Toward Predicting Global Seismicity of the Earth using Machine Learning Techniques and Solar Activity Data

Viacheslav Shkuratskyy

Submitted in accordance with the requirements for the degree of Master by Research

York St John University

School of Science, Technology and Health

June 2022

The candidate confirms that the work submitted is their own, except where work which has formed part of jointly authored publications has been included. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

The work in Chapters 1 and 2 of the thesis has appeared in publication as follows:

Handbook of Research on AI Methods and Applications in Computer Engineering. Chapter 11 “The Application of Machine Learning for Predicting Global Seismicity”. DOI:10.4018/978-1-6684-6937-8.ch011. 2023 pages: 31

I was responsible for conducting the study, literature review, data collection, experiment, and writing. Dr Aminu Usman Responsible for Supervision, conceptualization, designing the study, and overseeing the project. Dr Mike O’Dea responsible for supervision, overseeing project, reviewing, and editing, and methodology.

Acknowledgments

This study was conducted by a group that included myself, Viacheslav Shkuratskyy, and my supervisory team. My supervisory team included my main supervisor Dr Aminu Usman, co-supervisor Dr Michael O'Dea, and co-supervisor Dr Adi Baumgartner.

First and foremost, I'd like to express my gratitude to my supervisory team for their unwavering support, direction, and supervision throughout the project. Dr Aminu Usman and Dr Michael O'Dea helped me with editing, recommendations, proofreading, offering insight, and verifying the quality of the work. Dr Adi Baumgartner offered assistance in the form of proofreading and idea generation.

I would also like to express my gratitude to the entire staff of York St John University's Computer Science course for their guidance and support throughout my degree.

Finally, I would like to express my gratitude to my family and friends for their encouragement and support in my decision to return to school, as well as their patience and generosity throughout my studies.

Abstract

An earthquake is one of the deadliest natural disasters. Forecasting an earthquake is a challenging task since natural causes such as movement of tectonic plates, volcanic eruptions, rainfall, and tidal stress all play an important part in earthquakes. Earthquakes can also be caused by human beings, such as mining, dams, nuclear bomb testing, etc.

Solar activity has also been suggested as a possible cause of earthquakes. Solar activity and earthquakes occur in different parts of the solar system, on the Sun's surface and the Earth's surface, separated by a huge distance. However, scientists have been trying to figure out if there are any links between these two seemingly unrelated occurrences since the 19th century.

In this study, four machine learning algorithms k-nearest neighbour, support vector regression, random forest regression, and Long Short-Term Memory network were applied to understand if there is a relationship between solar activity and earthquakes. The study employed three types of solar activity: sunspot number, solar wind, and solar flares, as well as worldwide earthquake frequencies that ranged in magnitude and depth.

The study's findings imply that the Long Short-Term Memory network model predicts earthquakes more accurately than other models. There's a chance that earthquakes are influenced by solar activity. Earthquakes with a magnitude less than 5.5 are more linked to solar activity than earthquakes with a magnitude equal to or higher than 5.5. Solar activity has a bigger impact on earthquakes of lower depths.

Table of Contents

List of Tables

List of Figures

Abbreviation Key

Table 1.1 Glossary of Meaning

Abbreviation	Meaning
Earthquake Magnitude	M
Sunspot number	SSN
Machine Learning	ML
Supervised learning	SL
K-Nearest Neighbours	KNN
Simple Linear Regression	SLR
Multiple Linear Regression	MLR
Support Vector Machine	SVM
Support Vector Regression	SVR
Logistic Regression	LR
Naïve Bayes	NB
Random Forest Regression	RFR
Principal Component Analysis	PCA
Neural networks	NN
Recurrent Neural Network	RNN
Long Short-Term Memory network	LSTM
R-squared	R^2
Mean absolute percentage error	MAPE
Mean absolute error	MAE
Mean squared error	MSE
Root mean squared error	RMSE
Normalised Mean absolute error	NMAE
Normalised Root mean squared error	NRMSE

1 Chapter One Introduction

This chapter discusses the study's background as well as the key motives for conducting it. The research hypotheses, as well as the research contributions, are all presented in this chapter.

1.1 Background

Since ancient times, cataclysmic disasters such as droughts, floods, earthquakes, volcanic eruptions, storms, and many other types of natural catastrophes, had a profound impact on humans at the cost of countless lives. These disasters are classified as natural disasters (Wirasinghe *et al.*, 2013). The most severe natural disaster in recent history was the flood of the Yangtze–Huai River in China in the summer of 1931. Up to 25 million people were affected by the effects of this flood (National Flood Relief Commission, 1933); hence, it is considered the deadliest natural disaster since 1900, excluding epidemics and famines.

Although the China flood of 1931 was the most widescale disaster in terms of mortality, all other natural disasters also took a toll on human life. After the 2010 Haiti earthquake, the death toll reached more than 200,000 people (Daniell, Khazai and Wenzel, 2013).

The number of deaths from natural disasters may change depending on the type of disaster and the affected area. But, from the average point of view, around 40,000 people per year are killed by natural disasters. For example, Figure 1.1 shows the yearly average of global annual deaths from natural disasters between 1900 and 2010s. The graph was created based on data from (*OFDA/CRED International Disaster Data*, 2021).

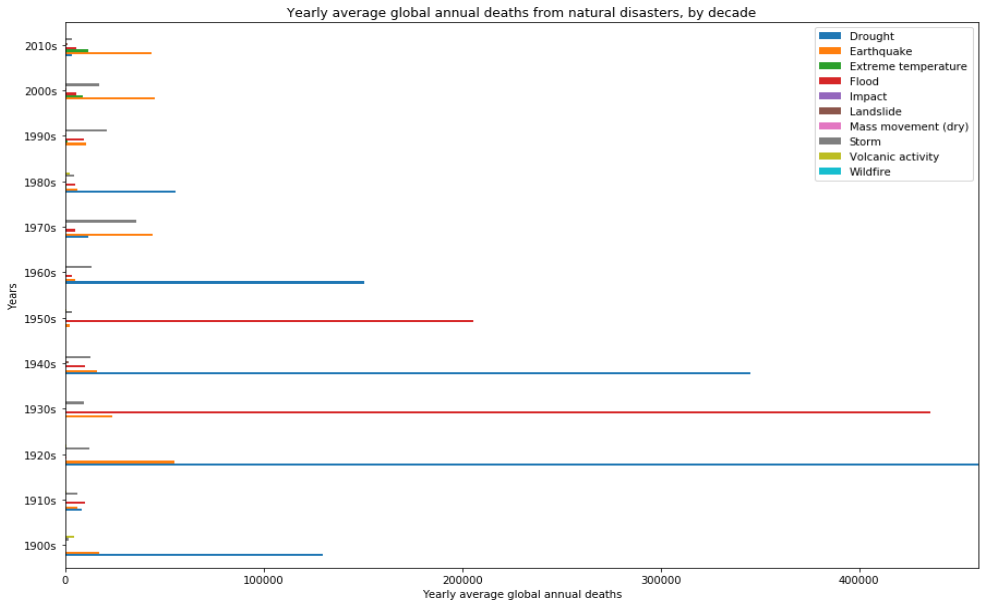


Figure 1.1 Yearly average global of annual deaths from natural disasters, by decade.

As seen in Figure 1.1, the three deadliest natural disasters are droughts, floods, and earthquakes. However, in the last few decades, the most dangerous natural disasters for people have been earthquakes, extreme temperatures, and floods. Even though the average global death toll from natural disasters in the 21st century is lower than in the previous century, the average death rate is still high.

Moreover, in addition to the loss of human lives, natural disasters also greatly influence human society and the economy. For example, Lesk, Rowhani and Ramankutty (2016) suggested that extreme natural disasters such as floods, droughts, and extreme temperatures drastically reduce cereal production. In addition, another study Shabnam (2014) provided a comprehensive review of the effect of natural disasters on macro- and microeconomic spheres.

Subsequently, understanding and exploring ways to understand and predict the occurrence of natural disasters might be useful in saving human lives, particularly in the area of anti-crisis services. Prediction of cataclysms, prediction of economic recession, and many others are also vital. Thus, the investigation of disaster management became crucial.

That is why numerous studies have been conducted to predict natural disasters and determine what factors influence each disaster. For example, several studies such as Wang *et al.* (2015), Muis *et al.*, (2018), Barnard *et al.* (2015) found that climate variability such as teleconnection pattens have a huge impact on floods throughout the world. Different studies established the relationship between solar activity and earth atmospheric–oceanic circulation mechanisms. Hassan *et al.* (2016) using the Markov chain method, found relationship between sunspots and the Pacific Ocean El Niño–Southern Oscillation. Gruzdev and Bezverkhnii (2018) used wavelet analysis to demonstrate the correlation between Central England temperature, the index of the North Atlantic Oscillation, and sunspot numbers during the solar cycle.

In contrast, Mohamed and El-Mahdy (2021) did not find a strong relationship between sunspot number and rainfall patterns over Eastern Africa and proposed that teleconnection pattens have an influence on rainfall.

However, most of the Earth's meteorological processes are localised and make good limited-area weather forecasts. Space weather is always global on the planetary scale (Koskinen *et al.*, 2001).

Further, the assumption that solar activities could influence Earth's natural disasters is not new. Back in 1853, the astronomer Wolf (1853) suggested sunspots might influence earthquake events. Since then, several studies, using statistical methods, showed the

correlation between solar activity and earthquakes. Odintsov *et al.* (2006) reported that seismic activity is related to the sunspot maximum during the solar cycle. Marchitelli *et al.* (2020) proved the correlation between solar activity and earthquakes with a magnitude of more than 5.6 on the Richter scale.

The modern solar activity data and natural disaster data, as well as worldwide data, which exponentially increase every year with improved or new technologies – contain a plethora of different parameters for solar but also natural disaster events. Reinse and colleagues stated that the International Data Corporation predicts an increase of the global dataset from 33 ZB in 2018 to 175 ZB by 2025 (Reinse, Gantz and Rydning, 2018). To work with such a huge amount of data, computer processing power must be faster but also algorithms more intelligent. There is a part of computer science that tries to achieve this goal by employing artificial intelligence. Studies focussing on the intelligence of animals (Thorndike, 2000) and plants (Calvo *et al.*, 2020) proved that one of the most crucial requirements for intelligence is learning. High intelligence is based on comprehensive learning and artificial intelligence, is not an exception. That is why machine learning is one of the most important and vital parts of artificial intelligence (Dunjko and Briegel, 2018).

One of the first occasions that machine learning was mentioned, was back in 1959, in the Samuel (1959) study. Samuel (1959) created a checkers programme, where two "machine-learning procedures" were used, and the study provided a start for the development of learning methods that would exceed average human abilities and solve real life problems. A quote from the original article describes machine learning: *"Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort."*

This checkers programme was one of the first programmes that used reinforcement learning. Most of the mathematical groundwork for reinforcement learning was laid by Richard Bellman. There are many other outstanding pioneers of machine learning, such as Karl Steinbuch, from Germany, who gave the start of neural networks with leading further research carried out by scientists such as Frank Rosenblatt, Charles Wightman, Mark I Perceptron, David Rumelhart, Geoffrey Hinton, and Terry Sejnowski (Trappenberg, 2020).

Nowadays, machine learning with regard to space weather has become more and more popular since the increase in the volume of data and the continuous development of computer hardware. The study by Bobra and Couvidat (2015) used the Supervised Learning model, the Support Vector Machine algorithm, to forecast solar flares. The study Liu *et al.* (2019) built the Long Short-Term Memory (LSTM) networks to predict solar flares, one LSTM network for each flare class.

So, increasing the processing power of computer systems in parallel with the growing amount of solar and climate data, together with implementing powerful data analysis techniques, will allow more accurate predictions of risk levels and threats of disasters, as well as the time and place of the disasters.

However, there is a lack of studies, that use Machine Learning techniques, which try to find the most appropriate method in the prediction of natural disasters using solar activity. This can be attributed to the fact that solar and natural disasters data are often raw and unstructured, which makes them due to their volume difficult to analyse and challenging to process.

1.2 Motivation

Machine learning is increasingly being used nowadays. With a variety of different algorithms, the most difficult tasks can be solved. Despite the fact that the percentage of machine learning techniques used for space and natural disaster studies increases daily, there are only a few studies that use machine learning techniques for trying to predict earthquakes based on solar activity events.

It is still not completely confirmed that solar activity events affect natural disasters. However, in accordance with Love and Thomas (2013) the statement that solar activity events have an impact on natural disasters cannot be rejected, even though they did not find a strong correlation between solar activity and earthquakes.

That is why, there is a growing need to understand by using data and machine leaning algorithms – if solar activity can influence earthquake activity.

1.3 Problem Definition

At the first glance, solar activity and earthquakes seem to be unrelated events. However, some studies have consistently shown that there is a relationship between solar activity and earthquakes (Gribbin, 1971; Han, 2004) Also, there are suggestions, that earthquakes depend on solar activity and the 11-year solar cycle (Odintsov, Ivanov-Kholodnyi and Georgieva, 2007).

However, it is still not clear of how and to what extent solar activity affects earthquakes. Moreover, all above theories are not yet sufficiently developed to allow reliable predictions of the likelihood of future earthquakes. Therefore, this study attempts to build a model that tests the relationship between solar activity and earthquakes using machine learning techniques.

1.4 Research Questions

The research question is: “How effective is machine learning in predicting of earthquakes based on solar activity?”

To attempt the above question, sub questions need to be addressed:

- What characteristics and types of earthquakes and solar activity should be used in earthquake prediction?
- How to evaluate the efficacy and effectiveness of the machine learning algorithms used in the study to ensure the efficacy of the analysis?
- Which machine learning algorithms should be chosen to answer the research question, and which one would give the highest accuracy among those chosen?
- Do solar activity events have the same impact on different types of earthquakes, and what other factors are important, such as time delay?

1.5 Research Hypotheses

Null hypothesis:

“Solar activity events do not have any relationship to earthquake events and these two events are completely independent of each other”.

Alternative hypothesis:

“If solar activity events and earthquake events are related to each other, then earthquake events may change in one way or another due to changes in solar activities events.”

1.6 Research Contribution

The connection between solar activity and earthquakes isn't a novel one. However, only a few studies have used machine learning approaches to investigate this association. Because it is still unknown whether solar activity events have an impact on earthquakes. The statement that solar activity events have an impact on earthquakes, on the other hand, cannot be dismissed (Love and Thomas, 2013).

Using machine learning techniques, an attempt was made to uncover any probable links between two events: solar activity and earthquakes. So that it can serve as a foundation for future earthquake research. Furthermore, the study is based on seismology findings, which will assist them in employing machine learning approaches to support their results. The study given here is only a step towards earthquake prediction using machine learning techniques, but it has important long-term ramifications.

1.7 Thesis Outline

The thesis is structured as follows:

- The first chapter provides background information, motivation, and a characterization of the research problem. There are other research questions and hypotheses in this chapter.
- The second chapter discusses natural disasters, earthquakes, and solar activity occurrences, as well as machine learning methods. The chapter also discusses previous research on the link between solar activity and natural disasters. The chapter also gives related work over machine learning algorithms – solar activity, and machine learning – natural disasters.
- Testing the null hypothesis and the research methodology are presented in the third chapter. The data sources and data preparation are described in this chapter. The data preparation for the experiments is also shown in this chapter. The chapter describes how data is normalised, checked for linear and nonlinear relationships, and dimensionally reduced. Furthermore, the chapter defines a model measurement approach as well as machine learning algorithms.
- The fourth and fifth chapters contain a collection of the results.
- The discussion of the results is presented in the final chapter, which also contains a conclusion as well as a suggestion for further study.
- At the end, a bibliography and additional materials will be included.

2 Chapter Two Literature Review

This chapter explores the link between solar activity and natural disasters, particularly earthquakes. The chapter also discusses machine learning in general, as well as machine learning techniques that have been applied to natural disasters and solar activity studies. The chapter begins with an overview of natural disasters and how they are linked to processes both on and off the surface of the Earth. This is followed by a description of earthquakes and solar activity in order to gain a better understanding of these events. The chapter then goes on to explain solar activity events and how and if they are linked to earthquakes. In the chapter, various techniques and algorithms for machine learning were examined, and it was discussed how machine learning has been used in natural disasters and solar activity, as well as how frequently it is used with solar activity and earthquakes.

2.1 Natural disasters

The word "disaster" is defined in the Oxford English Dictionary (2021) as *"An event or occurrence of a ruinous or very distressing nature; a calamity; esp. a sudden accident or natural catastrophe that causes great damage or loss of life."* Based on the Wirasinghe *et al.*, (2013) study, there are two main classifications for all disasters. They are classified as

a) natural disasters – events that are natural;
b) human-made disasters – events that occurred as a result of human activity.

Natural disasters are grouped as biological, geophysical, or hydrological events, depending on their cause. The origin of disasters can range from terrestrial to extra-terrestrial events. Disasters are further categorised into many types and sub-types, including drought, flood, tornado, tsunami, etc. According to Wirasinghe *et al.*, (2013), natural disasters such as cyclones/hurricanes/storms, earthquakes, floods, fires, landslides, lightning strikes, meteorite impacts, tsunamis, and volcanic eruptions have an impact on a global level on Earth.

Also, the Earth's climate has large-scale climate irregularities. Feldstein and Franzke (2017) study has shown that anomalies in weather conditions in different parts of the world are linked to each other. As an example, usually dry regions endure floods, while normally wet regions can have severe droughts. These remote connections are called teleconnection patterns. The examples of the teleconnection patterns are the El Niño–Southern Oscillation (ENSO), North Atlantic Oscillation (NAO), Arctic Oscillation (AO), etc (Allan, 2006).

There are several studies that have found a link between natural disasters and different processes that occur on and off the Earth. Donat *et al.* (2010) showed that most storm days

in Central Europe were related to the NAO. Dewitte *et al.* (2012) demonstrated that ENSO impacts rainfall flooding in parts of South America. Thompson and Wallace (1998) found the link between warming air temperatures and AO. Studies carried out by Dätwyler *et al.* (2019) and Pararas-Carayannis and Zoll (2017) confirmed that the relationship between teleconnection patterns and earthquakes, volcanic eruptions, and tsunamis is not clear and should be investigated in the future.

As for the processes that happen outside of the Earth, Laurenz, Lüdecke and Lüning (2019) investigated the impact of the solar cycle on rainfall in Europe. In the period from 1901 to 2015, they calculated the Pearson correlation coefficient based on sunspot number data and monthly precipitation series in Europe. They did not find the strong correlation that would have allowed them to assert that solar sunspots influence rainfall. However, they assumed that the solar cycle may have an influence on rainfall along with the NAO.

Also, Mohamed and El-Mahdy (2021) using a simple linear regression model, showed a weak negative correlation between sunspot number and rainfall patterns over Eastern Africa. They proposed that processes such as the ENSO have a greater impact on rainfall patterns as an explanation for this result.

What is more, the study in Sytinskii, (1973) claimed that the total seismicity of the Earth, expressed through the total energy of earthquakes and the number of catastrophic earthquakes per year, depends on the phase of the 11-year solar cycle. Also, the time of occurrence of individual strong earthquakes with a Richter magnitude $M \geq 6.5$ depends on the position of active regions on the Sun. Earthquakes occur mainly 2-3 days after the passage of the active region through the central solar meridian.

2.2 Earthquake

2.2.1 Earthquake General Information

In the last decades, as can be seen from Figure 1.1, the most dangerous natural disaster for people have been earthquakes. According to Kanamori and Brodsky (2004) the simple definition of an earthquake is an event shaking the Earth's surface. Worldwide, earthquakes are one of the most severe natural disasters. According to Dong and Shan (2013), about 60% of all deaths that occurred due to natural disasters were caused by earthquakes.

An earthquake occurs as a result of global tectonic plate movement. An earthquake is defined by several basic parameters, including its depth, hypocentre, and magnitude. (Table 2.1):

- Earthquake Depth: The depth indicates where an earthquake can occur between the Earth's surface and 700 kilometres below the surface. This sub-surface region is divided into three zones: shallow (0 – 70 km), intermediate (70 – 300 km), and deep (300 – 700 km).
- The hypocentre is the point of initiation of an earthquake.
- Earthquake Magnitude: The magnitude is the measure of the size of an earthquake source. Depending on the magnitude an earthquake be subdivided into classes.

Table 2.1 Earthquake Magnitude Scale (*Earthquakes magnitude scale and classes*, 2021)

Magnitude	**Earthquake Effects**	**Each Year Estimated Number**
2.5 or less	Not felt, can be recorded by seismograph.	900,000
2.5 to 5.4	Often felt, but only causes minor damage.	30,000
5.5 to 6.0	Slight damage to buildings and other structures.	500
6.1 to 6.9	A lot of damage in very populated areas.	100
7.0 to 7.9	Serious damage.	20
8.0 or greater	Can destroy communities.	One every 5 to 10 years

2.2.2 Earthquake map

Earthquakes happen at conservative and collisional plate margins (Merle, 2011):

- Conservative – the tectonic plates are **sliding past** each other.
- Continental – tectonic plates are moving **towards** each other.

The significant earthquakes that took place during the 23rd solar cycle are shown in Figure 2.1. Together with the plate tectonics map in Figure 2.2, they show that earthquake events mainly take place at the plate tectonic boundary. In Figure 2.1, earthquakes are distributed by magnitude based on the group from Table 2.1.

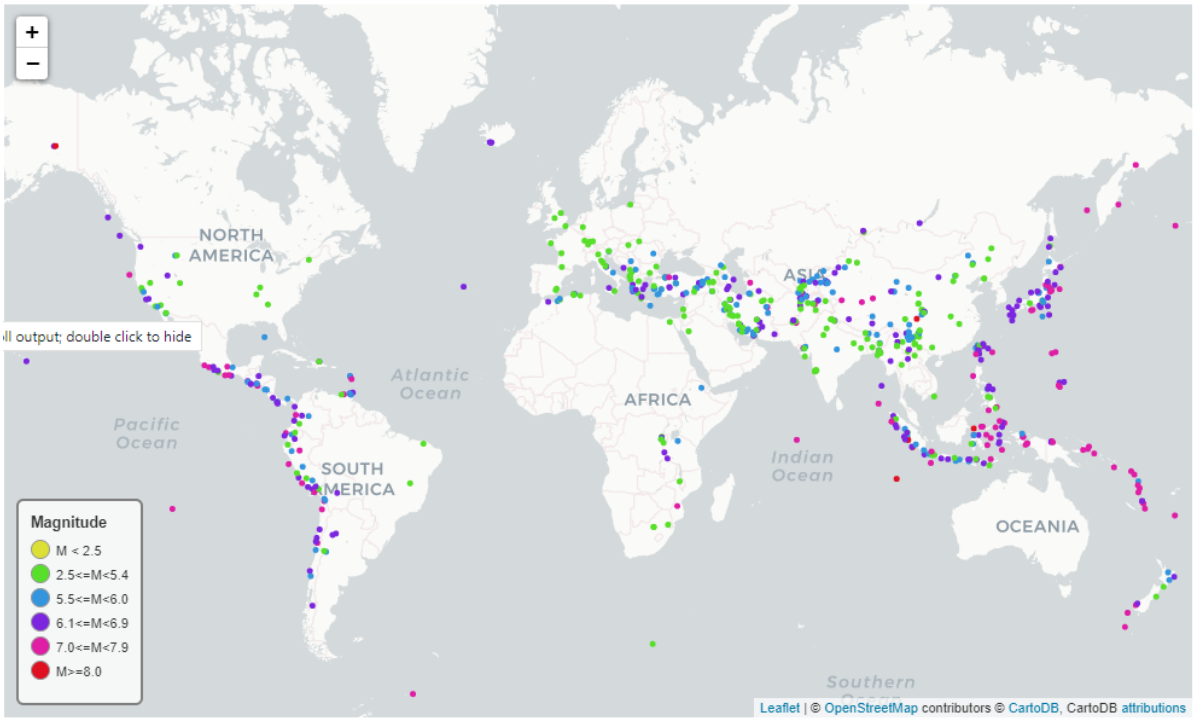


Figure 2.1 Earthquake events map during 23rd solar cycle, distributed by magnitude (python code for the map in Appendix A, Figure A - 1).

The earthquake data for Figure 2.1 were collected from the open-source National Geophysical Data Center / World Data Service (NGDC/WDS) (1972). The dataset consists of information about significant earthquakes that meet one of the criteria, such as "caused deaths, caused moderate damage (approximately $1 million or more). The minimum magnitude in the dataset is M = 1.6, which is proof that even the earthquake from the first group (Table 2.1) can lead to significant damage.

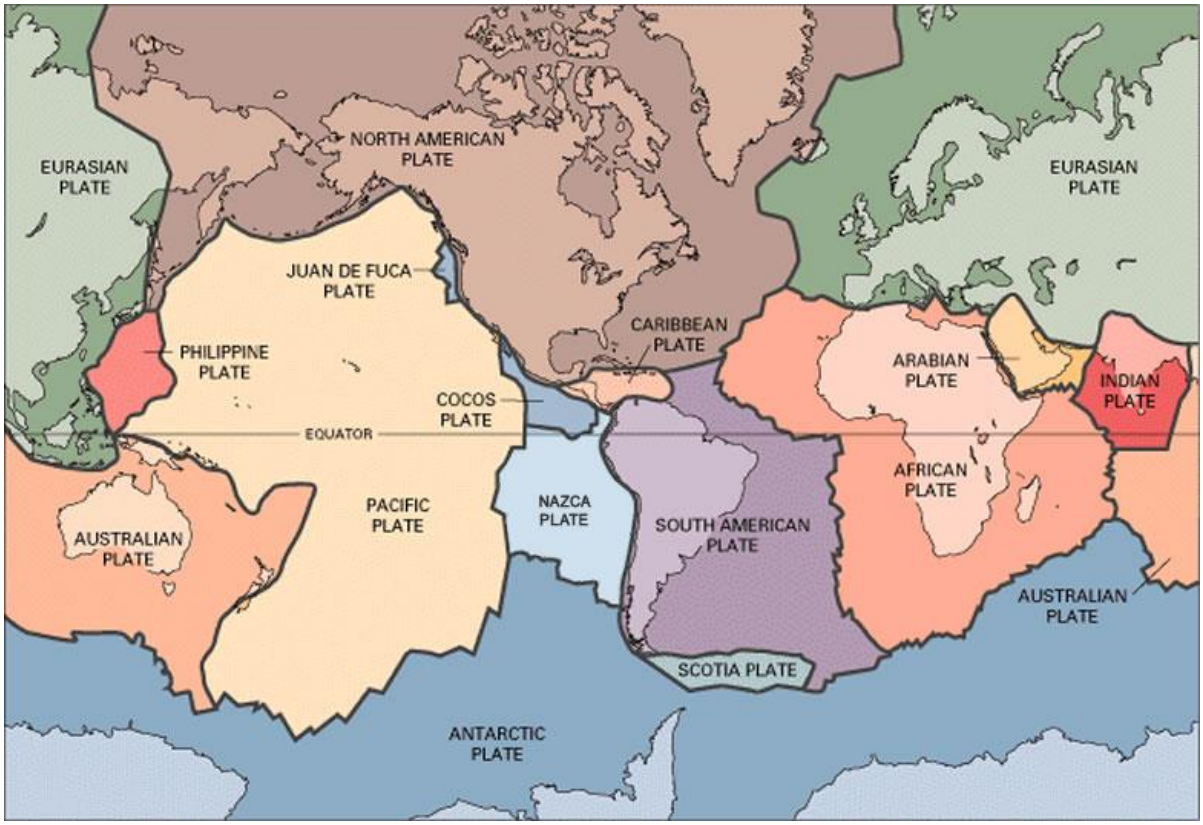


Figure 2.2 Plate Tectonics Map - Plate Boundary Map. Source: Plate Tectonics Map - Plate Boundary Map (2021)

2.2.3 Signs and Common Events that Trigger of Earthquakes

There are few ways to predict an earthquake. Studies based on the examination of different animal behaviours showed the possibility of predicting earthquakes. For example, Fidani (2010) found that one hour before the earthquake, all dogs started to bark and stopped barking right after the main shock of the earthquake. Li *et al.*, 2009, in their study, noted the change in the behaviour of mice before and after an earthquake. Others (Yamauchi *et al.*, 2017) found that about three weeks before an earthquake, cows produced less milk than usual.

