD thesis. Univ M'sila, 2020.

[9] Alberto Eduardo Gabás Royo. "Solar irradiance forecasting using neural networks". MA thesis. Universitat Politècnica de Catalunya, 2019.

[10] Muriele Souza et al. "Determination of Diffused Irradiation from Horizontal Global Irradiation - Study for the City of Curitiba". In: *Brazilian Archives of Biology and Technology* 62 (Jan. 2019). DOI: 10.1590/1678-4324-smart-2019190014.

[11] Kesh B Pun. "Short term forecasting of solar power with machine learning and time series techniques". PhD thesis. Wichita State University, 2020.

[12] Sobrina Sobri, Sam Koohi-Kamali, and Nasrudin Abd Rahim. "Solar photovoltaic generation forecasting methods: A review". In: *Energy conversion and management* 156 (2018), pp. 459–497.

[13] Giacomo Sbrana and Andrea Silvestrini. "Random switching exponential smoothing and inventory forecasting". In: *International Journal of Production Economics* 156 (2014), pp. 283–294.

[14] Liljana Ferbar Tratar and Ervin Strmčnik. "The comparison of Holt–Winters method and Multiple regression method: A case study". In: *Energy* 109 (2016), pp. 266–276.

[15] Mathieu David et al. "Probabilistic forecasting of the solar irradiance with recursive ARMA and GARCH models". In: *Solar Energy* 133 (2016).

[16] Razin Ahmed et al. "A review and evaluation of the state-of-the-art in PV solar power forecasting: Techniques and optimization". In: *Renewable and Sustainable Energy Reviews* 124 (2020), p. 109792.

[17] Rina Haiges et al. "Forecasting Electricity Generation Capacity in Malaysia: An Auto Regressive Integrated Moving Average Approach". In: *Energy Procedia* 105 (2017), pp. 3471–3478. URL: https://www.sciencedirect.com/science/article/pii/S1876610217308639.

[18] Fei Wang et al. "Daily pattern prediction based classification modeling approach for day-ahead electricity price forecasting". In: *International Journal of Electrical Power & Energy Systems* 105 (2019), pp. 529–540. URL: https://www.sciencedirect.com/science/article/pii/S0142061518305301.

[19] Sajid Hussain and Ali Al Alili. "Day ahead hourly forecast of solar irradiance for Abu Dhabi, UAE". In: *2016 IEEE Smart Energy Grid Engineering (SEGE)*. IEEE. 2016, pp. 68–71.

[20] Ilhami Colak et al. "Multi-period prediction of solar radiation using ARMA and ARIMA models". In: *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*. IEEE. 2015, pp. 1045–1049.

[21] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[22] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

[23] Imane Jebli et al. "Prediction of solar energy guided by pearson correlation using machine learning". In: *Energy* 224 (2021), p. 120109.

[24] David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.

[25] Brownlee Jason. *Master Machine Learning Algorithms - Discover how they work and Implement Them From Scratch*. Machine Learning Mastery, 2016, p. 163.

[26] Souhaila Chahboun and Mohamed Maaroufi. "Performance Comparison of Support Vector Regression, Random Forest and Multiple Linear Regression to Forecast the Power of Photovoltaic Panels". In: *2021 9th International Renewable and Sustainable Energy Conference (IRSEC)*. IEEE. 2021, pp. 1–4.

[27] https://nsrdb.nrel.gov. In: ().

[28] Bruce Johnson. *Visual Studio Code: End-to-End Editing and Debugging Tools for Web Developers*. John Wiley & Sons, 2019.

[29] Harsh Bhasin. *Python Basics: A Self-teaching Introduction*. Stylus Publishing, LLC, 2018.

[30] Dan Toomey. *Learning Jupyter*. Packt Publishing Ltd, 2016.

[31] URL: `https://www.tensorflow.org/`.

[32] *NumPy documentation*. URL: `https://numpy.org/doc/stable/`.

[33] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[34] URL: `www.scikit-learn.org/`.