Another way to predict an earthquake is to monitor the change in electromagnetic field signals. Amezquita Sanchez *et al.*, 2017 found anomalies in ultra-low frequency signals up to 8.5 hours before earthquakes. Also, Ida *et al.* (2008) showed that anomalies occur in the ultra-low frequency signals before an earthquake of Richter magnitude 6.1. According to Masci and Thomas (2015), other factors such as geomagnetic storms and solar activity are causes of anomalies in ultra-low frequency signals, and their influence on earthquakes should be investigated further. Other studies focused their attention on analysing the change in Earth's water levels. Orihara, Kamogawa and Nagao (2015) found a decrease in groundwater temperature and level three months before the 2011 earthquake in Japan. Also, the study of

Singh *et al.* (2010) concentrated their attention on analysing the chemical composition of ground waters and reporting changes before earthquakes. One of the natural triggers of earthquakes can be volcanic eruptions (McNutt and Roman, 2015), however, on the other hand, earthquakes can also be a trigger of volcanos (Nishimura, 2017). The other examples of natural triggers of earthquakes are rainfall (Hainzl et al., 2006), tidal stress (Métivier *et al.*, 2009), solar weather (Sytinskii, 1973).

What is more, earthquakes might potentially be caused by man's meddling with nature. A change in crustal balance caused by heavy water pressure in dams can produce earthquakes (Chander, 1999). Mining may be a source of concern since it involves the removal of large amounts of rock from various regions, which might result in an earthquake (Redmayne, 1988). Also, nuclear bombs and testing can cause particular sorts of shockwaves to travel over the earth's surface, disrupting tectonic plate alignment (Tian, Yao and Wen, 2018).

Arguably, the methods of using these signs to predict earthquakes are too far from being perfect (Schorlemmer *et al.*, 2018).

2.3 Solar activity

2.3.1 Sunspots and Solar cycles

A sunspot is a dark area that appears on the Sun's surface. The temperature within the dark area is cooler than the surrounding surface. Sunspots have various shapes and range in size, showing different diameters. The lifetime of sunspots depends on their size. The smaller the area, the shorter the lifetime. A sunspot with a diameter of 10 Mm may last for 2 – 3 days, but one with a diameter of 60 Mm can last up to 90 days (Priest, 2014).

Solar activities depend on a solar cycle. The sun is generating a magnetic field, which goes through a cycle. During this cycle, the magnetic field reverses and the north and south poles of the sun switch positions. During the next solar cycle, the poles revert back. One solar cycle has a period of approximately 11 years (Figure 2.3). The length of the solar cycle may vary, and throughout the cycle, the number of sunspots may fluctuate. The solar cycle has a solar minimum and a solar maximum. The solar minimum at the beginning and end of each solar cycle is associated with the minimum number of sunspots, while the solar maximum in the middle of the cycle is linked to the maximum number of sunspots (Priest, 2014). The first solar cycle was documented in the 18th century (Priest, 2014). The current solar cycle, the 25th, began in December 2019 (Potter, 2020).

2.3.2 Sunspot number

The state of the solar cycle is measured by calculating sunspots (Priest, 2014). Figure 2.3 provides a graphic representation of the solar cycles based on the average quantity of sunspot number per year. The graph was created based on data from the open-source resource SILSO World Data Center website (*SILSO | World Data Center for the production, preservation and dissemination of the international sunspot number* 2021). In Figure 2.3 the beginning and end of the solar cycle can be clearly seen, as can the rising and falling faces, minimum, and maximum of the solar cycle. The Python codes for the generation of SSN data and a graph are shown in Appendix A (Figure A - 3 and Figure A - 4).

According to Hathaway (2015), there is more than one method of counting the number of sunspots. The first method that yields the International Sunspot Number is a traditional way that uses the Wolf number. The Wolf number is a quantity that measures the number of sunspots. However, there are additional methods, often better ones. The Boulder Sunspot Number is provided by the US Air Force and the National Oceanic and Atmospheric Administration. The American Sunspot Number is provided by the American Association of Variable Star Observers. The Group Sunspot Number is used by Hoyt and Schatten (1998).

There are various methods for counting sunspots, but none of them can be said to be the optimal among them.

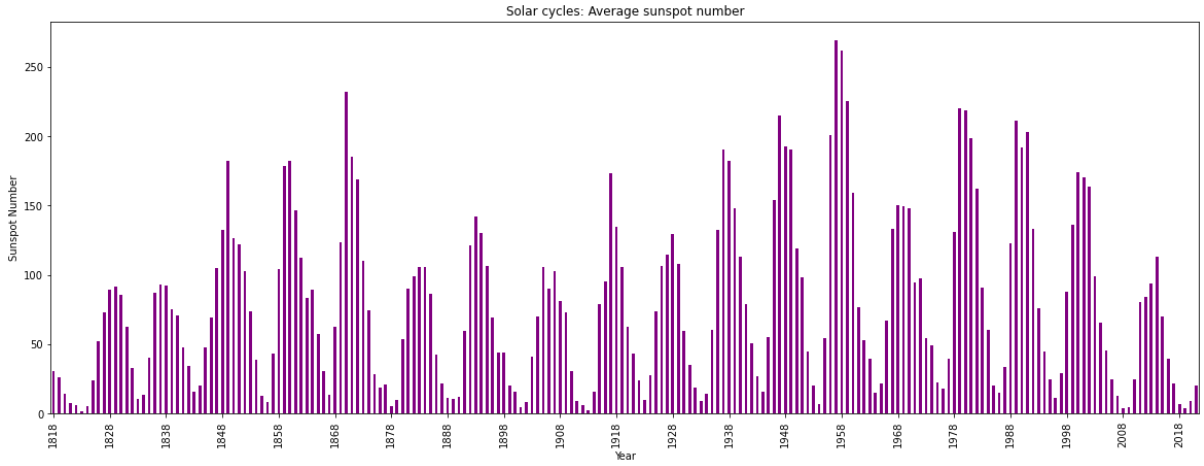


Figure 2.3 Solar cycle and sunspot number, data source: SILSO World Data Center website

Khan *et al.* (2020) made a sunspot number prediction for the 25th solar cycle in their study. As for a dataset, they used sunspot number data from 1818 to 2020. For machine learning technique – Long Short-Term Memory network. The study predicted that the 25th solar cycle would be reached between 2021 and 2025.

Dani and Sulistiani (2019) predicted the maximum sunspot number for the 25th solar cycle. They used four machine learning algorithms: Random Forest, Support Vector Machine, Linear Regression, and Radial Basis Function. As for the programming environment, they used open-source software called Weka. All four algorithms showed different results. Support Vector Machine and Radial Basis Function predicted similar sunspot numbers, while linear regression predicted the greatest number of sunspots. However, as the predicted maximum will be around the years 2023 and 2024, calculations might currently not be as accurate.

2.3.3 Solar Flares

A solar flare is an explosion of energy that is also accompanied by a coronal mass ejection. This explosion of energy occurs because magnetic fields intersect and reorganise near sunspots. Historically, a solar flare was defined using the H-alpha wavelength; later, with the onset of the space age, it was extended to the X-ray wavelength with the Geostationary Operational Environmental Satellites (GOES). Based on the flare's strength, flares are classified as A-class, B-class, C-class, M-class, and X-class. A-class is the smallest, and X-class is the largest solar flare by size. In turn, each flare class has a scale from 1 to 9; however, X-class flares can exceed the top scale of 9 (Priest, 2014).

Asaly, Gottlieb and Reuveni (2021) based on two datasets, Ionospheric Total Electron Content data and solar flare data, used Support Vector Machine for solar flare prediction. They found a high probability of predicting large-size solar flares of the M and X classes. However, the method they chose did not work for the prediction of small-sized solar flares.

2.3.4 Solar Wind

According to the description in NASA/Marshall solar physics (2014), the solar wind is a "not uniform" stream of charged particles that flows from the sun in all possible directions at a speed of about 400 km/s. According to Wood *et al. (2009)*, the measurements of the solar wind are solar wind speed (velocity), proton density, and proton temperature. The source of the solar wind is the Sun's outer atmosphere, which is called the corona. The Sun's gravity cannot hold the charged particles when the temperature of the corona rises. All streams flow away from the sun, and their speeds change depending on different factors, for example, magnetic clouds. The highest speed of around 800 km/s occurred over regions where the corona is dark (corona holes), and the lowest speed of around 300 km/s was observed over large cap-like coronal structures with long, pointed peaks that usually overlie sunspots and active regions (streamers).

The mean distance from the Sun to the Earth is $1.5x10^{11}$ m (Meyer-Vernet, 2012). As a result, the average time for solar wind to reach Earth is 4.3 days (1.5×10^8 km 400 km/s = 375,000 sec). However, the real time between detection of the solar wind and its arrival on Earth may be shorter because of the location of the satellites that detect it. For example, the location of the ACE (Advanced Composition Explorer) satellite between the Earth and the Sun about $1.5*10^6$ km forward of the Earth (*ACE real-time solar wind | NOAA / NWS space weather prediction center*, 2021).

Solar wind source classification is critical for solar and heliospheric physics research. Feldman (2005) conducted research on this topic. When it comes to categorising solar wind plasma, it's usually as simple as dividing it into "fast wind" and "slow wind" based on the wind speed. However, according to Xu and Borovsky (2015), there are four basic forms of solar wind: coronal-hole-origin plasma (CHOP), streamer belt plasma (SBP), sector-reversal-region plasma (SRRP), and ejecta (EJECT). That is why the classification of solar wind using machine learning techniques is becoming more popular. As a result, solar wind classification using machine learning techniques is becoming increasingly popular. In a solar wind classification, Camporeale, Carè and Borovsky (2017) used a supervised learning technique – Gaussian process and attained an accuracy of around 96 percent for all categories.

2.4 Earthquakes and Solar Activity

The two events, solar activity and earthquakes, take place at different locations within the solar system, on the surfaces of the Sun and the Earth, separated by approximately 1.5×10^{11} m (Meyer-Vernet, 2012). However, starting from Wolf (1853), researchers have tried to find out if there are any connections between these two seemingly separate events.

However, there is an opposite opinion. Love and Thomas (2013) claimed that there is no statistically valid explanation proving that solar-terrestrial interaction favours earthquake incidence. For their study they used data from the *SPDF - OMNIWeb Service* (2021) and *Sunspot-numbers - monthly* (2021). They based their judgment using χ^2 and Student's *t* tests. On the other hand, Love and Thomas (2013) acknowledged that they do not have proof that the notion that solar activity has no effect is correct.

There is an assumption, that earthquakes are influenced by several factors. Bijan, Saied and Somayeh (2013) in their study classified earthquakes into two categories (Figure 2.4):

1. Earthquakes that occurred as a result of tectonic or internal earth effects, such as rainfall, volcanic eruptions, or landslides.

2. Earthquakes caused by non-tectonic effects or external effects of the earth, such as the gravitational pull of the sun and moon or solar activity.

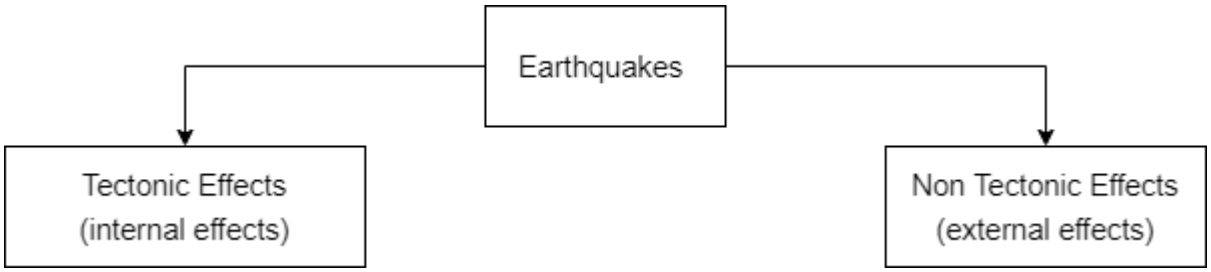


Figure 2.4 Classification of Triggers of Earthquakes, source: Bijan, Saied and Somayeh, 2013

What is more, Bijan, Saied and Somayeh (2013) claim, based on statistical graphs, that the number of earthquakes during the day is less than the number of earthquakes during the night. They explained this fact by saying that the atmospheric pressure during the day and night is different.

The study by Novikov *et al.* (2020) showed the possibility that strong solar flares impact the triggering of earthquakes. Novikov et al. (2020) used laboratory experiments and earthquake observations before and after a significant solar event (a solar flare of the X class). Novikov et

al. (2020) discussed the results of laboratory experiments in the first part of their study: electric current injected into the Earth by an artificial generator, what causes a telluric current, and the impact of ionising radiation from solar flares on earthquake sources. They noticed an increase in the quantity of earthquakes with a Richter magnitude less than 3 after the injection of electric current into the Earth's crust. It has been found that geomagnetic pulsations caused by X-rays from X-class solar flares, as well as geomagnetic storms, can generate geomagnetically induced currents in earthquake sources. What leads to the hypothesis being confirmed is that the electromagnetic mechanism of earthquake initiation under strong variations in space weather is confirmed.

Novikov et al. (2020) observed earthquake events that happened between August and September 2017. In this period of time, namely September 6, the X-class solar flare occurred. Novikov et al. (2020) used two groups of earthquakes: a quantity of global earthquakes with a Richter magnitude M≥4, and a quantity of regional (Greece) earthquakes with a Richter magnitude M≥3. The conclusion is based on a comparison of the number of earthquakes occurring before and after the solar flare X-class. An increase in the number of earthquakes in both groups of earthquakes, global (increased by 68%) and regional (increased by 120%), after the solar flare X-class allowed Novikov *et al.* (2020) assume the dependence of earthquakes on solar flares. Also, Novikov *et al.* (2020) mentioned that the current density in the Earth's crust depends on its electrical conductivity. It was demonstrated that if the electrical conductivity at 10 km depth is greater than that at the Earth's surface, the current density will increase.

Odintsov *et al.* (2006) and Odintsov, Ivanov-Kholodnyi and Georgieva (2007) tried to confirm the hypothesis of Sytinskii, (1973) that earthquakes with a Richter magnitude greater than 6.5 match with high-speed solar winds whose velocity is more than 500 km/s. For their study, they used a 27-year period, a daily number of earthquakes with a Richter magnitude M≥5.5, and solar wind with a velocity of 500 km/s and above. They identified 307 cases of solar wind with this velocity value. Odintsov et al. (2006) and Odintsov, Ivanov-Kholodnyi, and Georgieva (2007) examined the number of earthquakes on the day of solar wind arrival as well as a few days before and after. They found an increase in the quantity of earthquakes on the day of the high-speed solar wind arrival and the day after. Also, Odintsov, Ivanov-Kholodnyi and Georgieva (2007) observed nine full solar cycles to find out if there is a connection between earthquakes and solar activity. For this part they used quantity of earthquakes with Richter magnitude M≥7. They compared the average yearly sunspot number with the average yearly number of earthquakes and found that during the 11-year solar cycle, the number of earthquakes has two maxima. The first maximum is the same as the maximum sunspot

number, and the second maximum occurs during the descending phase of the 11-year solar cycle.

Furthermore, recent data-driven studies have discovered a link between global earthquakes and solar activity. Nishii, Qin and Kikuyama (2020) set out to determine whether solar activity is a source of earthquakes. They used solar activity data from *SPDF's OMNIWeb Service* (2021) and earthquake data from the *Usgs earthquake hazards program* (2021) catalogue for their study. They discovered a link between solar activity and earthquakes using support vector regression, notably for earthquakes with a Richter magnitude of less than 6. Using statistical methods Solar activity and earthquakes are linked, according to Marchitelli et al. (2020). They used two characteristics of solar wind for their research: proton density and velocity for solar activity data, and worldwide earthquakes with a Richter magnitude equal to or greater than 5.6 over a 20-year period. For the earthquake data, they used the Storchak et al. (2013) earthquake catalogue and solar activity data from the Solar and Heliospheric Observatory (SOHO) satellite.

According to previous research, it is still unclear whether solar activity events are the cause of natural disasters. On the other hand, the assertion that solar activity events have an impact on earthquakes cannot be dismissed. Furthermore, studies show that:

1. Earthquakes are influenced by the 11-year solar cycle.
2. Earthquakes are influenced by some solar activity.
3. Earthquakes are influenced by solar activity's electric current.
4. A correlation between solar activity and earthquakes can be established using statistical techniques.
5. Machine learning techniques can be used to determine whether solar activity is a cause of earthquakes.

2.5 Machine Learning

With the rise of computing, there are more and more studies that use machine learning in the prediction of earthquakes based on historical data analysis. There are a lot of ways to use machine learning in the earthquake sphere, from prediction of earthquake events to management of post-earthquake events. Mangalathu *et al.* (2020), as a part of post-earthquake management, classified earthquake-induced building damage. For their study, they had chosen four machine learning algorithms: linear discriminant analysis, k-nearest neighbour, decision trees, and neighbour forest. All four algorithms showed an accuracy prediction rate of around 60%; the highest accuracy of 66% was shown by using the random forest algorithm. Asim *et al.* (2017) based on the earthquake data and the seismic parameters, studied the prediction of earthquakes using four machine learning techniques: pattern recognition neural network, recurrent neural network, random forest, and linear programming boost ensemble classifier. Every algorithm showed different results when compared to each other.

The very first tasks of machine learning are to clarify a problem and explore data. Understanding the data is one of the most important parts of machine learning. A good knowledge of a problem and data will help choose the right machine learning technique. Without this understanding, the choice of the machine learning techniques would be random (Müller and Guido, 2016). The basic stages of the machine learning process are presented graphically in Figure 2.5.

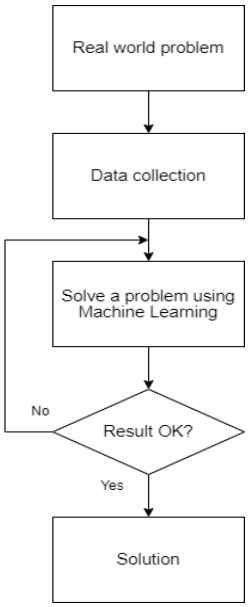


Figure 2.5 Basic stages for machine learning process, adapted from Kuncheva (2004).

There are three main types of what is called the learning process: supervised learning, unsupervised learning, and reinforcement learning (Kaplan, A., 2019), (Figure 2.6):

- Supervised learning is based on the relationship between inputs and their outputs, based on the result and knowledge gained, which allows for a future prediction. In supervised learning, data is pre-categorized or numerical (Kotsiantis, 2007).
- Unsupervised learning is used to know more about data. In unsupervised learning, input data points are not labelled and do not belong to categories. Unsupervised learning can be considered the process of finding patterns in data (Ghahramani, 2004).
- Reinforcement learning algorithm is also called the agent. The agent learns from an environment using feedback and compares actions based on feedback, trying to choose the most appropriate one (Sutton and Barto, 2018).

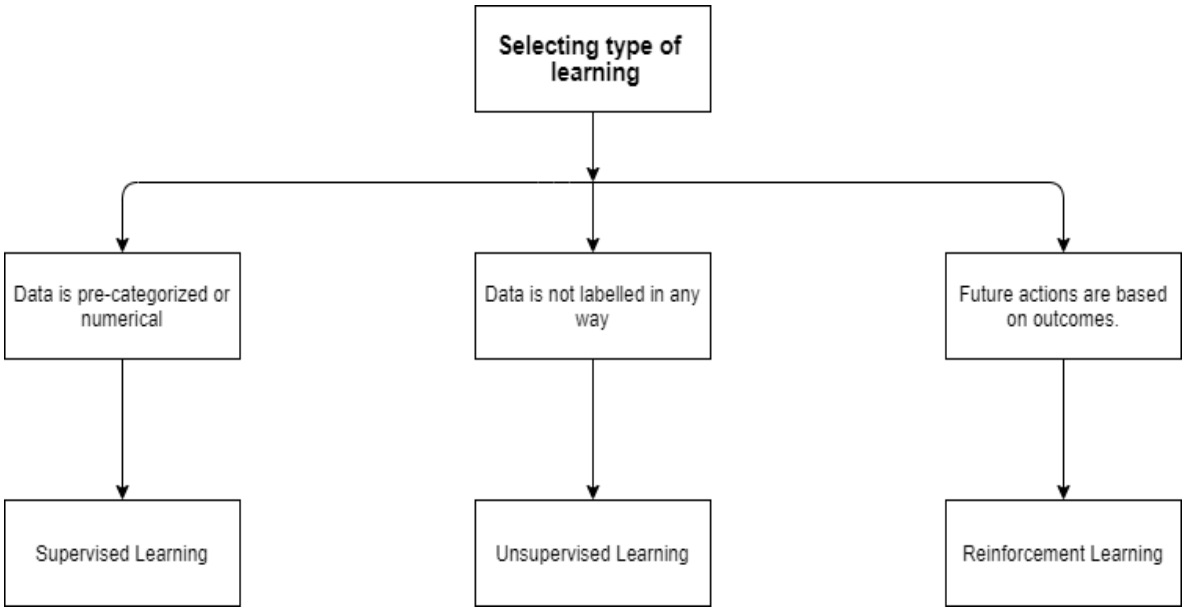


Figure 2.6 Taxonomy of machine learning algorithms.

2.5.1 Supervised Learning (SL)

The solar activity data and earthquake data are labelled. As previously stated, supervised learning uses labelled data. That is why supervised learning is the most appropriate option in these case. Supervised learning methods are based on the relationship between inputs and their outputs. For example, let the input variables be represented by a label "X" and the output variable "Y" and supervised learning algorithms are used to learn the mapping function from the input "X" to the output "Y". (Kotsiantis, 2007; Mohamed, 2017).

Since the inputs and outputs are known during the learning process, high accuracy in a prediction can be achieved. That is why supervised learning is highly used in the spheres of solar activity and natural disasters. Novianty *et al.* (2019) used SL to detect tsunamis, Nishii, Qin and Kikuyama (2020) used SL to find if solar activity affected earthquake events. Murwantara, Yugopuspito and Hermawan (2020) and Mallouhy *et al.* (2019) used SL to predict earthquakes. Rasouli, Hsieh, and Cannon (2012) and Aguilar-Martinez and Hsieh (2009) predicted teleconnections using SL.

There are two major types of SL, Figure 2.7. The first one is classification, and the other is regression (Mohamed, 2017; Müller and Guido, 2016). A classification problem predicts “a category”, while a regression problem predicts “a number”. (Kotsiantis, 2007).

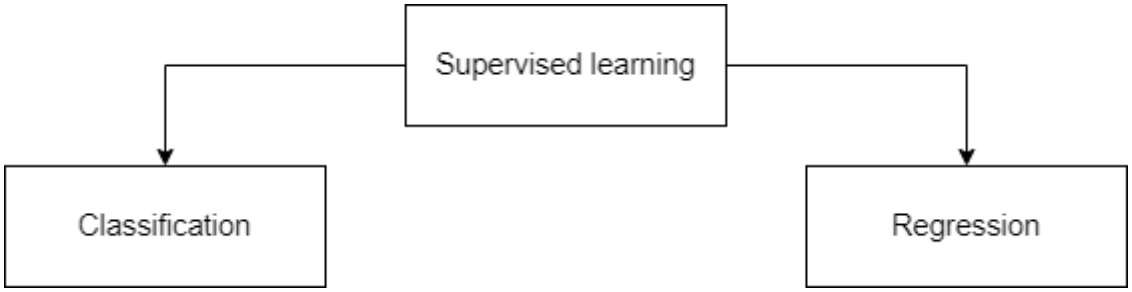


Figure 2.7 Supervised learning.

The main goal of classification supervised machine learning is to predict categorised outputs from previously learned input data. Each output is assigned to a specific category or class (Müller and Guido, 2016). The main goal of regression supervised learning is to estimate the value. The output data attribute of the input data attributes is a numeric value. Finding the value of an object is a common application of supervised learning regression (Alpaydin, 2014).

2.5.2 Evaluation metrics in supervised learning, regression

Evaluation metrics determine how accurate a prediction is. There are a few different types of metrics that are dependent on the task and the algorithms used. For example, classification metrics, regression metrics, and clustering metrics. Classification metrics are used to illustrate the quality of a prediction in supervised learning (classification type), whereas regression metrics are used to show how well a prediction is in supervised learning (regression type). Clustering metrics refer to unsupervised learning.

Regression is a type of predictive modelling that entails forecasting a numerical value. Calculating an error score to summarise a model's prediction ability is one of the regression metrics. The regression metrics demonstrate how closely the predicted values match the actual ones (Draper and Smith, 1998). According to Draper and Smith (1998) the challenges that require estimating a numeric value are known as "regression predictive modeling", like in the current situation. As a result, the regression metrics were examined in greater depth.

Error is used in regression metrics. Error is a metric that measures how close forecasts were to their predicted values on average. Witten and Frank (2017) have compiled a list of useful regression metrics. However, the R-squared error, mean absolute percentage error, mean absolute error, mean squared error, and root mean squared error are arguably the most extensively used metrics.

R-squared (R^2), *equation (1)* – is the fraction of the variance in the dependent variable that can be predicted by the independent variable. The closer R^2 to the "1" the better model fits data.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - \acute{p}_i)^2}{\sum_{i=1}^{n}(p_i - \bar{p})^2} \qquad (1)$$

Where:

n – the number of data points

$\acute{p}_1, \acute{p}_2, \dots, \acute{p}_n$ – predicted values

$p_1, p_2, \dots, p_n$ – actual values

$\bar{p}$ – mean of actual values

Mean absolute percentage error (MAPE), *equation (2)* – mean absolute percentage deviation. MAPE has a percentage value.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{p_i - \acute{p}_i}{p_i}\right| \qquad (2)$$

Where:

n – the number of data points

$\acute{p}_1, \acute{p}_2, \ldots, \acute{p}_n$ – predicted values

$p_1, p_2, \ldots, p_n$ – actual values

Mean absolute error (MAE), *equation (3)* – the average of the difference between the predicted and actual values. MAE has the same units as the original data. MAE shows how close the predicted values were to the actual values.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - \acute{p}_i| \qquad (3)$$

Where:

n – the number of data points

$\acute{p}_1, \acute{p}_2, \ldots, \acute{p}_n$ – predicted values

$p_1, p_2, \ldots, p_n$ – actual values

Mean squared error (MSE), *equation (4)* – the difference between estimated and actual values, expressed as an average squared difference. MSE has the squared units of the original data.

$$MSE = \frac{\sum_{i=1}^{n}(\acute{p}_i - p_i)^2}{n} \qquad (4)$$

Where:

n – the number of data points

$\acute{p}_1, \acute{p}_2, \ldots, \acute{p}_n$ – predicted values

$p_1, p_2, \ldots, p_n$ – actual values

Root mean squared error (RMSE), *equation (5)* – square root of MSE. RMSE has the same units as the original data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\acute{p}_i - p_i)^2}{n}} \qquad (5)$$

Where:

n – the number of data points

$\acute{p}_1, \acute{p}_2, \dots, \acute{p}_n$ – predicted values

$p_1, p_2, \dots, p_n$ – actual values

The same rule applies to MAPE, MAE, MSE, and RMSE: the lower the error, the better the model matches the data.

2.5.3 Types of Supervised Learning Algorithms.

For efficient implementation of a SL algorithm, there is a Scikit-learn library for Python (Pedregosa *et al.*, 2011). Scikit-learn is a free machine learning library, that was developed for Python (*Supervised learning — scikit-learn 0.24.2 documentation*, 2021). Depending on the task, supervised learning algorithms can be used for both classification and regression learning types. The descriptions of the summary of supervised ML algorithms are provided below.

i. **The K-Nearest Neighbors Algorithm (KNN)** is one of the simplest algorithms to implement and is used in natural disaster studies. However, with the increase in data size, it becomes slower. For example, Novianty *et al.* (2019) measured the accuracy of the identification of tsunamis based on earthquake events using the KNN algorithm with an earthquake dataset and a tsunami dataset. They used three dataset variations and different "K" values and discovered that as the "K" value increases, so does the accuracy of the identification tsunami; however, after a certain value of "K," there is no significant change in accuracy.

KNN implementation requires only two parameters: the *"K"* value, which means the number of nearest datapoints to the new data point, and the distance function. The value for *"K"* depends on a dataset. However, the higher the *"K"* value, the less noise influences the classification, and the forecast becomes more accurate, although boundaries between classes are less clear. There is no need to build a model, and new data can be easily added. KNN can be used for both classification (mostly) and regression learning types (Alpaydin, 2014). The measure between two data points calculated using Euclidean distance, *equation* (6), is the square root of the sum of the squared differences between two points, new and existing. Manhattan Distance – distance between real vectors using the sum of their absolute differences – is an alternative variation. For the regression task, the "K" closest data points

are chosen based on their distance from the new point, and the average of these data points is used to make the final forecast for the new point.

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad (6)$$

Where:

$p,\ q$– data represented in n-dimensional vector

n – dimension

ii. **Simple and Multiple Linear Regression (SLR and MLR)** are statistical methods that create relationships between independent and dependent data variables. The goal of SLR is to determine how much an independent variable influences a dependent variable (outcome) (Zou, Tuncali and Silverman, 2003). A model for SLR is a line function, equation *(7)*.

$$p = \sum_{i=1}^{n} a + bq_i \quad (7)$$

Where:

p– dependent data point

q – independent data point

a, b – coefficients, (a – y-intercept, b – slope of a line)

n – dimension

SLR is also one of the methods for determining if the connection between the dependent and independent variables is linear or non-linear. To do so, run the dataset through SLR and evaluate the least square error. If the least square error is closer to one, the dataset is linear; otherwise, the dataset is non-linear. MLR is an extended version of SLR. MLR creates a linear relationship between more than one independent variable and one dependent variable. Ma et *al.* (2019) study investigated how solar activity (sunspots) impacts the El Niño-Southern Oscillation (ENSO) and ENSO-related events. In their study, Ma et *al.* (2019) used MLR. They

used sunspot and ENSO data, which they divided into two sections based on high or low sun activity. They discovered that solar activity influenced ENSO phases using MLR, but they did not discover mechanisms by which solar activity modulated ENSO events.

iii. **Support Vector Machine Algorithm (SVM)** is one of the most popular algorithms used to solve data analytics problems. The main goal of SVM is to find a function that separates classes by line if a dataset has two features and puts new datapoints into the appropriate classes (Smola and Schölkopf, 2004). SVM is to find the most appropriate line, which has the maximum distance between datapoints. This line is called a hyperplane. The dimension of a hyperplane depends on the dimension(s) of a dataset. If a dataset has n features, a hyperplane will have (n-1) dimensions. Also, there are data points that are the closest to the hyperplane; these are called support vectors (Figure 2.8). SVM can be used for classification (mostly) and regression learning types. For regression tasks, SVM is called support vector regression (SVR).

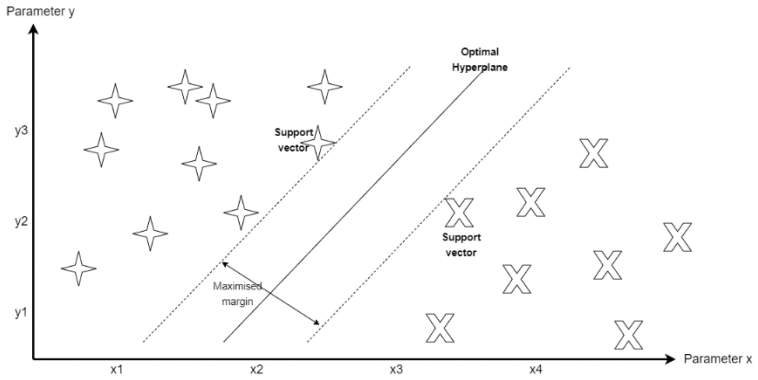


Figure 2.8 SVM, two dimensions.

SVM is based on a collection of mathematical functions known as the kernel. The kernel's job is to take data and turn it into the needed format. Different types of kernel functions are used by different tasks solved using SVM. A kernel is described by its form, *equation (8)*.

$$k(p,q) = (\Phi(p) * \Phi(\mathrm{q})) \qquad (8)$$

Where:

k– kernel

p, q – vectors

$\Phi(p), \Phi(\mathrm{q})$ – feature space

The kernel approach is the most useful feature of SVM, as it aids in solving the linearity and non-linearity of the equation in a very straightforward manner. There are different types of kernels, such as linear, *equation (9)*, polynomial, *equation (10)*, radial basis function (RBF), *equation (11)*, and sigmoid, *equation (12)* (Ghaedi *et al.*, 2016; Benkedjouh *et al.*, 2015; Loutas, Roulias and Georgoulas, 2013; Jacobs, 2012).

$$k(p,q) = p * \mathrm{q} \tag{9}$$

$$k(p,q) = ((p * q) + c)^{a} \tag{10}$$

Where:

c, a – kernel parameters, $c \geq 0$, $a \in \mathrm{N}$

$$k(p,q) = \exp\left(-\|p - q\|^2/\sigma^2\right) \tag{11}$$

Where:

σ – kernel parameters, $\sigma > 0$

$$k(p,q) = tangh(\lambda(p * q) + \psi) \tag{12}$$

Where:

λ, ψ – kernel parameters, $\lambda > 0$, $\psi < 0$

Nishii, Qin and Kikuyama (2020), using SVR, found that solar activity affected some earthquake events. An earthquake dataset was split into five groups, depending on earthquake magnitudes. As for the solar activity, they used nine physical measurements. They also used two vectors for earthquakes and solar activities, error terms, functions were given by a weighted sum of Gaussian kernels. They found that solar activity affects earthquakes with a magnitude(M) of $3 \leq M < 5$ most strongly.

iv. **Logistic Regression Algorithm (LR)** is based on probability. For the logic function, LR uses a logistic function, that gives probabilistic values which are between 0 and 1 (*Logistic regression: from introductory to advanced concepts and applications - sage research methods*, 2010), *equation (13)*.

$$prob = f(a + bq) \quad (13)$$

Where:

f– logistic function

a, b – petametres, which need to be fitted

q – data point

Korsós *et al.* (2021) used LR to predict solar flares based on sunspot and solar flare data. As for the programming environment, they used Python and the Scikit-Learn library. They used two models for training, the first based on a simple characteristic by using only solar flare intensity, and the second more complicated. They found out that LR can predict 70 – 75 % of the flares accurately.

v. **Naïve Bayes Algorithm (NB)** is based on Bayes theorem, *equation (14)*. NB assumes that all features (variables) are independent of each other. Then, it predicts a result based on the probability of an object. NB is used for the classification learning type (Verdhan and Kling, 2020).

$$prob(p|q) = \frac{prob(p|q) * prob(p)}{prob(q)} \quad (14)$$

Where:

$prob(A), prob(B)$– probabilities of the events p and q

$prob(A|B)$ – probability of the event p given the event q

$prob(B|A)$ – probability of the event p given the event q

Murwantara, Yugopuspito and Hermawan (2020) compared three algorithms to predict earthquakes in Indonesia in their study. The algorithms were multinominal LR, SVM, and NB. They made predictions based on available earthquake data. As for the programming environment, they used R with its machine learning library and methods. They discovered that SVM produced the highest accuracy in earthquake prediction, while NB produced the least reliable results.

The following algorithm is a tree-based algorithm. Tree-based algorithms can be used for both classification and regression learning types. Tree-based algorithms consist of nested "If-Else" conditions, Figure 2.9. Tree-based algorithms start with the full population and split the data based on some condition. The splitting will continue until the stopping criteria is met (Verdhan and Kling, 2020).

vi. **Random Forest Algorithm (RF)** is a tree-structured algorithm, based on ensemble learning conception. RF is a classifier that contains a number of tree-structured classifies – decision trees, which consist of independent vectors. A raw dataset is separated into randomly selected sub-features, and then particular subtrees are generated (Breiman, 2001; Alpaydin, 2014).

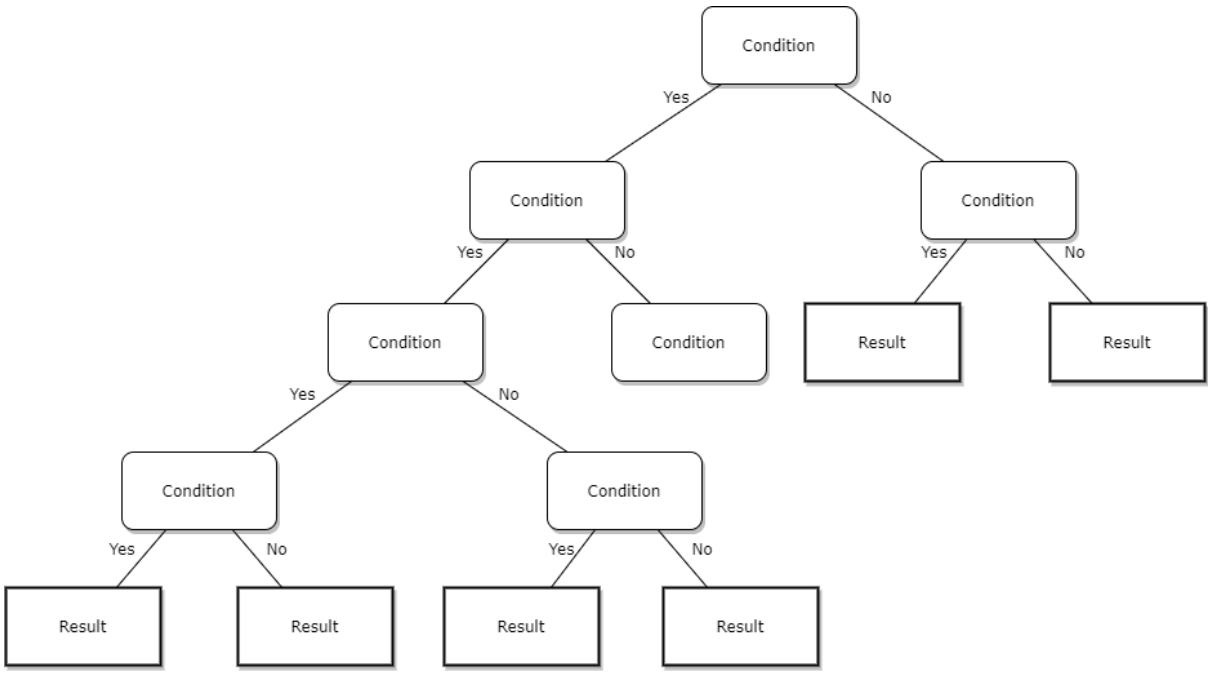


Figure 2.9 Tree-based algorithm, split by some conditions

Each tree has one vote. If the problem is a classification problem, the class with the most votes is the final result. For a regression problem, the average of all subtrees results is the final outcome. The greater the number of trees, the higher the accuracy of the prediction. RF does not have an overfiring problem; it uses a random subspace method (Breiman, 2001;Hothorn, Hornik and Zeileis, 2006). Mallouhy *et al.* (2019) compared in their study eight different algorithms for predicting earthquake events, based on earthquake data. As for the programming environment, they used Matlab. They found that the highest prediction percentage was RF. KNN was very close to RF. NB and LR yielded the lowest prediction percentage.

2.5.4 Dimension Reduction

When analysing data of moderate or high dimension, it is frequently beneficial to look for ways to restructure the data and lower its dimension while retaining the most relevant information or preserving some trait of interest. Dimension reduction is the process of reducing the number of traits, variables, and characteristics (Alpaydin, 2014). Reddy *et al.* (2020) in their study used the supervised algorithm Linear Discriminant Analysis (LDA) and the unsupervised algorithm PCA to reduce the size of a dataset and their impact on the final outcome. To train the reduced dataset, they used four machine learning techniques, including Random Forest and SVM. They discovered that PCA outperformed LDA in terms of final results. They also noticed that the dimensionally reduced dataset showed better results than the original one. However, they also indicated that when the data size is too small, dimensionality reduction methodologies have a negative impact on the performance of machine learning algorithms. One of the most popular unsupervised dimension reduction algorithms is **Principal Component Analysis** (PCA), *equation (15)*. An orthogonal transformation is used in PCA, which is a statistical process. A set of correlated variables is converted to a group of uncorrelated variables using PCA. For exploratory data analysis, PCA is utilised (Reddy *et al.*, 2020).

$$x = W\chi \qquad (15)$$

Where:

x– the observations

W – is the mixing matrix

χ – the source or the independent components

There are four stages in PCA. The standardisation of the raw data is the first stage. The second step is to calculate the raw data's co-variance matrix. The third step is to calculate the eigenvector and eigenvalue of the covariance matrix. The final stage is to project raw data into a new dimensional subspace. Dimensional reduction is unsupervised learning. For efficient implementation of the unsupervised learning algorithm, there is a Scikit-learn library for Python (*Unsupervised learning — scikit-learn 0.24.2 documentation*, 2021).

2.5.5 Neural Networks

Neural networks (NN) can solve both supervised and unsupervised problems. Also, NN are a great method for developing nonparametric and nonlinear classification/regression (Verdhan and Kling, 2020).

There are a lot of different types of NN, such as Recurrent Neural Network, Convolutional Neural Networks, Feed Forward Neural Networks, and Generative Adversarial Networks.

One of the well-known neural networks is the Recurrent Neural Network (RNN). RNN is a neural network that has connections between passages related to sequences and lists and is dependent on previous states. The standard RNN, on the other hand, has a weakness: the gradient vanishes as information is lost over time. The Long Short Term Memory network (LSTM) was created to avoid the long-term dependency problem. The structure of LSTM is similar to that of standard RNN, but the repeating module is different (Hochreiter and Schmidhuber, 1997). According to Hochreiter and Schmidhuber (1997), there are few steps of LSTM. The first step in LSTM is deciding what information from the cell state will be removed, *equation (16)*.

$$f_t = \sigma\left(W_t x_t + U_f h_{t-1} + b_f\right) \quad (16)$$

The second step in LSTM is deciding what "new" information will be stored in the cell state, *equations (17)(18)*.

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (17)$$

$$\tilde{c}_t = \sigma(W_c x_t + U_c h_{t-1} + b_c) \quad (18)$$

And finally, output, *equations (19)(20)(21)*:

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (19)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (20)$$

$$h_t = o_t \sigma_h(c_t) \quad (21)$$

Where:

x_t – input vectors

h_t – hidden vectors

f_t – forget gate's activation vector, between 0 and 1

σ – sigmoid function

W, U – weights

b_t – bias vector

i_t – update gate's activation vector, between 0 and 1

$\tilde{c}_t$ – cell input activation vector, between -1 and 1

c_t – cell state activation vector

o_t – output gate's activation vector, between 0 and 1

The above steps, which included fitting the model and obtaining prediction values, can be completed using Python libraries. *Keras: the Python deep learning API* (2021) is a popular library that comes highly recommended (Verdhan and Kling, 2020). One of the most important advantages of NN compared to traditional ML algorithms is that NN can work well with the increasing size of data. The bigger the training data size, the better the accuracy will get in the final result.

Zhang *et al.* (2017) applied LSTM network to forecast sea surface temperature, based on the sea surface temperature dataset. Aguilar-Martinez and Hsieh (2009) used a Bayesian neural network, support vector regression, and linear regression to forecast sea surface temperature. Yuan *et al.* (2019) based on historical North Atlantic Oscillation index data, created an LSTM network to predict the North Atlantic Oscillation index. Rasouli, Hsieh and Cannon (2012) used a Bayesian neural network, support vector regression, Gaussian processing, and multiple linear regression to try to find out which algorithms better predicted the behaviour of different teleconnection patterns.

2.5.6 Data splitting

Building computational models with good prediction and generalisation skills is one of the most important needs in machine learning (Alpaydin, 2014). To forecast the output, a model should first be trained, and then the model should be evaluated. A dataset should be divided into training and testing sets for this purpose. This causes two issues: with a smaller training data set, the data parameter estimations are more variable, and with a smaller test data set, the performance statistics are more variable. Therefore, the data should be separated such that none of the variances are very large (Kononenko and Kukar, 2007).

According to the previous studies, the most popular ratios for training/testing sets are 70/30 (Dao *et al.*, 2020) and 80/20 (Pham *et al.*, 2020; Das *et al.*, 2011). In their study, Nguyen *et*

al. (2021) claimed that 70/30 is the most appropriate ratio; however, they used a relatively small dataset of 538 samples. However, Rácz, Bajusz and Héberger (2021) suggested that an 80/20 ratio is likely to be superior, especially for large datasets.

2.5.7 Types of normalising

Data normalising is the process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values (Muhamedyev, 2015). Normalising helps reduce data redundancy. Normalising helps to remove anomalies and minimise null values, which are found in large numbers in the data used in the current study. Data redundancy can be reduced by normalising. Normalization aids in the removal of anomalies and the reduction of null values, both of which are common in the data. Furthermore, Raju *et al.* (2020) demonstrated that when data was normalised, the findings were more accurate when compared to the original data.

To normalise data, there are a number of different normalisation and standardisation methods that may be used (Raju *et al.*, 2020). Different techniques are used by different methods; some change the range of values while others change the distribution. In the current study, box plots were used to compare the results of these methods in order to find the highest normalising result.

MinMaxScaler – For each component, the base suggestion is set to 0, the most extreme value is set to 1, and all other values are set to a decimal between 0 and 1, *equation (22)*.

$$p_{scaled} = \frac{(p - p_{min})}{(p_{max} - p_{min})} \qquad (22)$$

Where:

p– data point

p_{min} – minimum value in a dataset

p_{max} – maximum value in a dataset

MaxAbsScaler – similar to MinMaxScaler, the range between 0 and 1, *equation (23)*.

$$p_{scaled} = \frac{p}{max(abs(p))} \qquad (23)$$

Where:

p– data point

StandardScaler – is usually used inside each component to scale it to the point where the distribution is currently centred around 0 with a standard deviation of 1, *equation (24)*.

$$p_{scaled} = \frac{(p - \mu)}{\sigma} \quad (24)$$

Where:

p– data point

μ – mean of a dataset

σ – standard deviation of a dataset

RobustScaler – eliminates the centre and scales the data according to the Interquartile Range (IQR). The interval between the first quartile (25th quantile) and the third quartile is known as the IQR (75th quantile), *equation (25)*.

$$p_{scaled} = \frac{(p - median)}{IQR_{1,3}} \quad (25)$$

Where:

p– data point

$median$ – median of a dataset

$IQR_{1,3}$ – the range between the first and the third quartiles (25th and 75th quantiles)

QuantileTransformer – is changed to follow a uniform or ordinary dispersion using this approach. As a result, in general, this alteration will spread out the most continuous attributes for a specific example. It also reduces the impact of (minor) deviations, making this a good pre-planning strategy. The update is applied independently to each case. QuantileTransform produces non-linear standardisation modifications by contracting the distance between minimal exceptions and inliers. The range between 0 and 1.

3 Chapter Three Methodology

Testing the null hypothesis and the research method used in this study are covered in this chapter. To carry out this study, a variation of design science was used to answer the research questions and prove or disprove the hypothesis as follows: After that, the method goes over what data will be used in the experiment, as well as what data pre - processing methods will be used to get the data ready for the experiment. The chapter discusses the relationship between earthquakes and solar activity. The machine learning algorithms that will be used in the experiment were also discussed in the chapter.

3.1 Testing Null Hypothesis

The null hypothesis is “Solar activity events do not have any relationship with earthquake events, and these two events are completely independent of each other”. For testing the null hypothesis, two variables have been chosen:

- The quantity of all earthquakes during the period, as for the earthquake events
- The sunspot numbers, as for the solar activity events.

These two variables were chosen because the all-earthquake variable describes all earthquake events and the sunspot number is the foundation for measuring a solar cycle; additionally, solar flares and solar winds are dependent on solar cycles, which are based on the sunspot number, as was discussed in Chapter 2.3. Based on a solar cycle, as a solar cycle affects each activity on the Sun’s surface, two periods were chosen for the events:

- The first period is during one solar cycle (23rd solar cycle from August 1996 until November 2008)
- The second period is two solar cycles (23rd and 24th solar cycles, cycle from 1996 till 2020).

Based on the graph in Figure 3.1 (an example of the code in Appendix B, Figure B - 2), it can be assumed that the number of earthquakes increases during the falling phase of the solar cycle, when the number of sunspots decreases from solar maximum to solar minimum. According to the Odintsov *et al.* (2006) study, based on solar activity, earthquakes can change dramatically, which leads to outliers. However, outliers can skew statistical analyses and violate their assumptions. That is why, two datasets were chosen to test the null hypothesis: the original data (Table 3.1) and the compact data, which contains no outliers (Appendix B, Figure B - 5 Compact dataset).

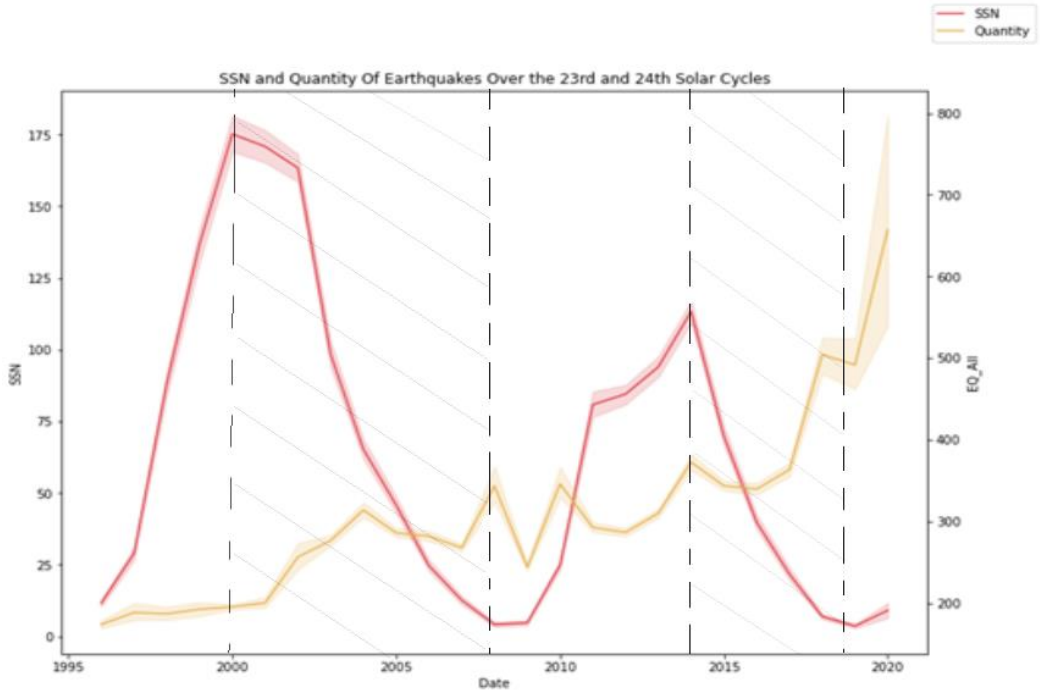


Figure 3.1 SSN and Quantity of Earthquakes Over the " 23rd and 24th Solar Cycles

Table 3.1 Original dataset

Date	Earthquake	SNN
1996-01-03	194	22
1996-01-04	226	35
1996-01-05	191	56
…	…	…
2020-01-08	759	4
2020-01-09	511	15
2020-01-10	445	4

3.1.1 Choosing the type of the test

To figure out which type of test to use for testing the null hypothesis, first it was checked to see if the sunspot number and earthquakes had a normal distribution. The next step was to determine whether the relationships between the sunspot numbers and earthquakes were linear or not. For testing the normality, the graph methods were used: box plots, distribution plots, and probability plots, as seen in Figure 3.2 through Figure 3.7 due to the size of the dataset with N=8718, as described in Chambers (1983) as supposed, such as the Shapiro-Wilk test, do not give the right result (Mohd Razali and Yap, 2011).

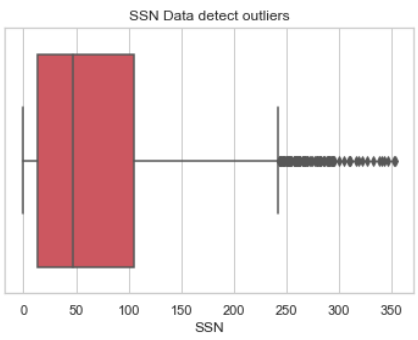


Figure 3.2 SSN: boxplot

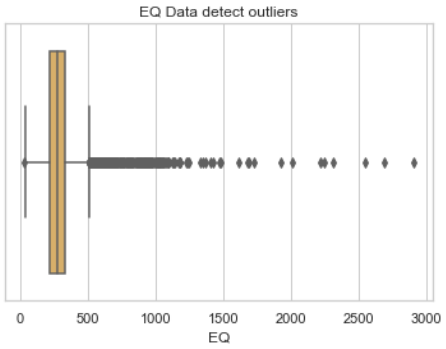


Figure 3.3 EQ: boxplot

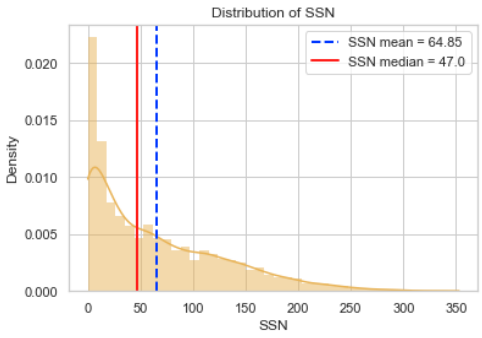


Figure 3.4 Distribution of SSN

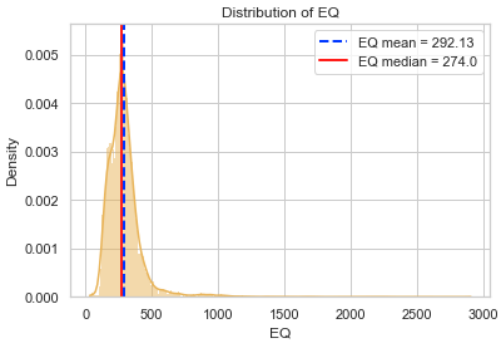


Figure 3.5 Distribution of EQ

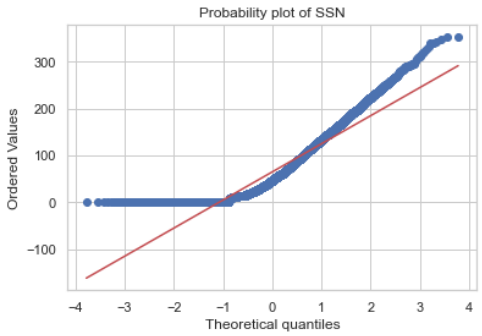


Figure 3.6 Probability plot of SSN

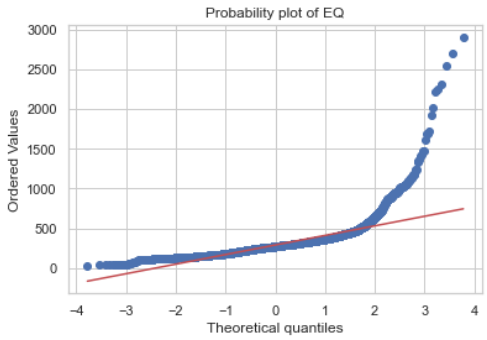


Figure 3.7 Probability plot of EQ

As can be seen from Figure 3.2 and Figure 3.4 the SSN data have outliers, and the distribution of the SSN data, due to the mean and median being left skewed, along with the probability plot in Figure 3.6, have shown that the SSN data is not normally distributed. The earthquake data (Figure 3.3, Figure 3.5, and Figure 3.7), while containing outliers, are nearly normally distributed. However, the presence of outliers raises concerns about the data's normality.

For testing if the relationships were linear or nonlinear, a machine learning simple linear regression algorithm with subsequent checking for least square error was applied. It was discovered that the least square error was equal to "17.246" (an example of the code in Appendix B, Figure B - 2), indicating that the SSN and earthquake have a nonlinear relationship. For further choosing the type of test for the null hypothesis, it was needed to determine whether the relationship is monotonic. The graph in Figure 3.8 shows that the variables tend to move in the same direction at a constant rate, which indicates a monotonic relationship.

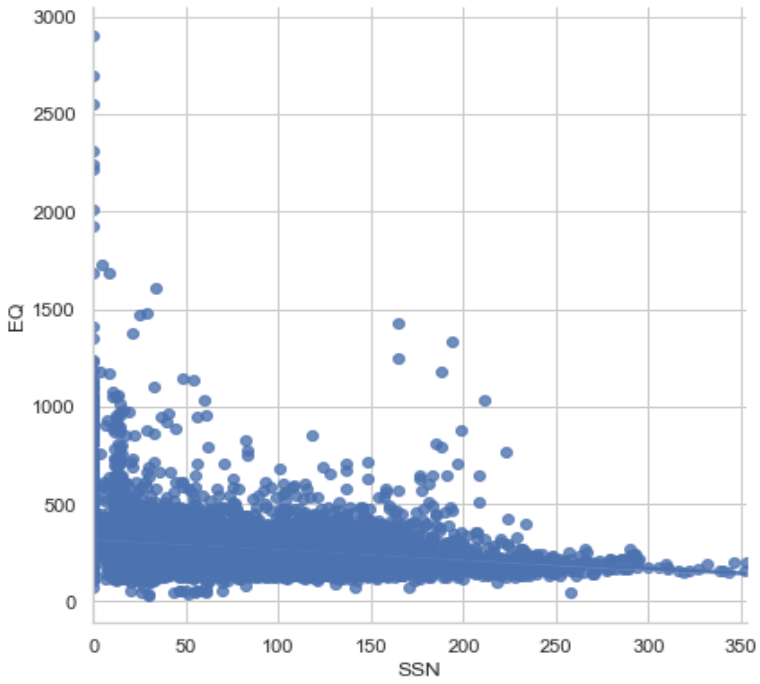


Figure 3.8 Sunspot Number and Earthquakes, original data, relationship

3.1.2 Spearman's rho correlation coefficient

To summarise the testing, it was found that the data are not normally distributed, at least the SSN is not normally distributed, and have nonlinear, monotonic relationships. Based on the de Siqueira Santos *et al.* (2014) study, for testing the null hypothesis, a Spearman's rho correlation coefficient was chosen. To reduce the possibility of a Type I error, α -level equals 0.01 was set (Frick, 1996). For the calculation of Spearman's coefficient, the statistical Python library was used (*Statistical functions (Scipy. Stats) — SciPy v1.7.1 Manual*, 2021). The null hypothesis was tested using both original data and compact data (Appendix B, Figure B - 5 Compact dataset). The results of the Spearman's rho correlation coefficient test are as follows:

- Original data: p-value equal 7.759e-124 is less than 0.01.
- Compact data: p-value equal 1.008e-74 is less than 0.01

Also, the null hypothesis was tested during the 23rd solar cycle. The p-values in both the original and compact datasets were less than 0.01. The dataset examples, graphs, p-value calculation, and codes are in Appendix B.

Based on the Spearman's coefficients, that were calculated during the testing, it was assumed that the null hypothesis could be rejected. Also, Spearman's coefficients showed that there is a possibility that solar activity influences earthquakes. However, the relationship is weak. Perhaps this is because solar activity affects earthquake events of different magnitudes in different ways (Odintsov, Ivanov-Kholodnyi and Georgieva, 2007; Odintsov *et al.*, 2006). Also, different variables of solar activity have different effects on earthquakes (Novikov *et al.*, 2020).

3.2 Research Method

3.2.1 Design of the study

The design for this study has been based on and changed from typical design science models to focus on a researcher's standpoint (Dresch, Lacerda and Antunes, 2015). The research design has seven stages (Figure 3.9). The first step is formalising the research questions. The second step is formalisation of the aspect of the problem, understanding of the outer environment, understanding why the study is important, and systematic literature review. The third step is locating acceptable data resources that can be relied upon in future implementation and satisfy all of the requirements identified in the previous steps. The fourth step is to carry out the experiment so that it may be analysed and debated in the context of the research and answer the research questions. The fifth step is examining the data gathering findings, analysing them, and presenting them in a way that will ideally answer the previously stated research questions. The sixth and final steps are conclusion and communication. These steps include debating the findings and responding to the conclusions drawn from the data collection and analysis. This process frequently leads to potential future studies that may be continuously improved based on the research findings.

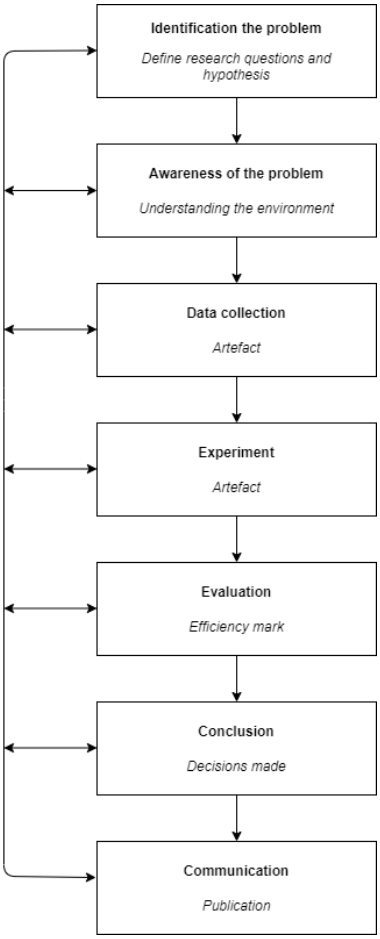


Figure 3.9 Research design

3.2.2 Experiment Method

For the experiment, it was decided to choose both traditional machine learning (KNN, SVR, and RFR) and deep learning (LSTM). Algorithms with a variety of backgrounds were chosen for the study. The selection algorithms were based on the notion that there is a non-linear relationship between the earthquakes and solar activity (that relationship was defined in Chapter 3.7). The experiment is based on the findings of previous seismological studies. Novikov *et al.* (2020) and Odintsov, Ivanov-Kholodnyi and Georgieva (2007) showed the relationship between strong earthquakes (Richter magnitude greater than 5.5) and solar activity. However, in their studies, they did not use earthquakes with a Richter magnitude lower than 5. They compare the number of earthquakes before and after solar activity events in their studies. They discovered that after solar activity events, the number of earthquakes increased. As for solar activity, Novikov *et al.* (2020) used solar flares, and Odintsov, Ivanov-Kholodnyi and Georgieva (2007) used solar wind velocity.

On the other hand, Nishii, Qin and Kikuyama (2020) using Support Vector Regression demonstrated that solar activity affects earthquakes with a Richter magnitude less than 5, but that earthquakes with a Richter magnitude of 3 and 4 have the strongest influence from solar activity. They used a number of earthquakes with a Richter magnitude of at least three and six variables of various types (ratio types and integers) for their study. Also, they used nine measurements of solar activity, including sunspot number, solar wind velocity, proton temperature, and others that were ratio types. They used correlation for the evaluation, which showed that the correlation for earthquakes with Richter magnitudes of 3 and 4 was 0.4777 and 0.5298, respectively. Also, Nishii, Qin and Kikuyama (2020) mentioned that not all nine measurements of solar activity affected earthquakes.

Also, Novikov *et al.* (2020) notice the increasing number of earthquakes with a Richter magnitude less than 3 after the influence of electric current on the Earth's crust, which is similar to solar exposure. The Earth's crust has different electrical conductivity in different regions. According to Novikov *et al.* (2020) and Novikov *et al.* (2017), the higher the electrical conductivity, the higher the current density in the lower Earth crust levels, resulting in an earthquake. It can be assumed that in different regions of the Earth, solar activity may cause earthquakes depending on their depth.

The sunspot number, solar wind (speed (velocity), proton density, and proton temperature), and solar flares (A, B, C, M, X classes) were chosen for this study based on previous research. As for the earthquake data, earthquakes were divided into two parts. The first part is the number of earthquakes with a Richter magnitude less than 5.5, and the second part is earthquakes with a Richter magnitude of 5.5 and larger. Due to the fact that the study is studying global earthquakes, it was decided to use two earthquake options. The first option is to divide earthquakes by their depth, and the second option is not to divide them by their depth. So, all the models are separated by the depth of EQ categories, which, in turn, are separated by each magnitude category. Therefore, the independent variables are solar activity, and the dependent variables are the number of earthquakes. As was mentioned in Chapters 2.3.4 and 2.4, there is a time gap between solar activity events and earthquake events. That is why a few predictions based on time delay were used, ranging from two to seven days, based on Sytinskii's (1973) study and considering the time it takes a solar wind to get from the sun to the Earth.

3.3 Data collection

The data for the study consist of two parts: the earthquake data and the solar activity data. Solar activity depends on a solar cycle. That is why the data for the period between two completed solar cycles, namely the 23rd and 24th solar cycles, were collected from the year 1996 until the year 2020.

3.3.1 Earthquake Data Collection

Daily earthquake data were collected from the United States Geological Survey (USGS) website (Earthquakes, 2021). USGS was created on March 3, 1879. USGS is a US government-run scientific organisation. More than 8,000 employees work at the USGS. USGS notifies authorities, emergency responders, the media, and the general public about major earthquakes in the United States and across the world. The website contains scientific information about natural disasters that endanger people's lives and livelihoods, as well as about water, energy, minerals, and other natural resources. It also keeps long-term seismic data archives for scientific and technical study. The USGS keeps track of seismic activity all around the world.

The Earthquakes (2021) data consist of all earthquake events, regardless of the consequences caused by them. The data consist of 22 variables, but the main ones are: time of earthquakes, location of earthquakes (longitude and latitude), magnitude of earthquakes, and depth of earthquakes. Table 3.2 gives an example of the characteristics (time, location, magnitude, and depth) of earthquakes.

Table 3.2 Earthquake data from the source

time	latitude	longitude	depth	magnitude
1996-01-01 00:08:39	10.1250	-70.0910	10.00	4.2
1996-01-01 01:10:35	60.0308	-153.1522	129.70	2.5
…	…	…	…	…
2020-12-31 23:32:33	55.4040	-159.3768	7.10	3.3
2020-12-31 23:51:23	62.3613	-151.1366	72.40	2.9

3.3.2 Solar Activity Data Collection

Based on Moldwin, M. (2008), there are a number of excellent websites that are related to space weather and the Sun-Earth relationship, such as the National Centers for Environmental Information (NOAA), National Aeronautics and Space Administration (NASA) and websites

that are connected to NASA, Space Weather: The International Journal of Research and Applications and the High Altitude Observatory (HAO).

There are different solar activity events, which were collected from several open-source resources. For this study, the following solar activity data were selected: sunspot number, solar wind (solar wind speed, proton density, and proton temperature), and solar flares (A class, B class, C class, M class, and X class). Table 3.3 shows an example of a physical measurement of solar activity based on open-source data resources.

Table 3.3 Solar activity, physical measurements

Solar Activity	**Units**
Sunspot number	-
Solar wind speed	km/s
Proton density	N/cm^3
Proton temperature	Degrees, K
Solar flare	-

Daily total sunspot number data were collected from the open-source resource SILSO World Data Center website (*SILSO | World Data Center for the production, preservation and dissemination of the international sunspot number*, 2021). The data have two primary variables: date and daily total sunspot number; the whole set has six variables. Table 3.4 displays the daily total number of sunspots.

Table 3.4 Daily total sunspot number

Date	**Daily total sunspot number**
1996-01-01	0
1996-01-02	14
…	…
2020-12-30	32
2020-12-31	34

For the solar wind data, daily averages of the characteristics of the solar wind chosen for the study were taken, including solar wind speed, proton density, and proton temperature. The data were collected from the open-source NASA OMNIWeb website (*SPDF - OMNIWeb Service*, 2021). The daily averages consist of information about solar activity events and their coordinates in the Heliographic Intertial Coordinate System and the Real Time Network

Coordinate System. The coordinates are not needed for the current study. Table 3.5 shows the solar wind characteristics.

Table 3.5 Summarises daily averages of the solar wind measurements

Date	Solar Wind Speed	Proton Density	Proton Temperature
1996-01-01	403	7.9	72020
1996-01-02	398	8.0	77660
…	…	…	…
2020-12-30	483	3.6	122507
2020-12-31	406	3.3	44521

Solar flare data were collected from the open-source NOAA National Center for Environmental Information website. Historical information about the solar flares is limited to the year 2016. Therefore, the data were collected from the Solar Flare Data | NCEI (2021) the period from the year 1996 until the year 2016. For the period from the year 2017 until the year 2020, the data were collected from GOES-R Space Weather | NCEI (2021). Solar flare data contain a wealth of information about the event, such as the date, coordinates, class, intensity, etc. However, for the current study, two main variables were chosen: the date of the event and the solar flare class. Table 3.6 lists the solar flare classes that were chosen for the study.

Table 3.6 Solar flares classes

Date	Solar flare class
1996-01-03	B
1996-01-03	B
…	…
2017-01-10	C
…	…
2020-01-10	B

3.4 Data Cleaning

3.4.1 Earthquake Data Cleaning

Based on the United States Geological Survey (USGS) website (Earthquakes, 2021), all sell in the data should not be empty. That is why, all data had been checked for empty items. The earthquake data did not have any missing values, except one row, that had been removed.

The precise location of the earthquake is not important in current investigation. As a result, the three variables employed in the current study are the date of the earthquakes, earthquake magnitude, and earthquake depth.

In this investigation, two different experimental setups were used. The first is a list of global earthquakes organised by Richter magnitude. Earthquakes were split into two groups based on the research of Odintsov, Ivanov-Kholodnyi, and Georgieva (2007). The first category of earthquakes includes those with a Richter magnitude of less than 5.5. The second category includes global earthquakes with a Richter magnitude of 5.5 or greater. According to the Novikov et al. (2020) experiment, electric current influences global earthquakes. For the second setting, in addition to sorting global earthquakes by Richter magnitude, the global earthquakes were sorted by depth, using the earthquake zones described in Chapter 2.2.1: shallow zone, intermediate zone, and deep zone. Table 3.7 shows an example of the earthquake data that has been cleaned. Figure 3.10 shows the data structure for earthquakes.

Table 3.7 Frequencies of Earthquakes by magnitude, shallow zone

Magnitude / t	M<5.5	M≥5.5	Shallow zone		Intermediate zone		Deep zone	
			M<5.5	M≥5.5	M<5.5	M≥5.5	M<5.5	M≥5.5
0	193	1	173	1	19	0	1	0
1	225	1	205	1	15	0	5	0
…	…	…	…	…	…	…	…	…
8716	510	1	475	1	31	0	4	0
8717	444	1	414	1	28	0	2	0

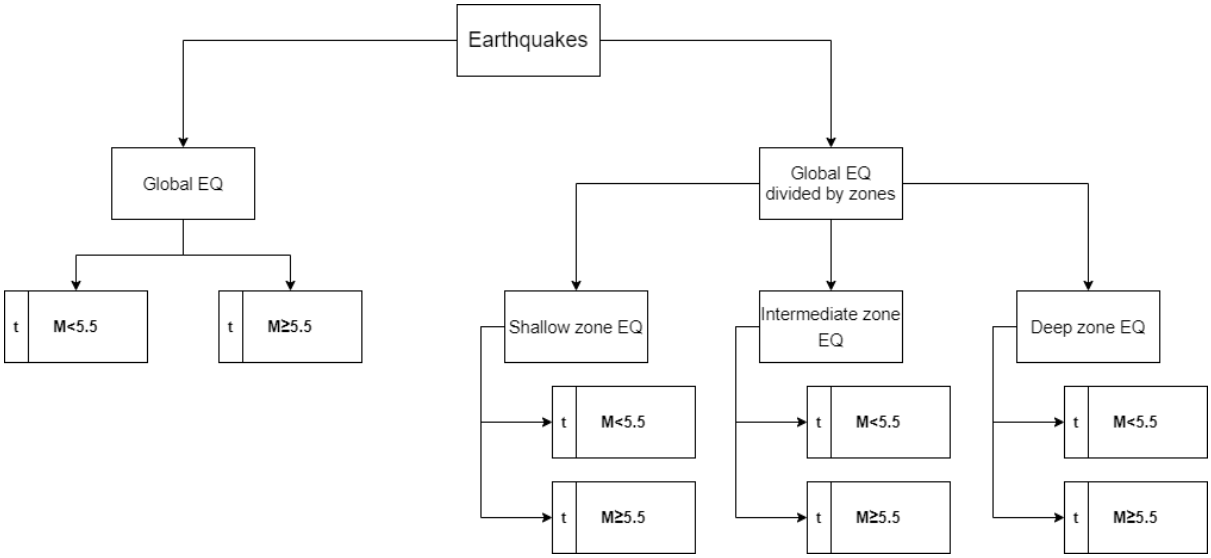


Figure 3.10 Structure of Earthquake data

In Figure 3.11 through Figure 3.18 all earthquake components are presented in box plots. It goes without saying that the data have outliers, and these outliers will have an impact on the final result. Moreover, a future fitted model can be significantly impacted by even a single value. Removing outliers from a dataset is one approach to deal with them, but it is not always the best approach (Spiegelhalter, 2019; Agresti, Franklin and Klingenberg, 2018). The current data outliers contain valuable information. The data outliers represent important events in the data. Removing them can lead to a loss of information and can affect the validity of the analysis (Novikov *et al.*, 2020). That is why normalisation techniques were used here. Normalisation is a technique that can be used to deal with outliers without removing them from the data (Muhamedyev, 2015). Normalisation involves transforming the data so that it has a specific distribution or range of values (Chapter 3.5). After the process of cleaning the data, the Earthquake dataset had 8718 records.

To summarise, for the earthquakes, as dependent variables there were two parts. The first part is global earthquakes divided by their magnitude:

- Global earthquakes with a Richter magnitude less than 5.5 – daily number of earthquakes
- Global earthquakes with a Richter magnitude greater than 5.5 – daily number of earthquakes

The second part is shallow depth earthquakes divided by their magnitude:

- Shallow depth earthquakes with a Richter magnitude less than 5.5 – daily number of earthquakes
- Shallow depth earthquakes with a Richter magnitude greater than 5.5 – daily number of earthquakes

The second part is intermediate depth earthquakes divided by their magnitude:

- Intermediate depth earthquakes with a Richter magnitude less than 5.5 – daily number of earthquakes
- Intermediate depth earthquakes with a Richter magnitude greater than 5.5 – daily number of earthquakes

The second part is deep depth earthquakes divided by their magnitude:

- Deep depth earthquakes with a Richter magnitude less than 5.5 – daily number of earthquakes
- Deep depth earthquakes with a Richter magnitude greater than 5.5 – daily number of earthquakes

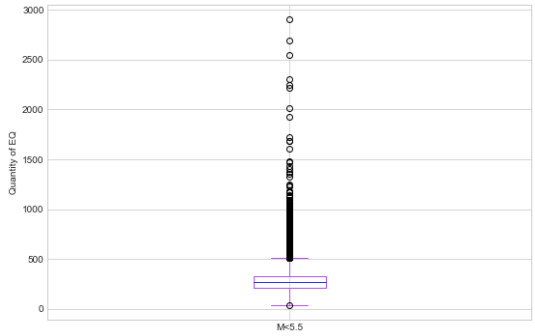


Figure 3.11 Global EQ, M < 5.5

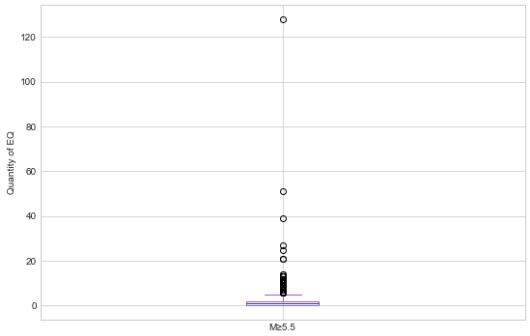


Figure 3.12 Global EQ, M ≥ 5.5

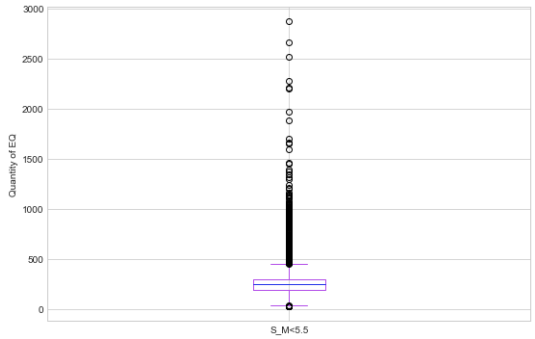


Figure 3.13 Shallow zone EQ, M < 5.5

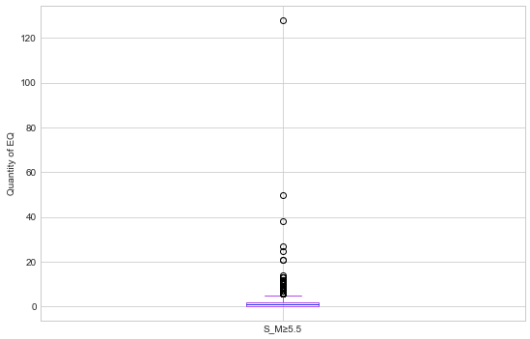


Figure 3.14 Shallow zone EQ, M ≥ 5.5

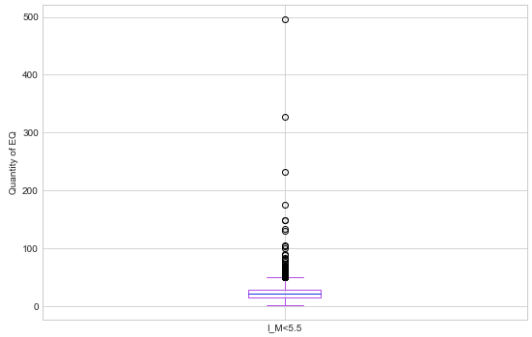


Figure 3.15 Intermediate zone EQ, M < 5.5

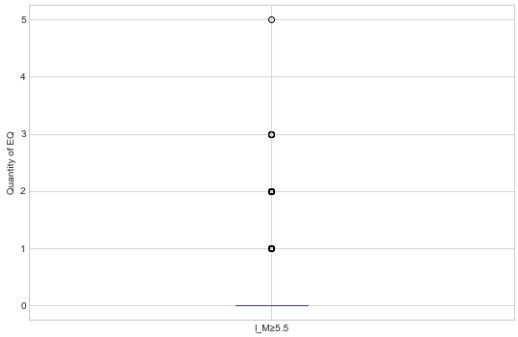


Figure 3.16 Intermediate zone EQ, M ≥ 5.5

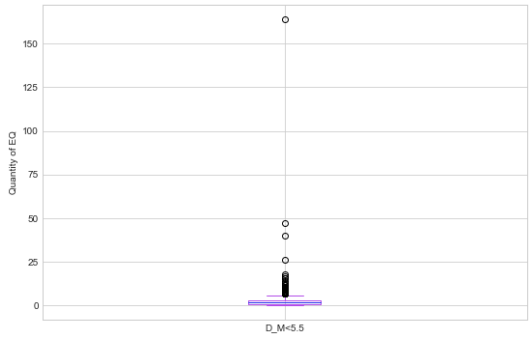


Figure 3.17 Deep zone EQ, M < 5.5

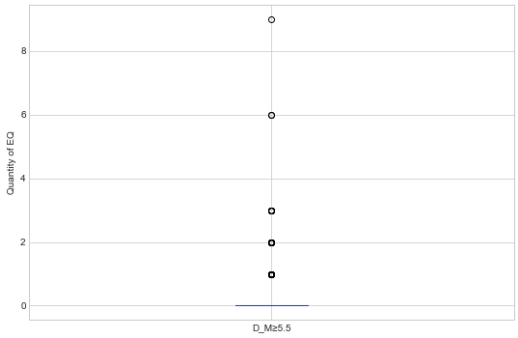


Figure 3.18 Deep zone EQ, M ≥ 5.5

3.4.2 Solar Activity Data Cleaning

According to the fact that solar activities were collected from three different open-source resources, solar activity cleaning data was broken into three parts, one for each resource.

The sunspot number data contain information about sunspot number from the year 1818 until the present. That is why the unnecessary data for the current study, which is outside the chosen period of the 23rd and 24th solar cycles, has been removed. According to the dataset description, the daily sunspot number column has negative values, which indicates that "no

numbers are available for that day" (*SILSO | World Data Center for the production, preservation and dissemination of the international sunspot number* 2021). That is why the data had been checked for negative values, which showed that there were not any negative values in the selected period.

The data for the three solar wind characteristics, which were chosen for the study (solar wind speed, proton density, and proton temperature), also contained values that were outside the chosen period, which had also been removed. The data did not contain any empty values; however, the data could include the values "999.99," "9999.99," and "9999999.99," which indicate the absence of the values. That is why these empty values were found and changed to "NaN" values for later removal.

A few completely empty rows in the solar flare data were deleted. In addition, in the current study, the frequency of the solar flares was classified by their classes. Table 3.8 gives an example of the solar activity data.

Table 3.8 Solar activity data

Date	Sunspot Number	Solar Wind Speed	Proton Density	Proton Temperature	A-class	B-class	C-class	M-class	X-class
1996/01/01	0	403	7.9	7202	NaN	NaN	NaN	NaN	NaN
1996/01/02	14	398	8.0	77660	NaN	NaN	NaN	NaN	NaN
…	…	…	…	…	…	…	…	…	…
2020/12/30	32	483	3.6	122507	NaN	NaN	NaN	NaN	NaN
2020/12/30	34	406	3.3	44521	NaN	NaN	NaN	NaN	NaN

As the solar activity data had empty values, these empty values have been removed. Figure 3.19 illustrates the structure of the solar activity data. The final step, after merging data and removing empty values, was the union of earthquake data and solar activity data. For future use, the dataset was saved in "*.tsv" format. The final version of the earthquakes and solar activity dataset is shown in Appendix B Figure B - 1.

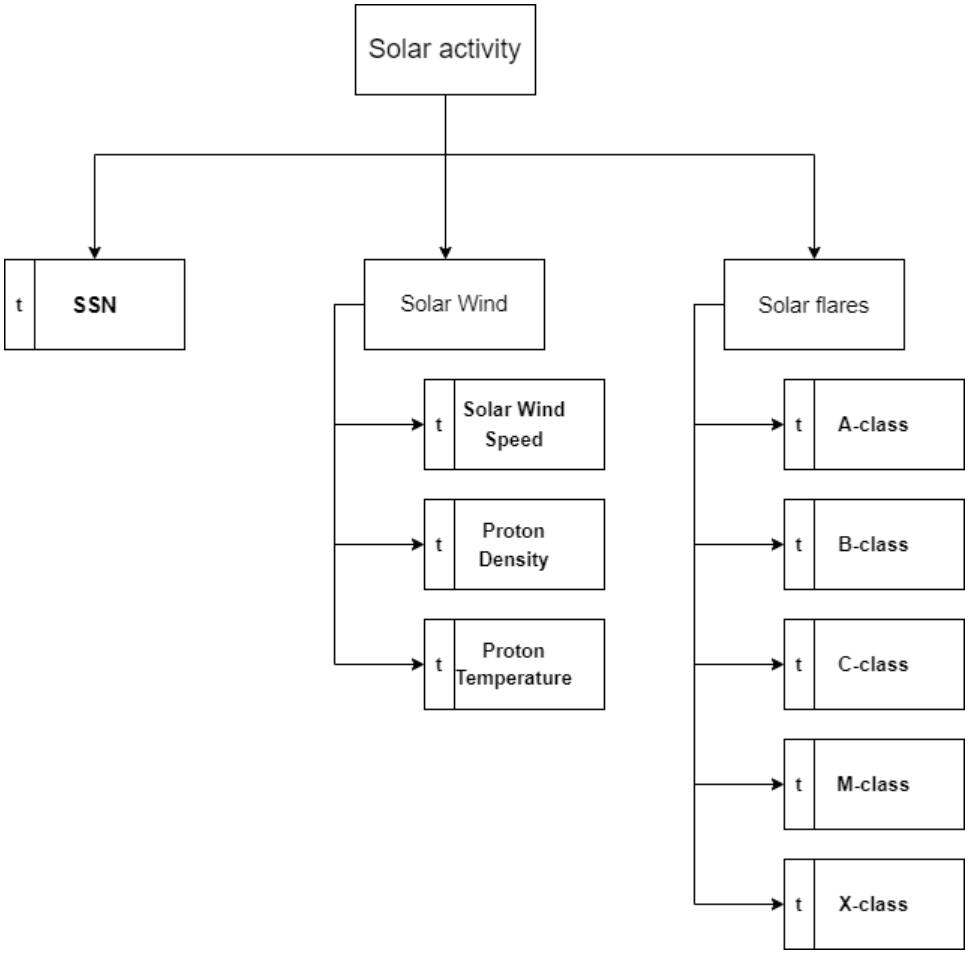


Figure 3.19 Structure of Solar activity data

The boxplots of the original solar activity data shown in Figure 3.20 through Figure 3.28 show that each independent variable includes outliers. As was mentioned in Chapter 3.4.1 the outliers will impact the final result, and it was a decision to leave outliers. In the case of solar activity, it can be a high-speed solar wind, which has a significant impact on earthquakes (Odintsov *et al.*, 2006). After the process of cleaning the data, the Solar activity dataset had 8718 records.

To summarise, for the solar activity, as independent variables there were chosen:

- Sunspot number – daily total sunspot number
- Solar wind speed – daily averages
- Proton density – daily averages
- Proton temperature – daily averages
- Solar flares A-class – quantity of solar flares A-class per a day
- Solar flares B-class – quantity of solar flares B-class per a day
- Solar flares C-class – quantity of solar flares C-class per a day

- Solar flares M-class – quantity of solar flares M-class per a day
- Solar flares X-class – quantity of solar flares X-class per a day

These solar activity events were chosen based on the previous seismological studies' findings (Odintsov *et al.*, 2006; Novikov *et al.*, 2020; Novikov *et al.* 2017; Odintsov, Ivanov-Kholodnyi and Georgieva 2007; Sytinskii's 1973).

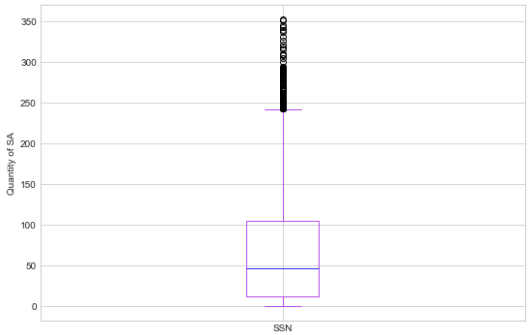


Figure 3.20 SSN

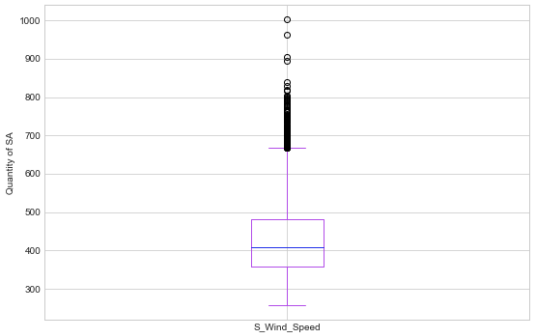


Figure 3.21 Solar wind speed

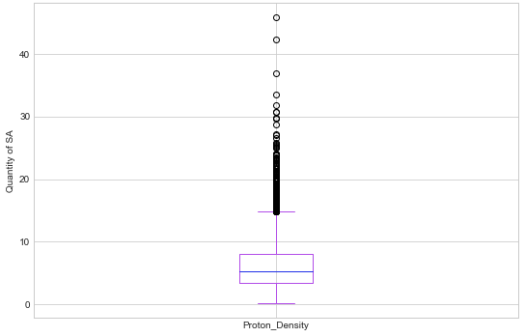


Figure 3.22 Proton density

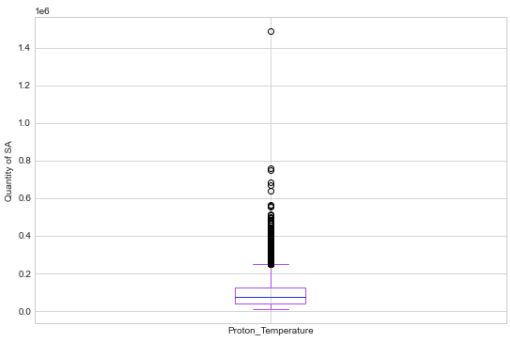


Figure 3.23 Proton temperature

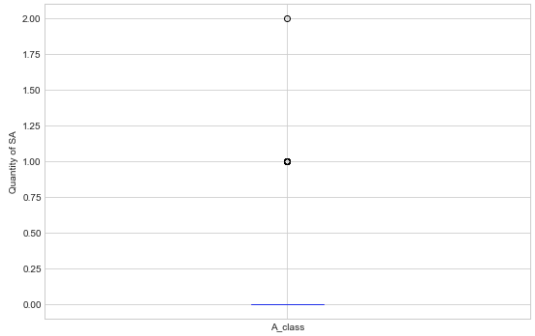


Figure 3.24 Solar flares A class

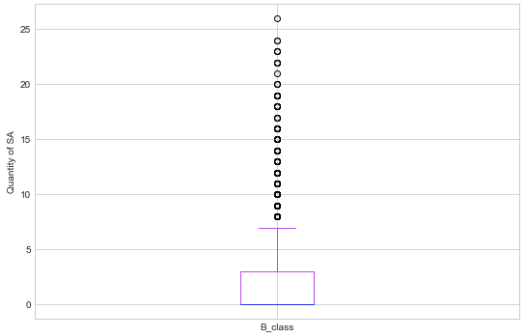


Figure 3.25 Solar flares B class

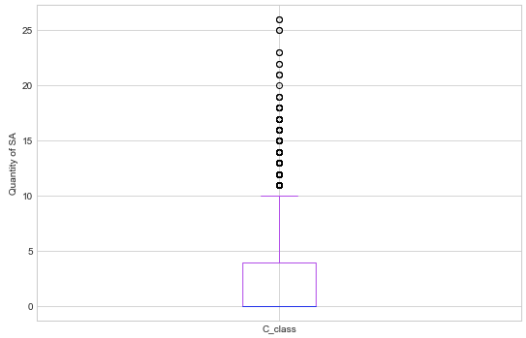


Figure 3.26 Solar flares C class

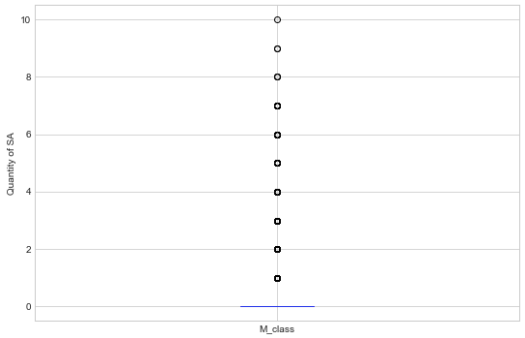


Figure 3.27 Solar flares M class

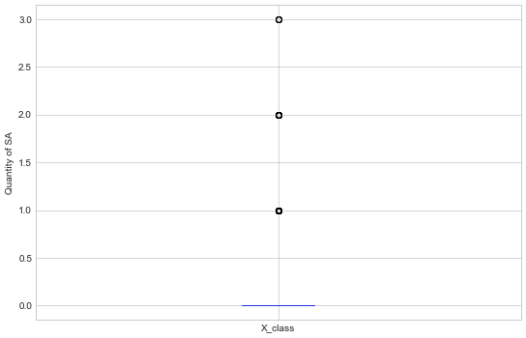


Figure 3.28 Solar flares X class

3.5 Normalising data

3.5.1 Normalising earthquake data (dependent variables)

As was mentioned in Chapter 3.4.1, it was decided to leave earthquake outliers, based on Novikov *et al.* (2020) study. To determine which normalisation scaler is optimal for the data, box plots were used to compare the scaler findings. From Figure 3.29 through Figure 3.34, the box plots of the normalised results are shown. As can be seen from these graphs, the "Quantile Transformer" scaler produced the highest normalising result.

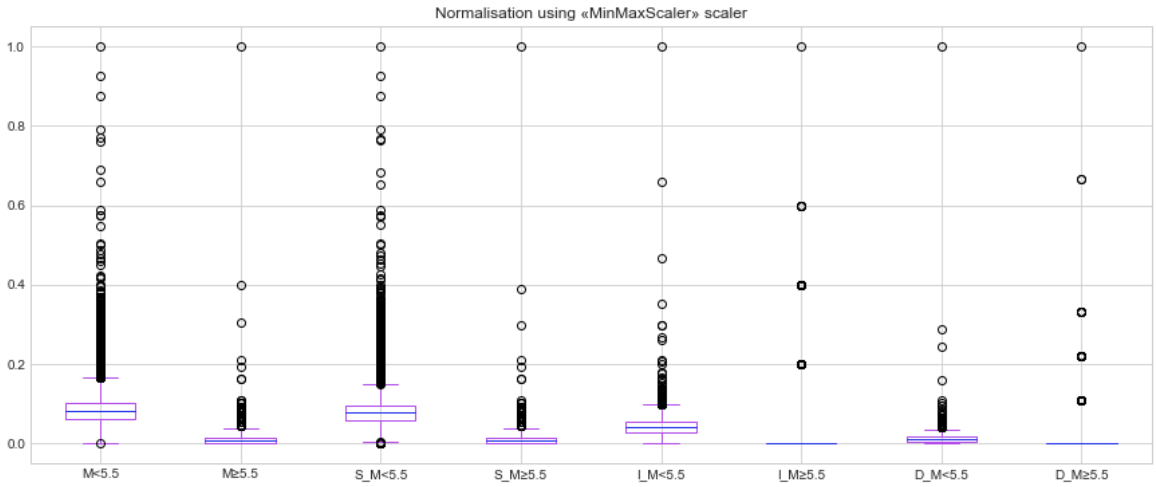


Figure 3.29 Dependent variables: Normalisation using "MinMaxScaler" scaler after normalising

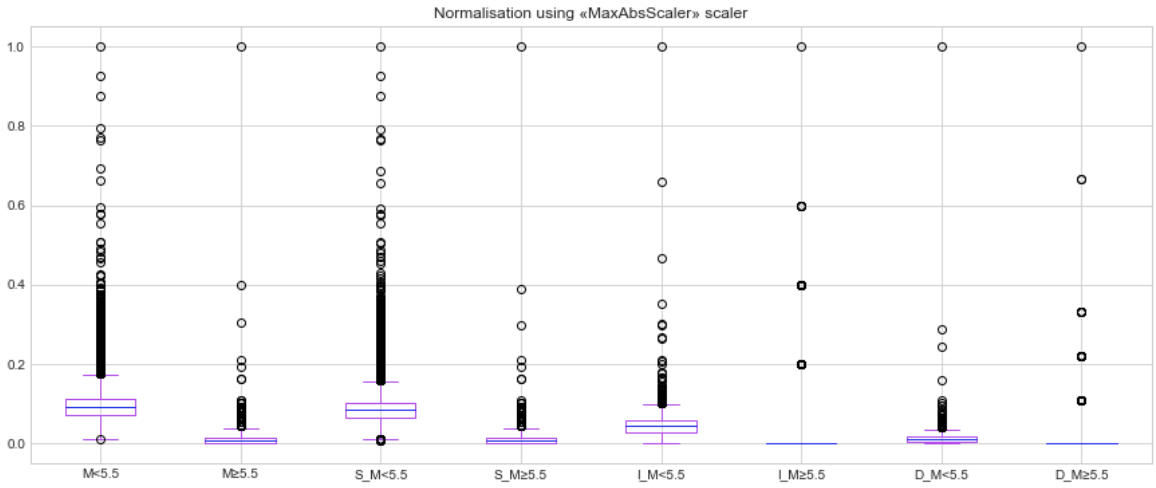


Figure 3.30 Dependent variables: Normalisation using "MaxAbsScaler" scaler after normalising

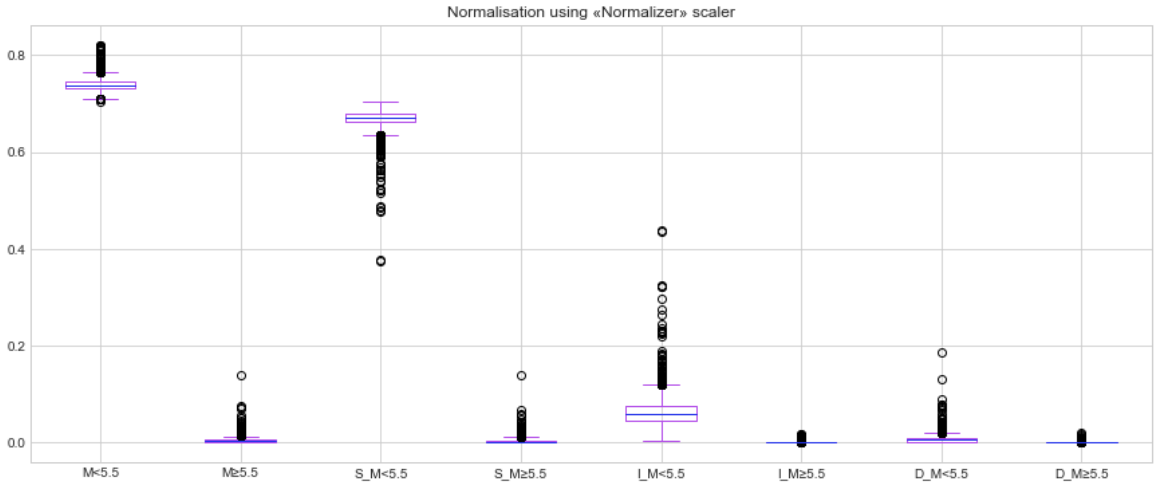


Figure 3.31 Dependent variables: Normalisation using "Normalizer" scaler after normalising

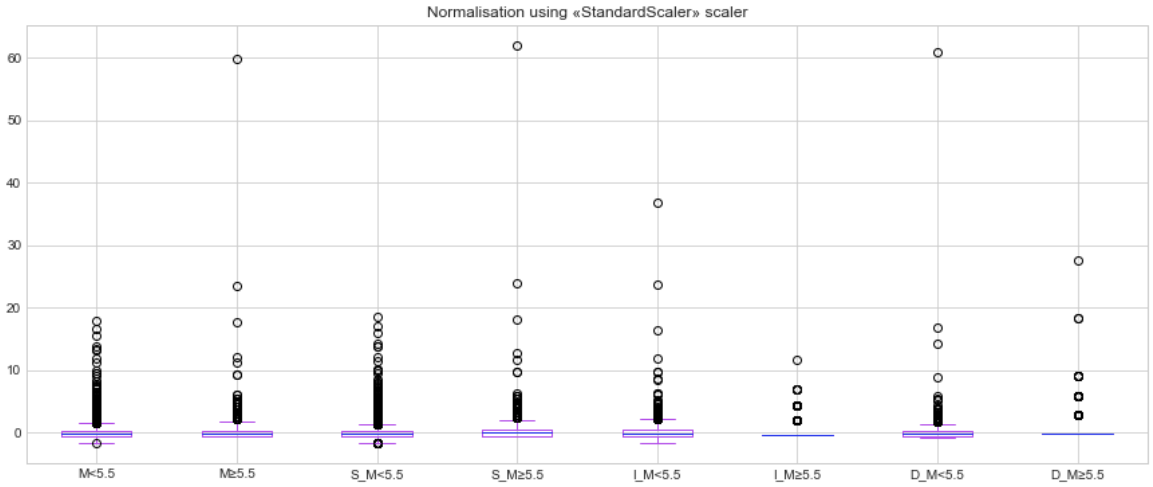


Figure 3.32 Dependent variables: Normalisation using "StandardScaler" scaler after normalising

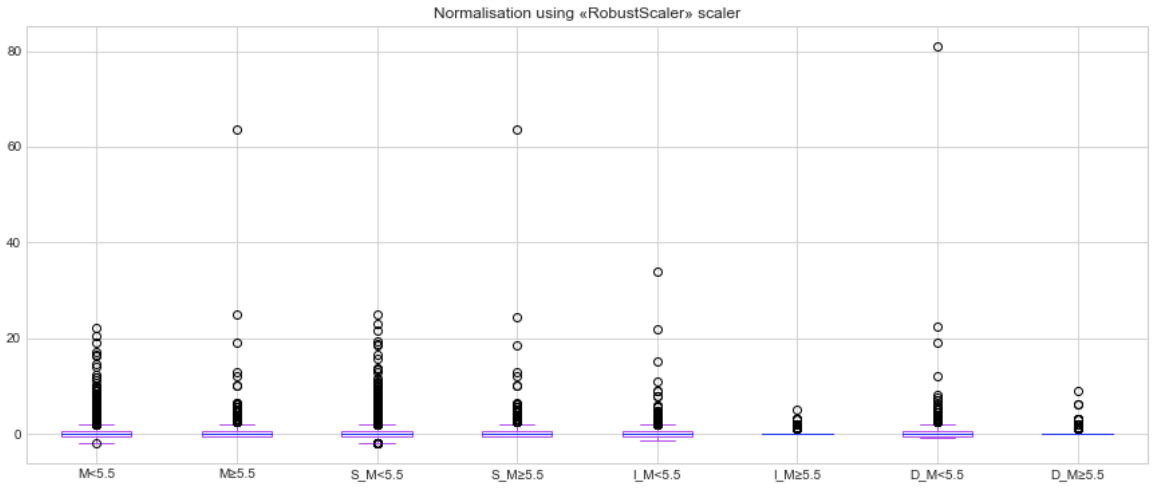


Figure 3.33 Dependent variables: Normalisation using "RobustScaler" scaler after normalising

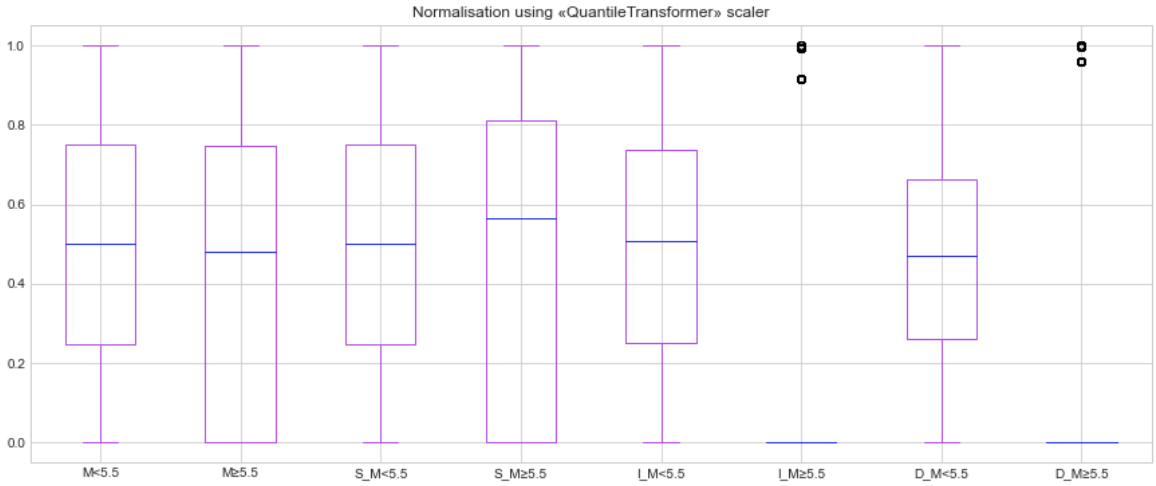


Figure 3.34 Dependent variables: Normalisation using "Quantile Transformer" scaler after normalising

3.5.2 Normalising solar activity data (independent variables)

The same normalising comparison had been run with the independent variables, solar activity, as it had been done with the dependent variables, earthquakes. From Figure 3.35 until Figure 3.40, box plots depict the outcomes of the normalising scalers. Similar to the earthquake data, the "Quantile Transformer" scaler produced the highest normalisation results. That is why the "Quantile Transformer" scaler for the normalising process was employed in the experimental section of the study.

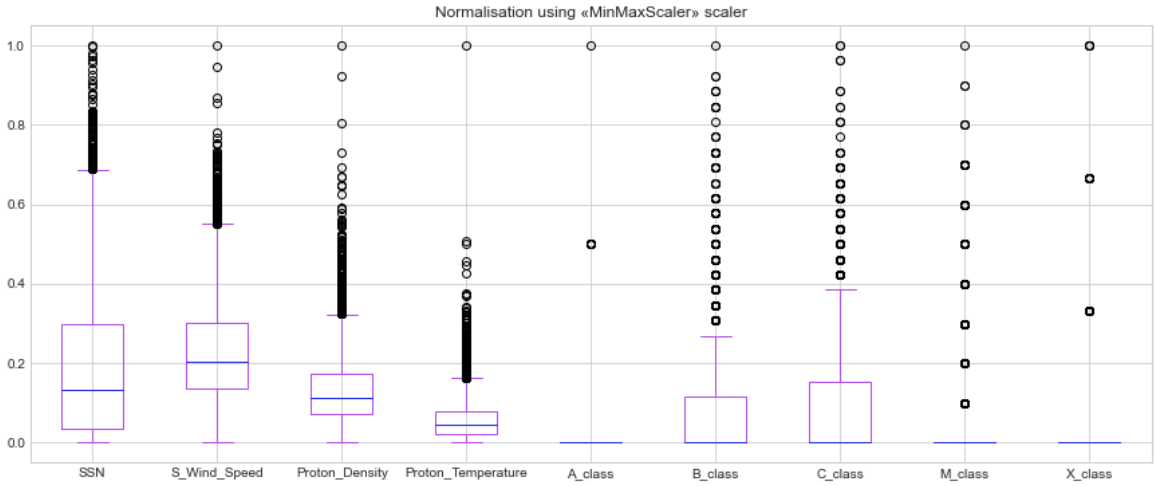


Figure 3.35 Independent variables: Normalising using "MinMaxScaler" scaler after normalising

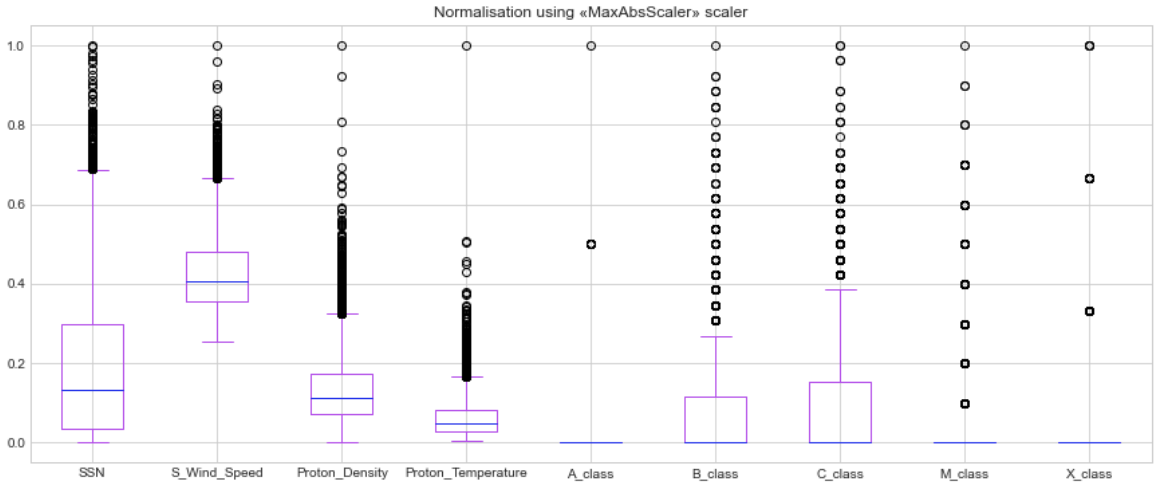


Figure 3.36 Independent variables: Normalising using "MaxAbsScaler" scaler after normalising

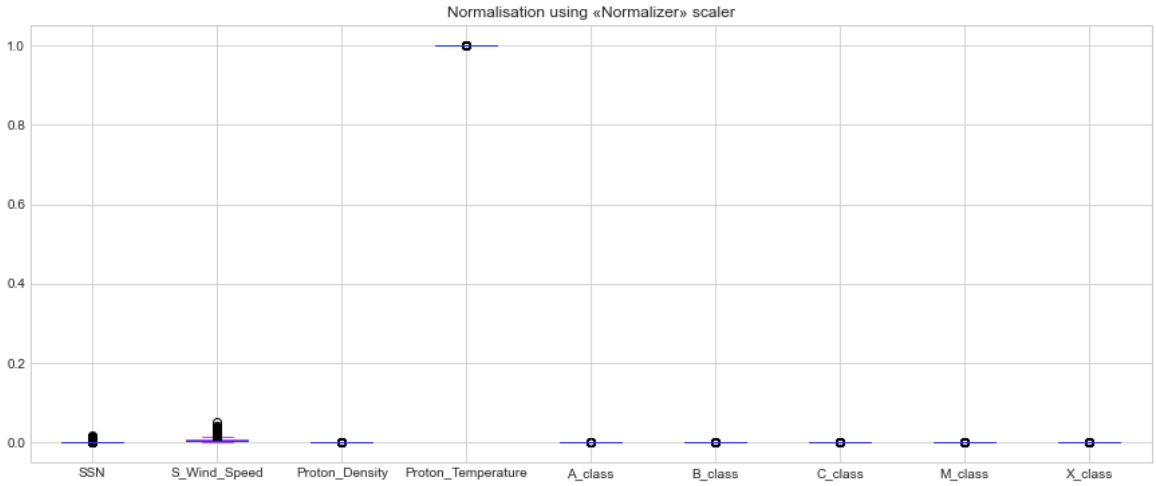


Figure 3.37 Independent variables: Normalising using "Normalizer" scaler after normalising

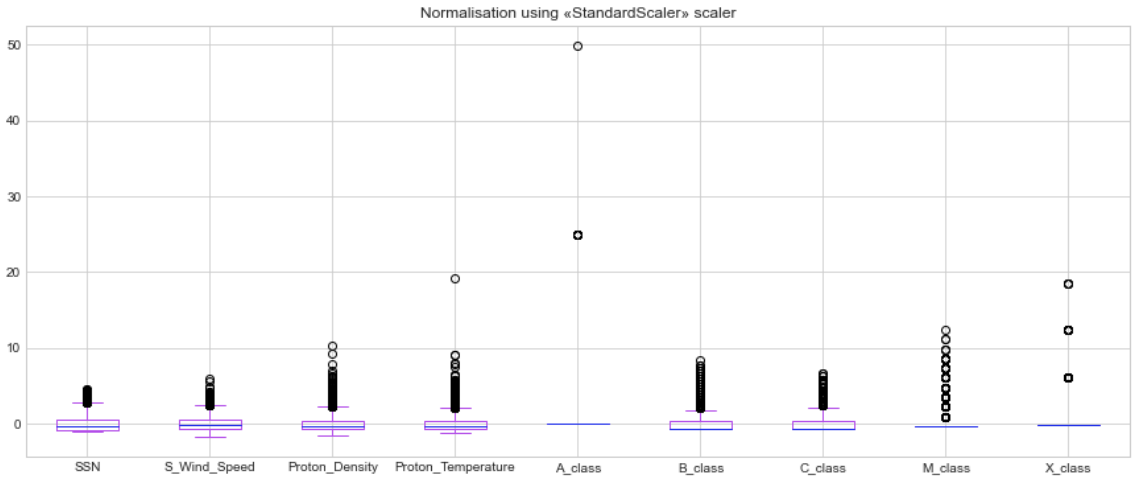


Figure 3.38 Independent variables: Normalising using "StandardScaler" scaler after normalising

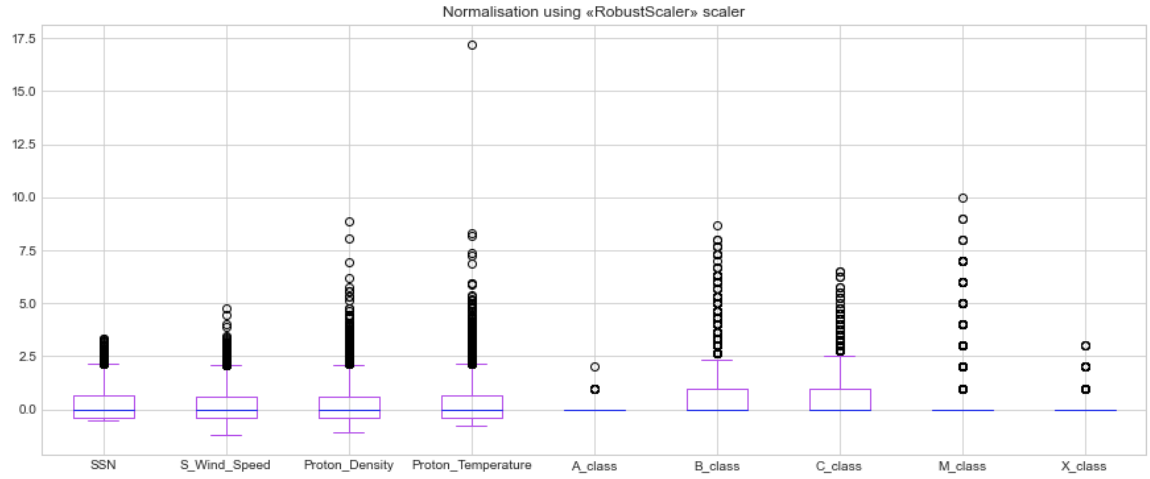


Figure 3.39 Independent variables: Normalising using "RobustScaler" scaler after normalising

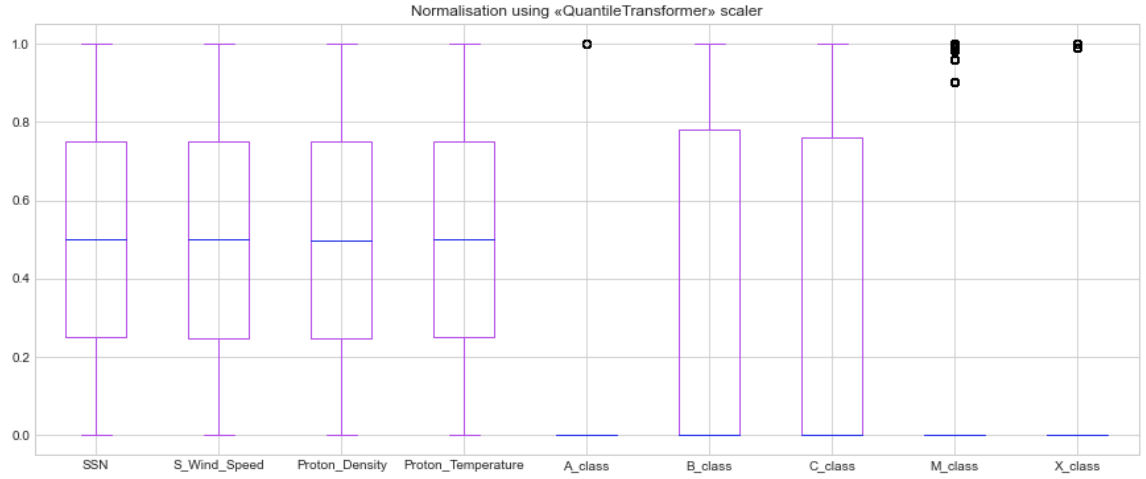


Figure 3.40 Independent variables: Normalising using "Quantile Transformer" scaler after normalising

3.6 Dimensional reduction of solar activity data

As previously stated, Nishii, Qin, and Kikuyama (2020) proposed in their variable reduction study that not all solar activity events influence earthquakes equally. The solar activity dataset has nine variables. Moreover, the aim is to explain the earthquake (dependent) variables as a function of the solar activity (independent) variables. That is why the earthquake variables should not be included in the PCA.

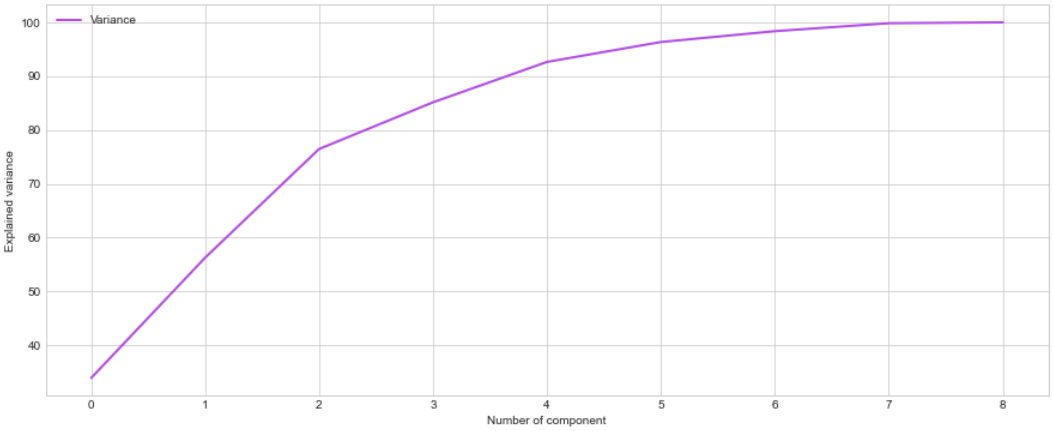


Figure 3.41 Choosing the number of SA variables

Firstly, it was examined whether all solar activity variables have an effect on earthquake variables. Figure 3.41 shows that the first six principal components keep over 96 percent (96.343 percent) of the variability in the independent variables. The remaining three variables may be co-correlated or not contribute much. That is why, the independent variables can be reduced by three features (Figure 3.42).

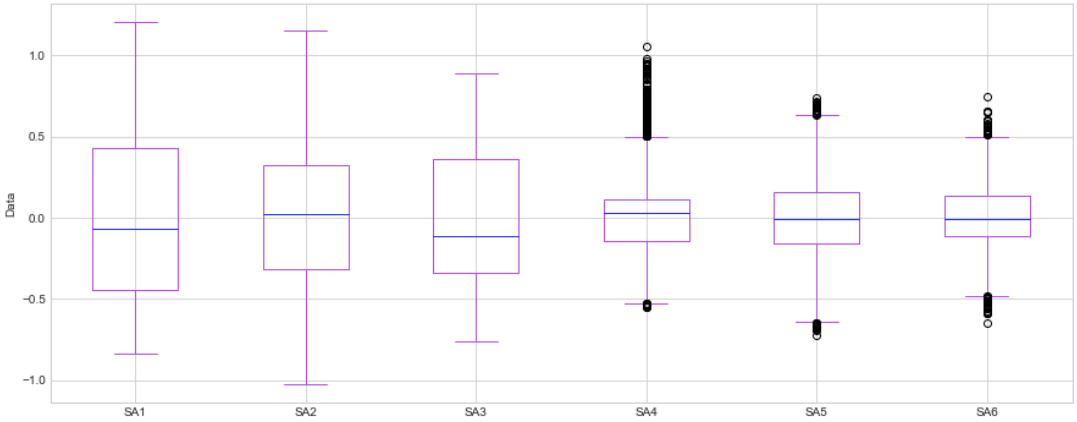


Figure 3.42 Independent variables after dimensionally reduction

3.7 Linear and non-linear relationships between earthquake and solar activity data

Before applying machine learning algorithms to each type of data, the variables in the data were checked for linear/nonlinear relationships using the Linear Regression machine learning algorithm, *equation (26)* and the R-squared (R^2) error, *equation (27)*.

$$\acute{EQ} = a_0 + a_1 * SA1 + a_2 * SA2 + a_3 * SA3 + a_4 * SA4 + a_5 * SA5 + a_6 * SA6 \qquad (26)$$

Where:

$\acute{EQ}$ – predicted earthquake frequency

$SA1 - SA6$ – solar activity data points

$a_0, a_1, a_2, a_3, a_4, a_5, a_6$ – coefficients

$$R^2 = \frac{\sum_{i=1}^{n}(\acute{EQ} - \overline{\acute{EQ}})^2}{\sum_{i=1}^{n}(EQ - \overline{EQ})^2} \qquad (27)$$

Where:

n – data size

$\acute{EQ}$ – predicted earthquake frequency

EQ – actual earthquake frequency

All of the above calculations and coefficient determinations were carried out using the Python library Scikit-learn (Pedregosa et al., 2011; Supervised learning—Scikit-learn 0.24.2 documentation, 2021). The results are presented in Table 3.9. All R^2 values in Table 3.9 are negative, and the highest value of R^2 equals -13.01. A negative R^2 value in a linear regression model indicates that the model is performing worse than a simple mean model that uses the mean value of the dependent variable to predict its value. An R^2 value of -13.01 is particularly low and suggests that the model is not a good fit for the data (Kutner, 2005). It can be assumed that the relationships between solar activity data and earthquake data in all earthquake

categories are non-linear. That is why algorithms that work well with nonlinear relationships are preferred.

Table 3.9 Linear and nonlinear relationships, R^2 values

Global						
Solar Activity / EQ category	SA: 2 days delay	SA: 3 days delay	SA: 4 days delay	SA: 5 days delay	SA: 6 days delay	SA: 7 days delay
M<5.5	-13.49	-13.21	-13.01	-13.28	-13.33	-13.41
M≥5.5	-1619.85	-1376.71	-811.66	-630.0	-1075.8	-1431.76
Shallow Zone						
Solar Activity / EQ category	SA: 2 days delay	SA: 3 days delay	SA: 4 days delay	SA: 5 days delay	SA: 6 days delay	SA: 7 days delay
M<5.5	-14.24	-13.8	-13.56	-13.85	-13.98	-14.08
M≥5.5	-965.43	-813.24	-892.07	-916.75	-809.8	-741.1
Intermediate Zone						
Solar Activity / EQ category	SA: 2 days delay	SA: 3 days delay	SA: 4 days delay	SA: 5 days delay	SA: 6 days delay	SA: 7 days delay
M<5.5	-16.22	-16.16	-16.35	-16.45	-16.33	-16.11
M≥5.5	-1446.15	-1709.9	-625.39	-480.75	-883.75	-1050.33
Deep Zone						
Solar Activity / EQ category	SA: 2 days delay	SA: 3 days delay	SA: 4 days delay	SA: 5 days delay	SA: 6 days delay	SA: 7 days delay
M<5.5	-55.47	-52.88	-53.37	-51.87	-50.83	-54.61
M≥5.5	-1683.91	-1136.28	-1129.27	-1074.42	-1809.56	-3828.17

3.8 Determining the method for the model measurement error

Spiess and Neumeyer (2010) study showed, that R^2 error should not be used in nonlinear data analysis. The study used various simulation models and found that R-squared leads to false conclusions about which nonlinear models are better. A good way of comparing different models is by using errors, which provide a relative measure of the percentage. It was found that, as the "Deep zone" category of EQ has a lot of zero values, MAPE, *equation (2),* is not suitable here since using MAPE would give a division by zero. However, there is no division by zero using the RMSE *equation (28)* and MAE *equation (29)* for the calculations of margin error in the data.

$$RMSE_{EQ} = \sqrt{\frac{\sum_{i=1}^{n}(\acute{EQ}_i - EQ_i)^2}{n}} \qquad (28)$$

Where:

n – the number of data points

$\acute{EQ}_1, \acute{EQ}_2, \dots, \acute{EQ}_n$ – predicted earthquake frequency

$EQ_1, EQ_2, \dots, EQ_n$ – actual earthquake frequency

$$MAE_{EQ} = \frac{\sum_{i=1}^{n}\left|\acute{EQ}_i - EQ_i\right|}{n} \qquad (29)$$

Where:

n – the number of data points

$\acute{EQ}_1, \acute{EQ}_2, \dots, \acute{EQ}_n$ – predicted earthquake frequency

$EQ_1, EQ_2, \dots, EQ_n$ – actual earthquake frequency

However, Willmott and Matsuura (2005) indicated that RMSE is not a good measure of model performance and may be a deceptive indicator of average error. That is why they suggested MAE is a preferable metric. On the other hand, Chai and Draxler (2014) stated that avoiding RMSE is not the appropriate practice. Also, RMSE avoids using absolute values, which is a benefit over MAE. Also, they suggested that for evaluating model performance, it's better to

use a variety of metrics. That is why it was decided to use both errors, MAE and RMSE, for the evaluation of the study's model.

As it was needed to compare the models with different dependent variables, MAE and RMSE are not useful for this purpose. That is why the normalised RMSE (NRMSE) was used here. There are few ways to normalise the RMSE (Shcherbakov *et al.*, 2013):

- Normalisation by the difference between the 75th and 25th percentile, the interquartile range of the data as presented in *equation (30)*.
- Normalisation by the difference between the maximum and minimum of the data, *equation (31)*
- Normalisation by the data mean, *equation (32)*.
- Normalisation by the standard deviation, *equation (33)*.

$$NRMSE_{EQ} = \frac{RMSE_{EQ}}{Q3 - Q1} \qquad (30)$$

Where:

$RMSE_{EQ}$ – RMSE of a model

$Q1, Q3$ – 25th and 75th percentile of the data

$$NRMSE_{EQ} = \frac{RMSE_{EQ}}{EQ_{max} - EQ_{min}} \qquad (31)$$

Where:

$RMSE_{EQ}$ – RMSE of a model

EQ_{max}, EQ_{min} – maximum and minimum of the data

$$NRMSE_{EQ} = \frac{RMSE_{EQ}}{\overline{EQ}} \qquad (32)$$

Where:

$RMSE_{EQ}$ – RMSE of a model

$\overline{EQ}$ – mean of EQ dependent variable

$$NRMSE_{EQ} = \frac{RMSE_{EQ}}{SD} \quad (33)$$

Where:

$RMSE_{EQ}$ – RMSE of a model

SD – standard deviation of the data

However, normalising the data by its interquartile range is not useful in this case because, in some cases, both the 25th and 75th percentiles are equal to zero, resulting in division by zero. Also, normalisation by the difference between maximum and minimum is not the appropriate option in this case, because the data were normalised using the Quantile Transformer scaler, $EQ_{max} = 1$ and $EQ_{min} = 0$. That is why RMSE will always be equal to NRMSE normalised by the difference between maximum and minimum. That is why normalisations by mean and standard deviation are suitable in the current case.

MAE, like RMSE, can be normalised using the mean, interquartile range, and difference between the maximum and minimum values. Normalisation MAE (NMAE) employing an interquartile range and difference between maximum and minimum, similar to RMSE, is not appropriate here. As a result, the mean was used to calculate NMAE, *equation (34)*.

$$NMAE_{EQ} = \frac{MAE_{EQ}}{\overline{EQ}} \quad (34)$$

Where:

MAE_{EQ} – RMSE of a model

$\overline{EQ}$ – mean of EQ dependent variable

3.9 Machine learning algorithm used in the study

For the experiment, classic algorithms as well as deep learning methods were used. There were two main requirements for selecting algorithms. The first requirement is that algorithms should have different approaches to solving a problem that help solve the problem using various techniques. The second is that these methods should be good at solving issues with non-linear relationships between the variables. Based on the above requirements, the classic algorithms K-Nearest Neighbour (KNN), Support Vector Regression (SVR), and Random Forest Regression (RFR) were chosen, and the Long Short-Term Memory Network (LSTM) was chosen for the neural networks. It should be noted that all these machine learning methods have been successfully applied in many fields.

3.9.1 K-nearest neighbour algorithm (regression)

KNN is one of the simplest algorithms, with one of the fastest training speeds. KNN employs Euclidean Distance. Even though the accuracy of prediction decreases as the number of predictors increases, the decision was made to start with the KNN algorithm because it is good for data exploration and works well with non-linear relationships. Moreover, the data have 8718 records, which is not critical for the KNN algorithm.

The first step is the calculation of the Euclidian distance between the new data point and the existing point equation *(6).* The second step is to find a predicted value of earthquakes that is equal to the average of earthquakes for the top "K" neighbours. All of the calculations related to fitting the model and determining the prediction values were performed using the Python library Scikit-learn (Pedregosa *et al.*, 2011; *Supervised learning — scikit-learn 0.24.2 documentation*, 2021).

To determine the appropriate value of "K", a plot of the RMSE values versus the "K" values was created in each scenario, Figure 1.1 – Figure 3.50. An array of "K" values ranging from 1 to 20 was used to create the charts. For each "K" value, a model was trained, and afterward KNN was used to make a prediction. After that, the RMSE value was determined. As can be seen from Figure 3.43 through Figure 3.50, after "K" equals 17, the value of the error does not change significantly. That is why the decision was made to choose the value of "K" to be equal to 17. Also, "K" must be a whole number, as "K" is the number of nearest datapoints to the new data point.

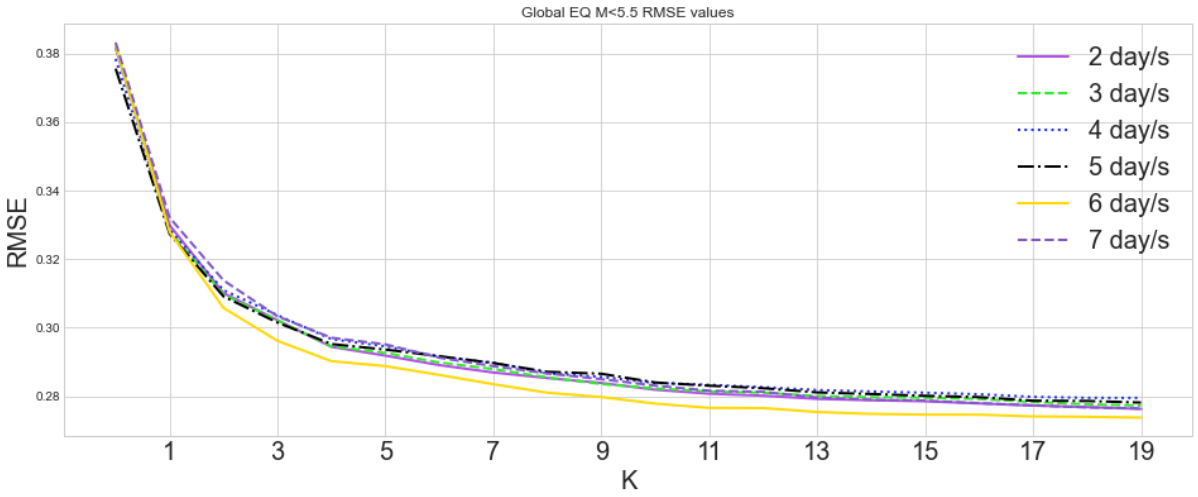


Figure 3.43 Finding the most appropriate value of “K” Global EQ M<5.5

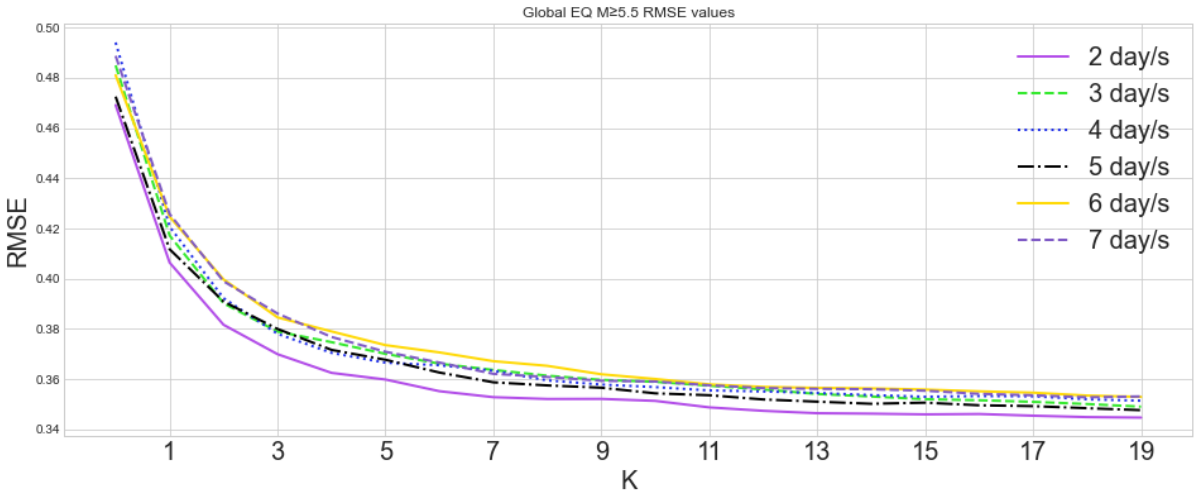


Figure 3.44 Finding the most appropriate value of “K” Global EQ M≥5.5

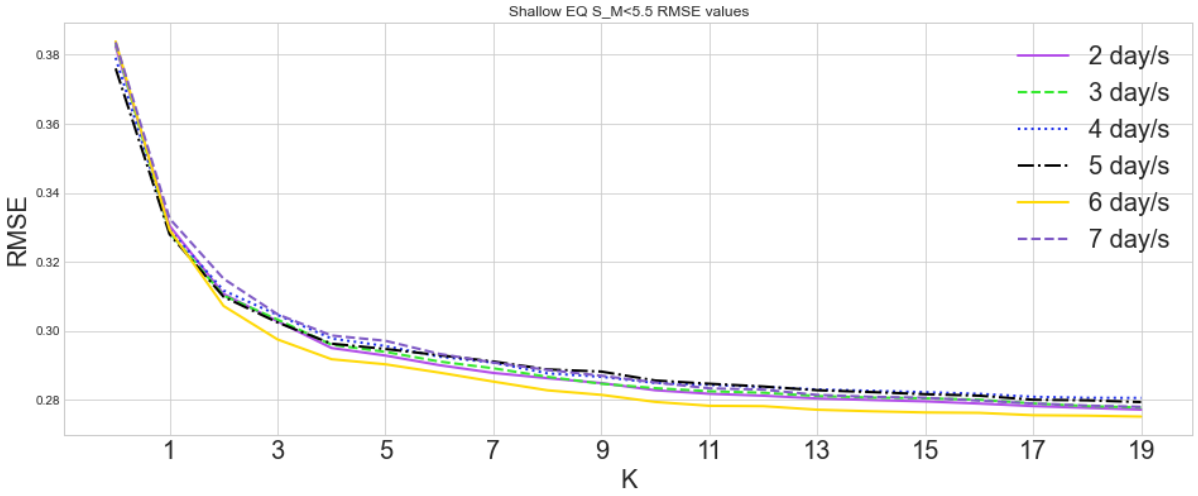


Figure 3.45 Finding the most appropriate value of “K” Shallow zone EQ M<5.5

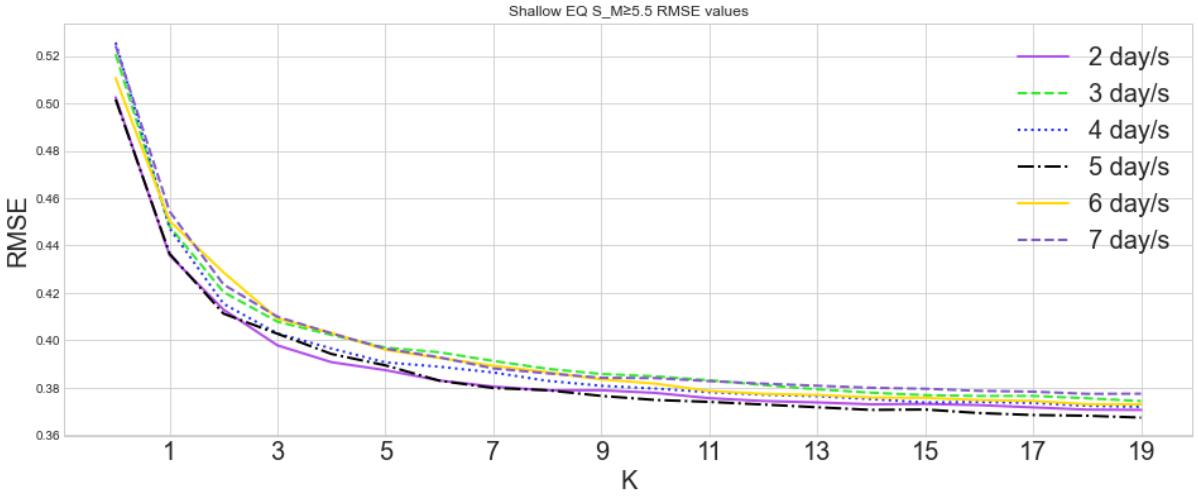


Figure 3.46 Finding the most appropriate value of “K” Shallow zone EQ M≥5.5

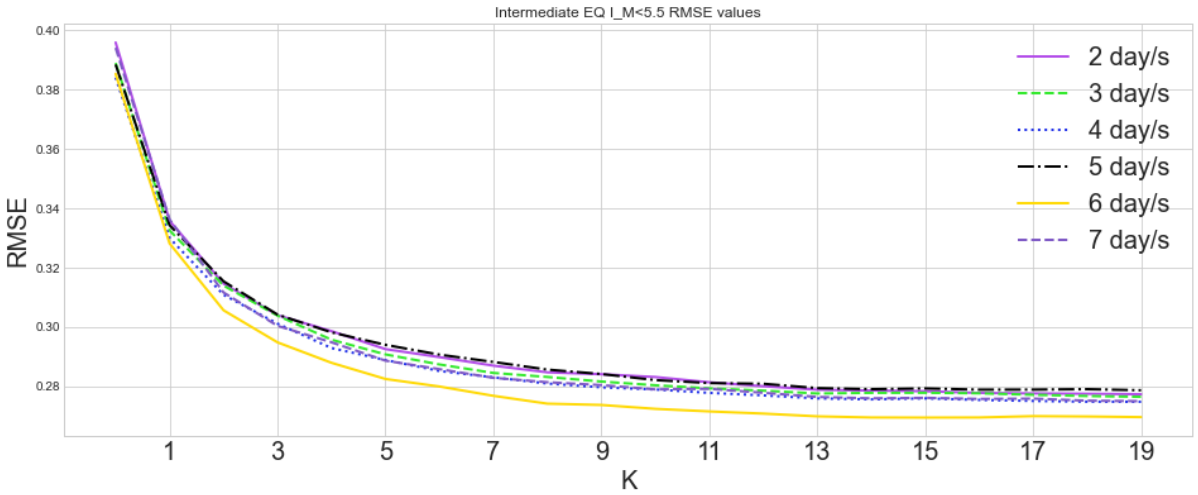


Figure 3.47 Finding the most appropriate value of “K” Intermediate zone EQ M<5.5

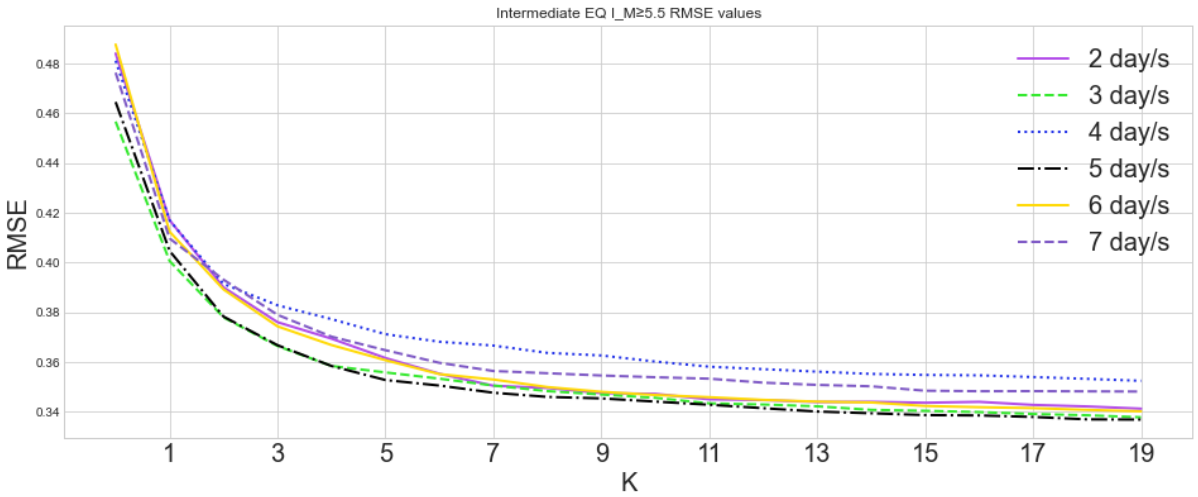


Figure 3.48 Finding the most appropriate value of “K” Intermediate zone EQ M≥5.5

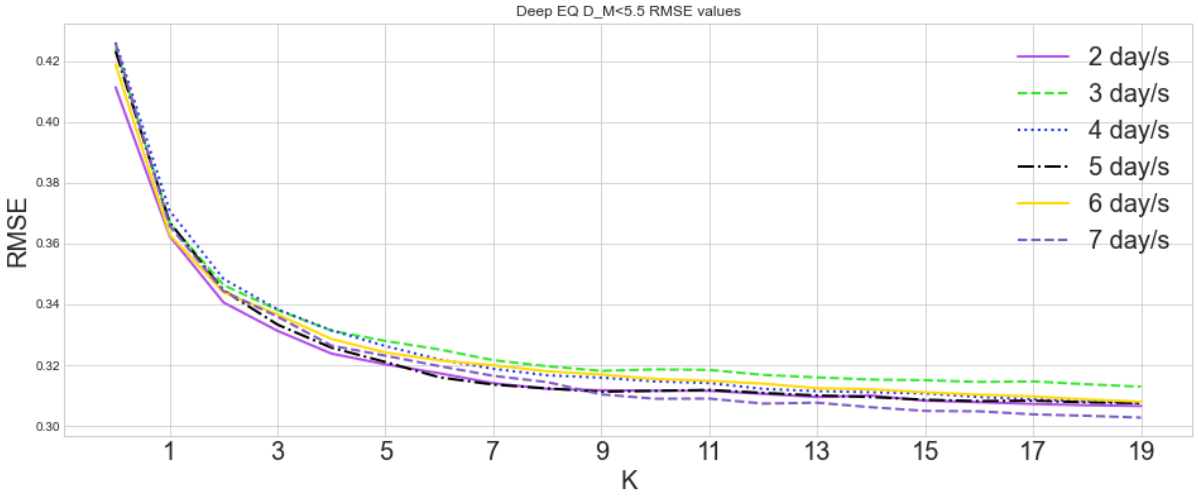


Figure 3.49 Finding the most appropriate value of “K” Deep zone EQ M<5.5

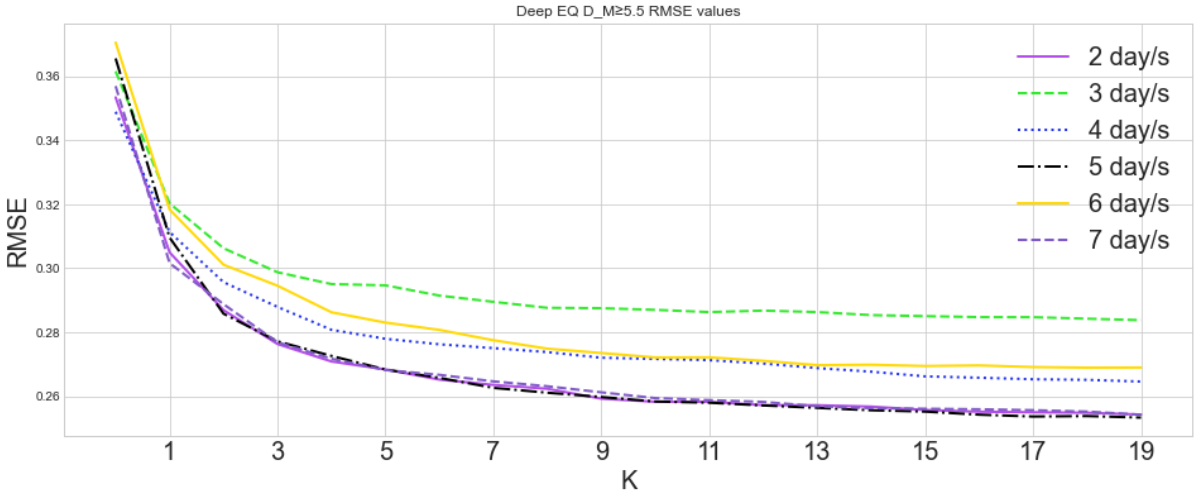


Figure 3.50 Finding the most appropriate value of “K” Deep zone EQ M≥5.5

3.9.2 Support vector regression algorithm

The SVR algorithm is a kernel-based algorithm. The most important parameter is the kernel type. Hyperparameters, such as kernel, can be set in SVR. That is what makes SVR so effective in dealing with non-linear relationships between dependent and independent variables. Due to the fact that the dependent and independent variables have non-linear relationships, the RBF kernel has been chosen; the equation is based on (Smola and Schölkopf, 2004).

Two vectors were defined for the experiment: vector x = (SA1, SA2, SA3, SA4, SA5, SA6), where SA1 – SA6 frequency of solar activity of the data points and vector y = (EQ), where EQ frequency of earthquakes of the data points. The Python library Scikit-learn (Pedregosa *et*

al., 2011; *Supervised learning — scikit-learn 0.24.2 documentation*, 2021) was used to accomplish all of the calculations related to fitting the model and generating the prediction values.

3.9.3 Random forest regression (RFR)

RFR is an example of ensemble learning. RFR has no formal distributional assumptions and works well with non-linear relationships. There are four steps involved in RFR. In the first stage, the data were divided into subsets. In the second phase, an individual decision tree was generated for each subgroup. The third step includes a result from each subset. The final step is an average of all subset outcomes. All of these steps related to fitting the model and obtaining the prediction values were completed using the Python library Scikit-learn (Pedregosa *et al.*, 2011; *Supervised learning — scikit-learn 0.24.2 documentation*, 2021).

According to Rodriguez-Galiano *et al.* (2012), the most important parameters for RFR are the number of regression trees and the number of features needed at each node to make regression trees develop. The number of regression trees was equal to 100 (the scikit-learn default number), as RFR does not have overfitting. The number of features equals two because, when the number of features is reduced, the correlation between trees is reduced, which improves the model's accuracy.

3.9.4 Long Short-Term Memory network

The LSTM algorithm is a neural network (NN) algorithm that works well with non-linear functions. A typical sigmoid transfer function was used to create a number of neural network models, which is why the sigmoid function was also used in the study, the sigmoid function is default in *Keras: the Python deep learning API* (2021) settings.

The number of hidden layers and nodes per layer need to be chosen as a structure before training the LSTM. The number of hidden layers is determined by no rule. In general, the more hidden layers there are, the better the network's ability to represent training data patterns. However, the large number of units in the hidden layer reduces the generalisation power of the networks and increases the cost (Verdhan and Kling, 2020). That is why, two hidden layers were chosen. The number of hidden nodes was calculated using the equation in *Keras: the Python deep learning API* (2021), *equation (35)*

$$N_h = \frac{N_s}{(\alpha * (N_i + N_o))} \quad (35)$$

Where:

N_s – number of samples in training data set

N_i – number of input neurons

N_o – number of output neurons

α – scaling factor (minimum 2, maximum 10), was calculated as mean between minimum and maximum.

To fit the LSTM model, the number of epochs needs to be chosen. The number of epochs specifies how many times the learning algorithm will iterate over the entire training dataset. As with hidden layers, there is no specific way to choose the number of epochs. The optimal number of epochs depends on various factors, including the complexity of the problem, the amount of data, and the selected hyperparameters. Thus, an experiment with different values and monitoring the training progress could help find the optimal number of epochs for a given problem (Lipton, Berkowitz and Elkan, 2015). However, this process is costly in time and energy. Therefore, prior studies and research can provide guidance on the range of reasonable values for the number of epochs based on similar datasets and problems. These studies can also provide insight into the relationship between the number of epochs and model performance, such as the risk of overfitting and the impact of early stopping. Istiake Sunny, Maswood and Alharbi (2020) compared various LSTM model settings and discovered that 100 epochs and 2 hidden layers produced the highest accuracy for their model. Kim *et al.* (2021) trained their LSTM model with epochs ranging from 1 to 30, and they discovered that for different types of data, LSTM with epochs of 25 and 30 produced the highest accuracy. While there is no definitive answer to the optimal number of epochs for LSTM training, based on prior studies and research that provided valuable guidance, 70 epochs were chosen for the current study.

3.9.5 Data splitting for the experiment

As was discussed in Chapter 2.5.6 the 80/20 split for training and testing is a commonly used practise in machine learning because it provides a good balance between having enough data for training and having enough data for testing (Pham *et al.*, 2020; Das *et al.*, 2011 Rácz, Bajusz and Héberger 2021). The idea is to use 80% of the available data for training the model and the remaining 20% for testing the model's performance.

The 80/20 training and testing split is typically selected randomly from the available data. This means that a random subset of the data is selected to be used as the training set, and the remaining data is used as the test set. The random selection helps to ensure that the training and test sets are representative of the overall data set and that the results are not biased towards a particular subset of the data. For the current study, an 80/20 ratio was chosen based on the previous studies findings, which were described in Chapter 2.5.6, as the most appropriate for the data. That is why, in terms of the number of random tests selected, typically, only one test set is selected here.

4 Chapter Four Influence of Solar activity on global earthquakes

The findings of the study, which show a possible link between solar activity and global earthquakes, are presented in this chapter. These findings are presented first, followed by a discussion of the findings as a whole in Chapter 6.

The findings are divided into two categories. The first section demonstrates the association between solar activity and earthquakes with a Richter magnitude of less than 5.5 on the Richter scale. The second half focuses on solar activity and large earthquakes with a Richter magnitude greater than 5.5. The results are presented in tables that include the RMSE and MSE as well as their normalised values.

4.1 Solar activity and global earthquakes with a Richter magnitude less than 5.5

Table 4.1 through Table 4.6, illustrate the outcomes of each part of the experiment. After each table, the summary shows which algorithm had the highest accuracy (the smallest error) and which method had the poorest result (the largest error). Figure 4.1 through Figure 4.12 show the graphical interpretation of the above tables and the contrasts between actual and predicted earthquake values. The data contains over 8,000 records; creating a line graph with good visual appeal using all of them is not possible. That's why a graph of the moving average (averaging every 100 points) was created.

Table 4.1 Global earthquakes M<5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2779	0.2734	0.2768	0.2733
NRMSE by SD	0.9659	0.9503	0.9623	0.9503
NRMSE by mean	0.5562	0.5472	0.5542	0.5472
Mean absolute error				
MAE	0.2339	0.2271	0.2362	0.2326
NMAE by mean	0.4683	0.4547	0.4729	0.4656

Table 4.1 and Figure 4.1 show the highest accuracy, evaluated by NRMSE, was LSTM, and the KNN algorithm had the lowest accuracy for a two-day delay between solar activity events and earthquake events. In terms of NMAE, the highest accuracy was SVR, and the lowest accuracy was RFR. LSTM and SVR are the first two spots in both NRMSE and NMAE, whereas RFR and KNN are the last two.

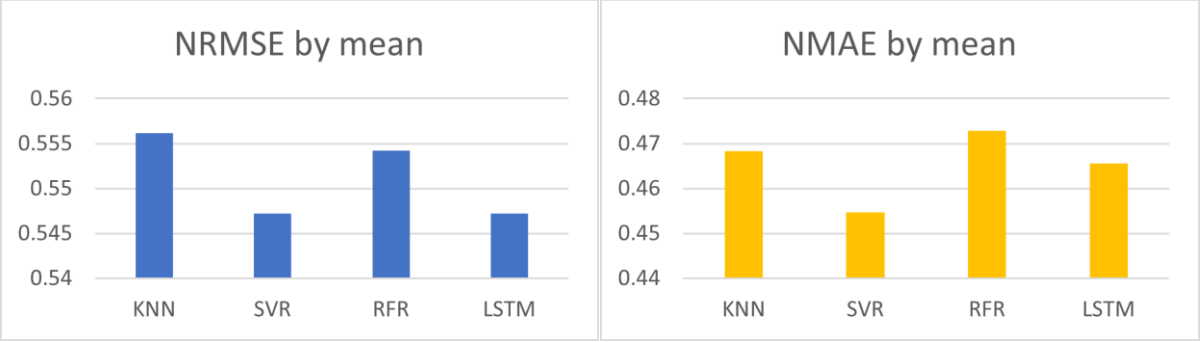


Figure 4.1 Errors: Global earthquakes M<5.5, Two Days Delay

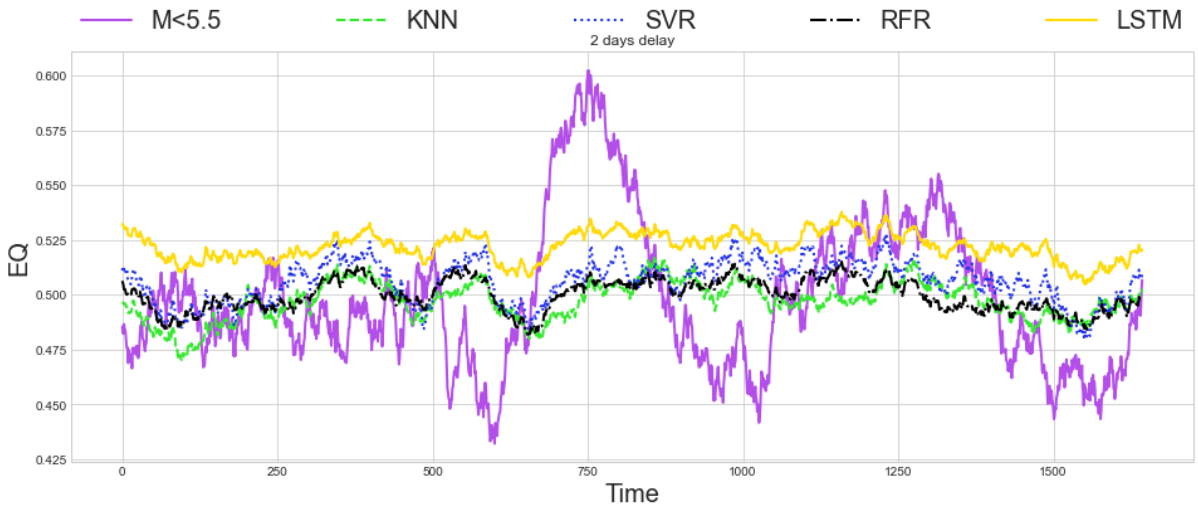


Figure 4.2 Global earthquakes M<5.5: Compare actual and predicted values, Two Days Delay

As can be seen in Figure 4.2, all three traditional machine learning algorithms are close to each other, but SVR repeats the original data graph better. As for LSTM, it repeats the original data similarly to SVR but moves separately and has more points of intersection with the original data, particularly the data points that are above average. This also explains the difference in the outcomes of errors. Also, it was observed that LSTM is more expensive in terms of both time and energy.

Table 4.2 Global earthquakes M<5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2792	0.2758	0.2777	0.2732
NRMSE by SD	0.9682	0.9566	0.9632	0.9476
NRMSE by mean	0.5533	0.5467	0.5505	0.5416
Mean absolute error				
MAE	0.236	0.2296	0.2373	0.2317
NMAE by mean	0.4677	0.455	0.4704	0.4592

Table 4.2 and Figure 4.3 show the results for a three-day delay part are the same as for a two-day delay part. LSTM and SVR have the highest accuracy, whereas RFR and KNN have the lowest. Furthermore, Figure 4.4 shows that the LSTM prediction line and the SVR prediction line are relatively close and that there is a much smaller difference between both errors than there was in the two-day delay part.

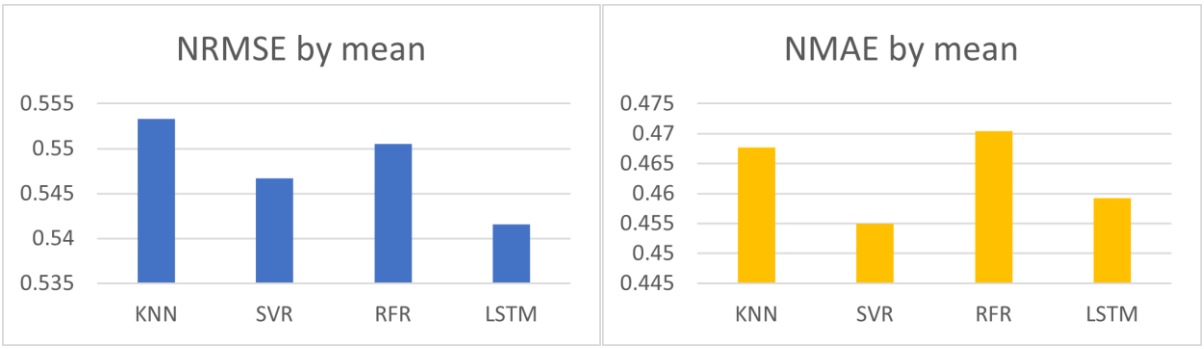


Figure 4.3 Errors: Global earthquakes M<5.5, Three Days Delay

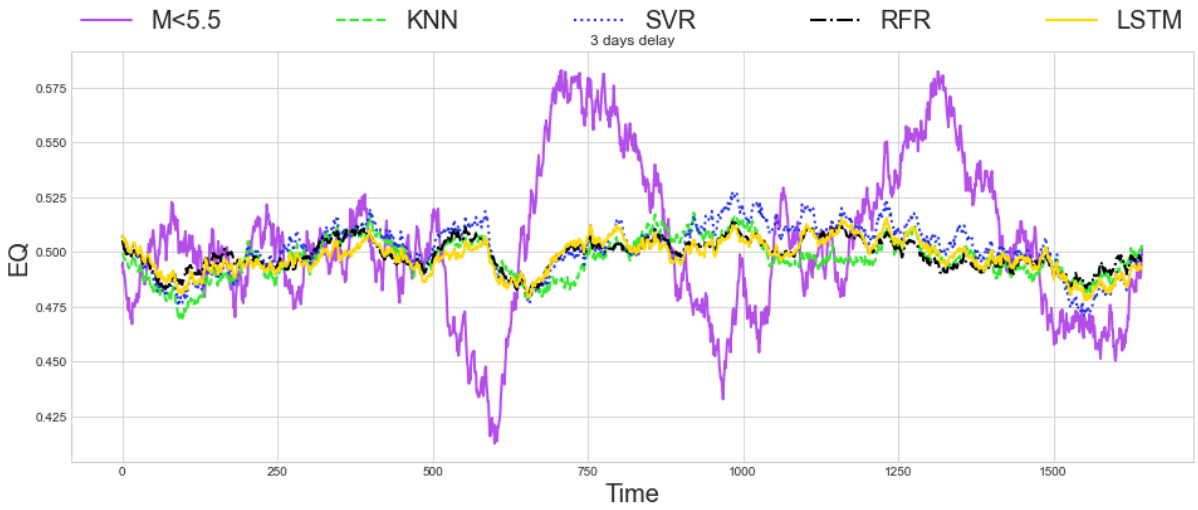


Figure 4.4 Global earthquakes M<5.5: Compare actual and predicted values, Three Days Delay

Table 4.3 Global earthquakes M<5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2805	0.2783	0.28	0.2764
NRMSE by SD	0.97	0.962	0.968	0.9555
NRMSE by mean	0.5563	0.5517	0.5551	0.548
Mean absolute error				
MAE	0.2365	0.2321	0.239	0.236
NMAE by mean	0.4689	0.4601	0.4738	0.4679

The results for a four-day delay part are the same as for a two-day delay and three-day delay parts, as shown in Table 4.3 and Figure 4.5. However, Figure 4.6 shows that the prediction lines repeat the prediction lines in the two-day delay part, and the LSTM prediction values are located above average.

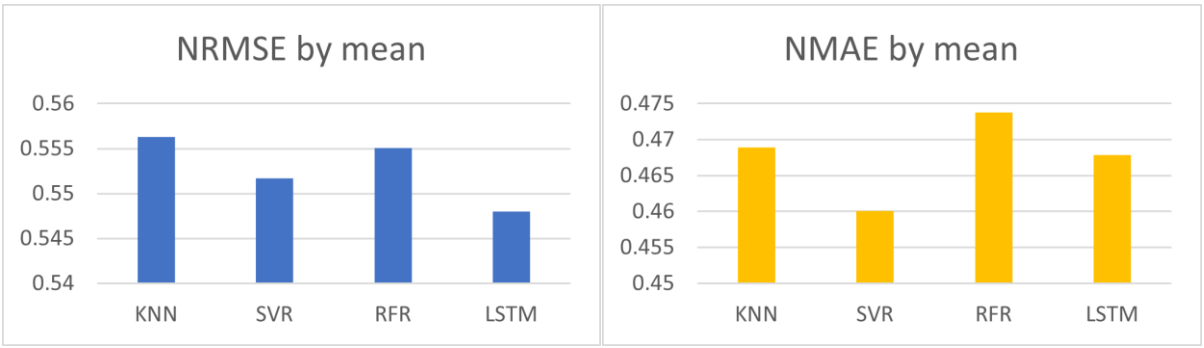


Figure 4.5 Errors: Global earthquakes M<5.5, Four Days Delay

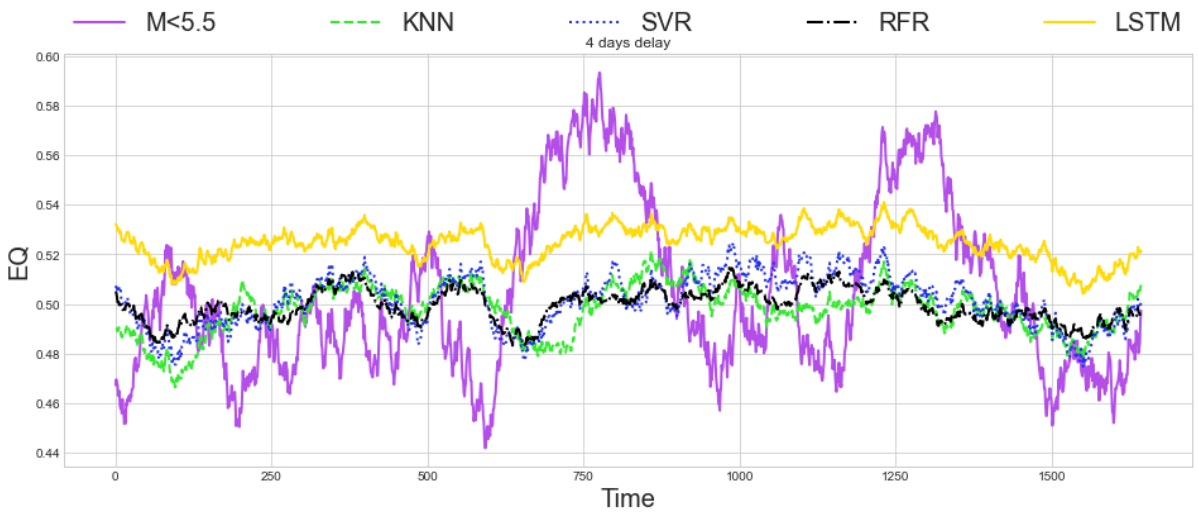


Figure 4.6 Global earthquakes M<5.5: Compare actual and predicted values, Four Days Delay

Based on Table 4.4 and Figure 4.7 the outcomes for a five-day delay component are similar to the previous sections in general. The two with the highest accuracy are LSTM and SVR, while RFR and KNN have the lowest accuracy. Figure 4.8, however, demonstrates that the prediction lines repeat the prediction lines in the three-day delay part.

Table 4.4 Global earthquakes M<5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2797	0.2752	0.2757	0.2729
NRMSE by SD	0.9787	0.9629	0.9646	0.9548
NRMSE by mean	0.5566	0.5476	0.5485	0.5429
Mean absolute error				
MAE	0.2366	0.2305	0.2346	0.231
NMAE by mean	0.4702	0.4586	0.4669	0.4596

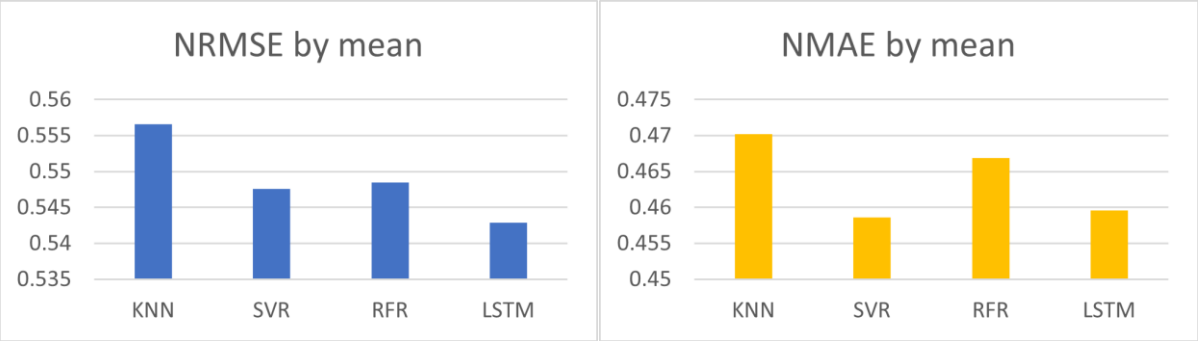


Figure 4.7 Errors: Global earthquakes M< 5.5, Five Days Delay

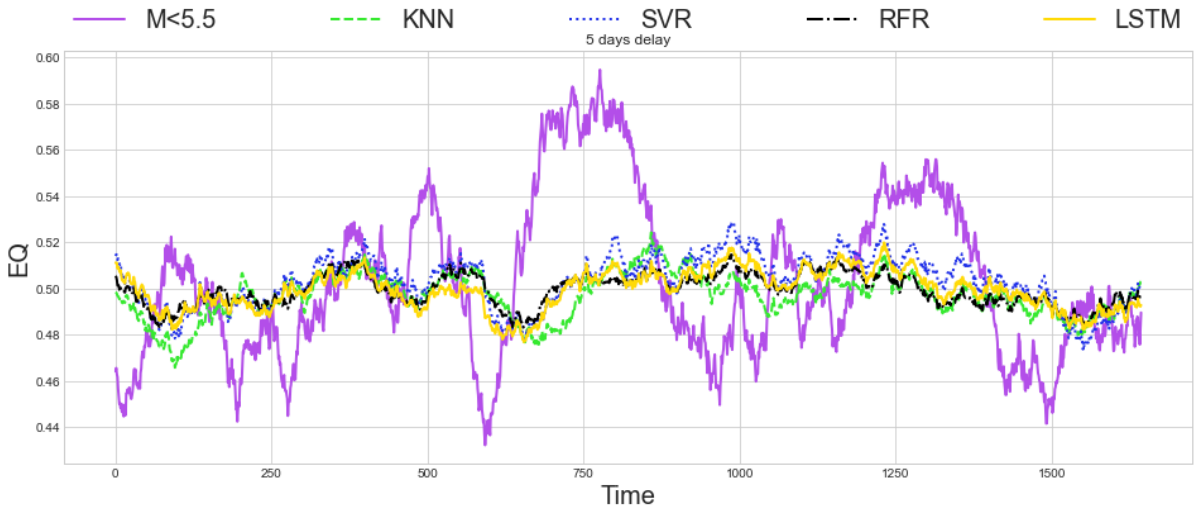


Figure 4.8 Global earthquakes M<5.5: Compare actual and predicted values, Five Days Delay

The results for the six-day delay part are identical to the previous parts, with all models remaining in their unchanged positions, as can be seen in Table 4.5 and Figure 4.9. However, Figure 4.10 depicts the LSTM prediction line in a different location. It has more crossing points with values that are below average and is located below the prediction lines of the conventional ML algorithms.

Table 4.5 Global earthquakes M<5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2746	0.2718	0.2734	0.2713
NRMSE by SD	0.9702	0.9604	0.9659	0.9585
NRMSE by mean	0.5469	0.5414	0.5445	0.5404
Mean absolute error				
MAE	0.2309	0.2255	0.2321	0.2275
NMAE by mean	0.4599	0.4492	0.4622	0.4531

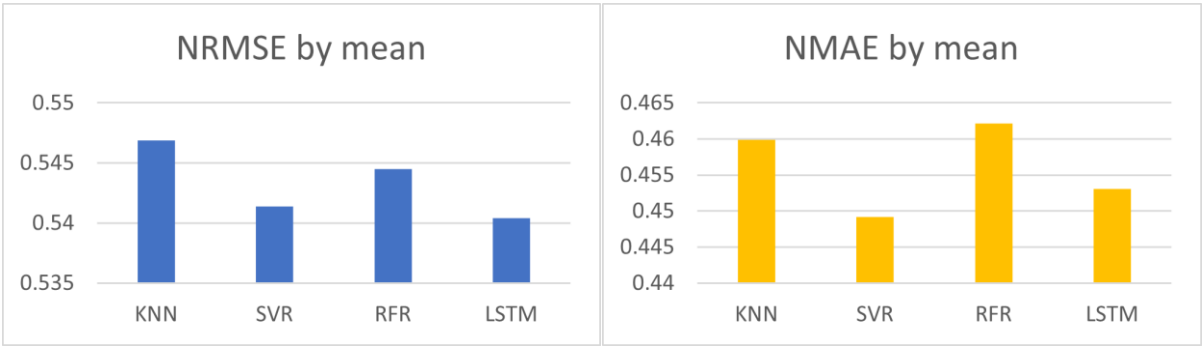


Figure 4.9 Errors: Global earthquakes M<5.5, Six Days Delay

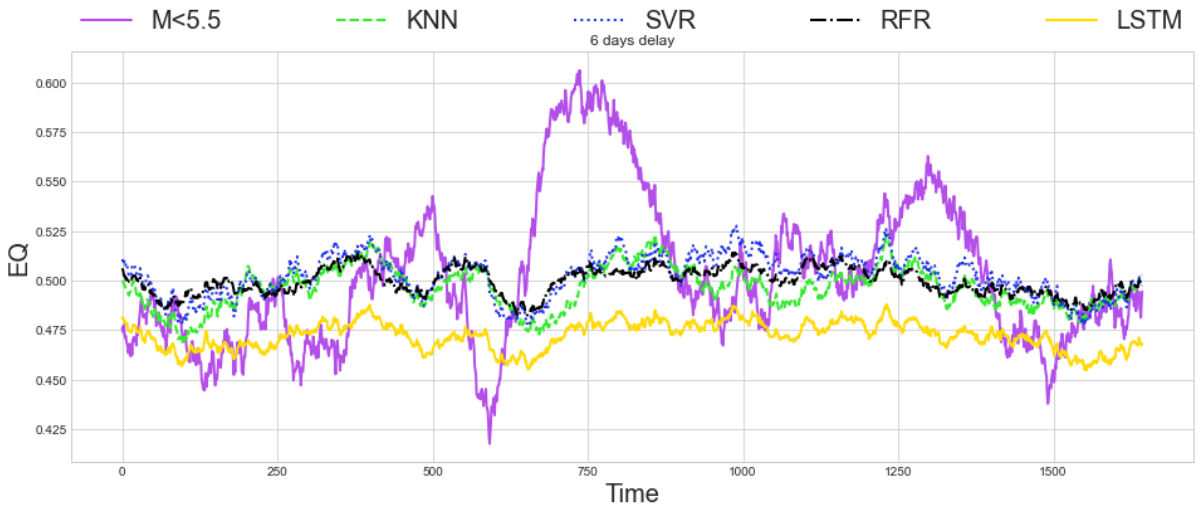


Figure 4.10 Global earthquakes M<5.5: Compare actual and predicted values, Six Days Delay

The results of a seven-day delay for the remaining components of the experiment are the same as the previous parts, Table 4.6 and Figure 4.11. LSTM consistently outperforms other algorithms in terms of NRMSE. In terms of NMAE, SVR and LSTM came in first and second, respectively, with RFR and KNN coming in third and fourth. All of the prediction lines are close

to one another, as seen in Figure 4.12, but LSTM and SVR are better at repeating the original values line, as in the three-day and five-day delay parts.

Table 4.6 Global earthquakes M<5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.278	0.2747	0.2744	0.2717
NRMSE by SD	0.9783	0.9667	0.9656	0.9564
NRMSE by mean	0.5556	0.5491	0.5484	0.5432
Mean absolute error				
MAE	0.2331	0.2281	0.2328	0.2288
NMAE by mean	0.466	0.456	0.4653	0.4574

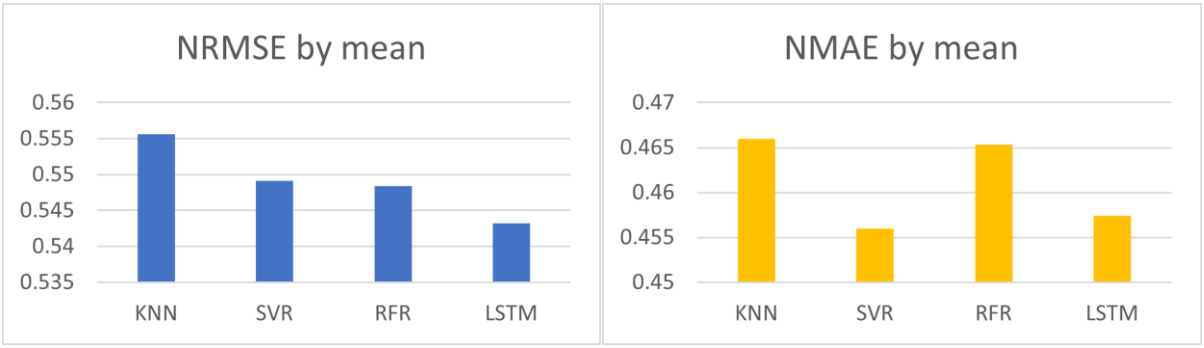


Figure 4.11 Errors: Global earthquakes M<5.5, Seven Days Delay

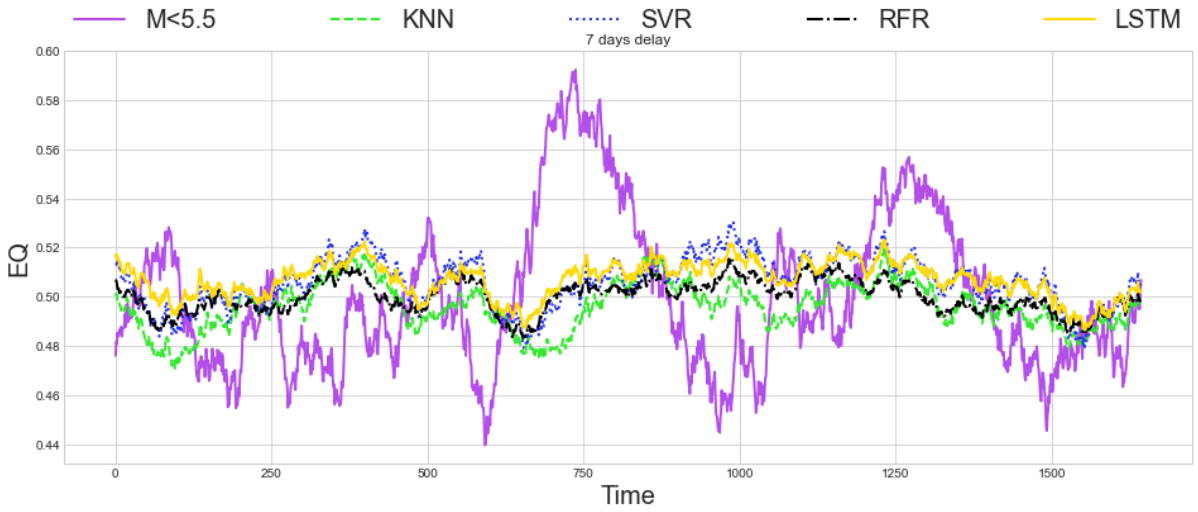


Figure 4.12 Global earthquakes M<5.5: Compare actual and predicted values, Seven Days Delay

The results are visualised in Figure 4.13. It is evident from these graphs that LSTM and SVR have the highest accuracy. Furthermore, the algorithms with the highest and lowest accuracy are generally the same in all parts of the experiment with solar activity and earthquakes with Richter magnitudes less than 5.5. Furthermore, the NRMSE metric values for all algorithms were discovered to be extremely close to one another. The same was seen in the NMAE measurements.

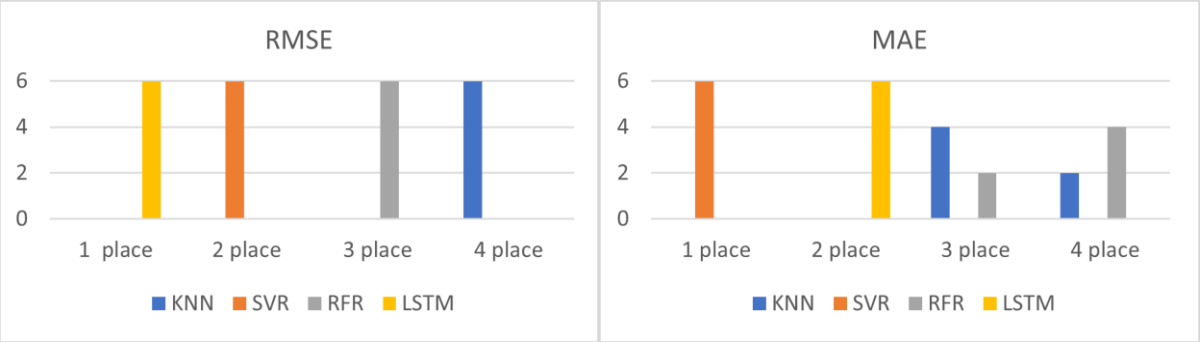


Figure 4.13 Global earthquakes M<5.5, summarising results

It was noted that, based on Figure 4.2 and the other similar graphs, the earthquakes have two peaks, which had been discussed earlier. That is why the values that are further from the mean are significant in the earthquake data. That is why RMSE values are more desirable in this situation. It had been found that, in terms of normalising error values, the three-day delay and six-day delay parts have the highest accuracy in terms of NRMSE, while the six-day delay part has the highest accuracy in terms of NMAE.

4.2 *Solar activity and global earthquakes with a Richter magnitude equal to or greater than 5.5*

The results of each part of the experiment are depicted in Table 4.7 – Table 4.12. From Figure 4.14 until Figure 4.25, the graphical interpretation of the above tables and the differences between actual and predicted values are shown.

Table 4.7 Global earthquakes M ≥ 5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3461	0.3532	0.3386	0.3385
NRMSE by SD	1.0213	1.0422	0.9989	0.9987
NRMSE by mean	0.7737	0.7895	0.7568	0.7566
Mean absolute error				
MAE	0.2897	0.2926	0.2793	0.2812
NMAE by mean	0.6476	0.654	0.6243	0.6286

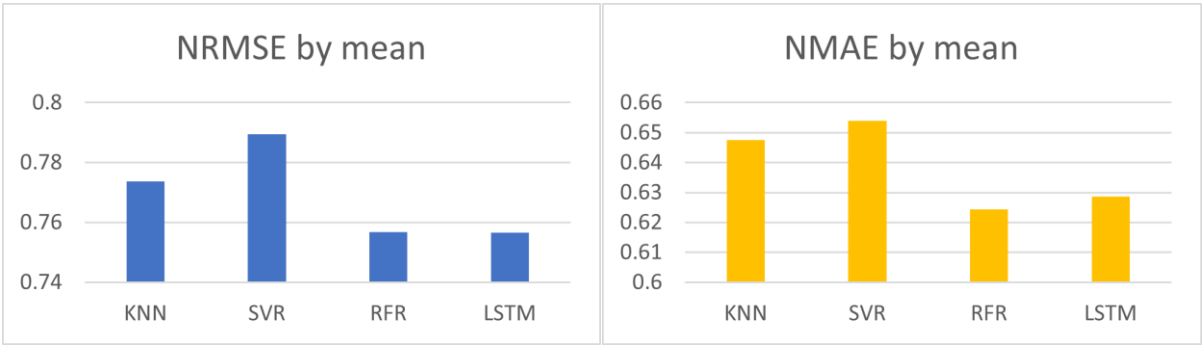


Figure 4.14 Errors: Global earthquakes M≥5.5, Two Days Delay

Table 4.7 and Figure 4.14 show the highest accuracy for a two-day delay in terms of NRMSE was LSTM, with a very close value to RFR. Also, RFR had the highest accuracy in terms of NMAE. The model with the lowest accuracy had SVR in both metrics (NRMSE and NMAE), and KNN came in third. The SVR result is explained by Figure 4.15, which depicts the SVR prediction line as being above the actual values line, while LSTM and RVR are situated close to averages.

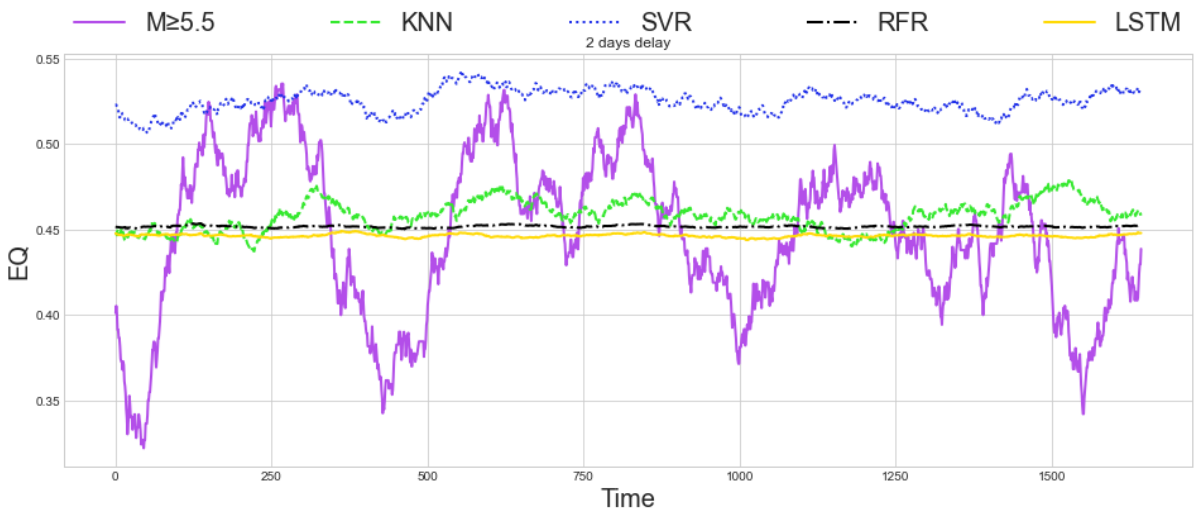


Figure 4.15 Global earthquakes M≥5.5: Compare actual and predicted values, Two Days Delay

Table 4.8 Global earthquakes M≥5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3516	0.3535	0.3399	0.3406
NRMSE by SD	1.0329	1.0383	0.9986	1.0005
NRMSE by mean	0.7739	0.7779	0.7482	0.7496
Mean absolute error				
MAE	0.2961	0.2931	0.2811	0.2864
NMAE by mean	0.6516	0.6452	0.6187	0.6303

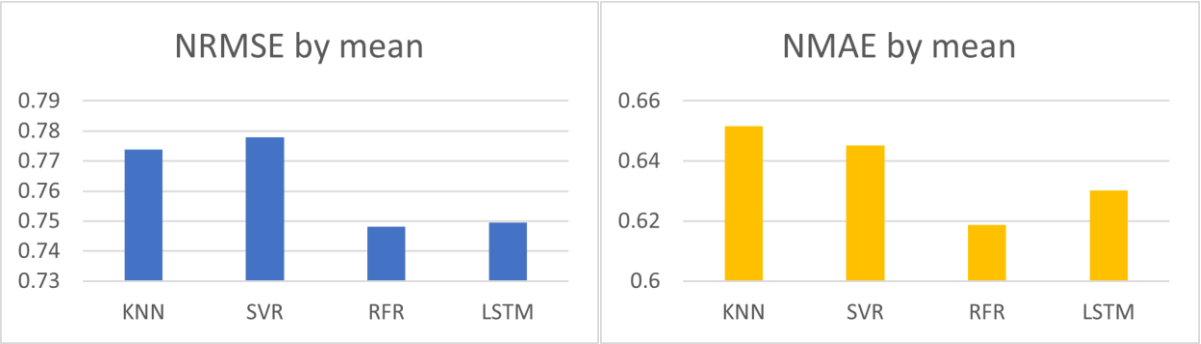


Figure 4.16 Errors: Global earthquakes M≥5.5, Three Days Delay

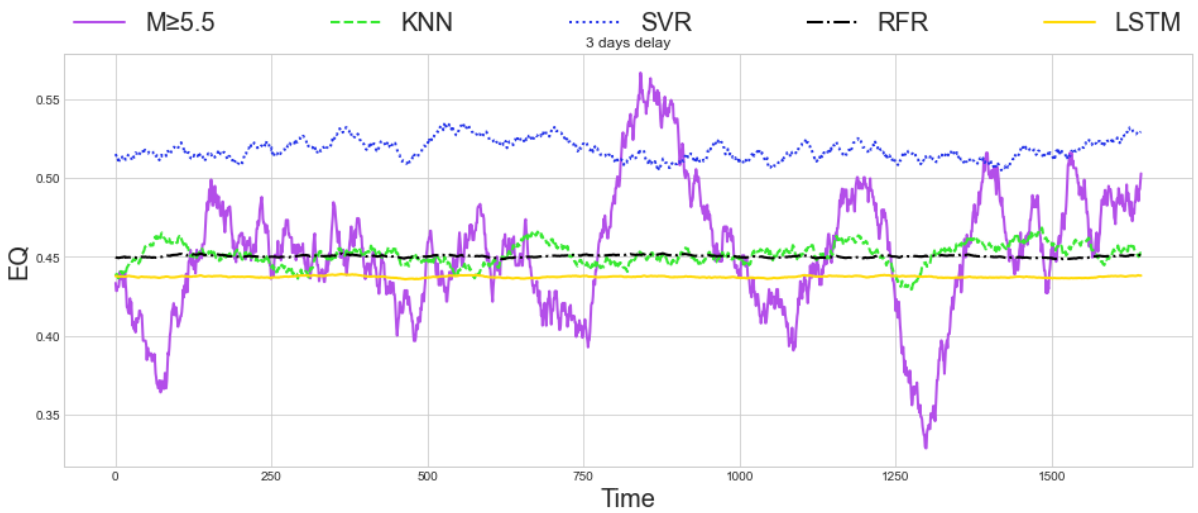


Figure 4.17 Global earthquakes M≥5.5: Compare actual and predicted values, Three days delay

For the three-day delay part, results are shown in Table 4.8 and Figure 4.16. For both metrics, RFR has the highest accuracy, followed by LSTM with very close values. SVR showed the lowest accuracy in terms of NRMSE, while KNN had the lowest accuracy in terms of NMAE. Figure 4.17 shows that the SVR prediction line, as in the previous part, is above the actual values lines and the KNN line does not properly repeat the actual values line. RFR and LSTM get their results because they are located near the averages and have more crossing points with the actual values line.

Table 4.9 Global earthquakes M≥5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3533	0.3577	0.3457	0.3456
NRMSE by SD	1.0215	1.0342	0.9995	0.9993
NRMSE by mean	0.7823	0.792	0.7654	0.7653
Mean absolute error				
MAE	0.3	0.2971	0.289	0.2916
NMAE by mean	0.6644	0.6579	0.64	0.6457

The results for the four-day delay part, from Table 4.9 and Figure 4.18, are the same as for a two-day delay part. LSTM and RFR had the highest accuracy, whereas SVR and KNN had the lowest accuracy. Figure 4.19 shows that SVR prediction lines are above the actual values line and KNN prediction lines do not repeat the actual values line property. Additionally, both SVR and KNN have some points where they cross the actual line. LSTM and RFR are both close to the mean of the actual values.

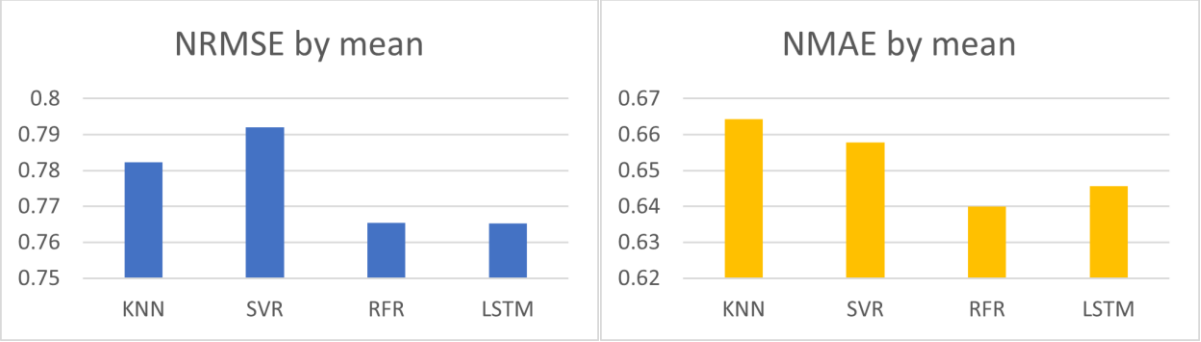


Figure 4.18 Errors: Global earthquakes M≥5.5, Four Days Delay

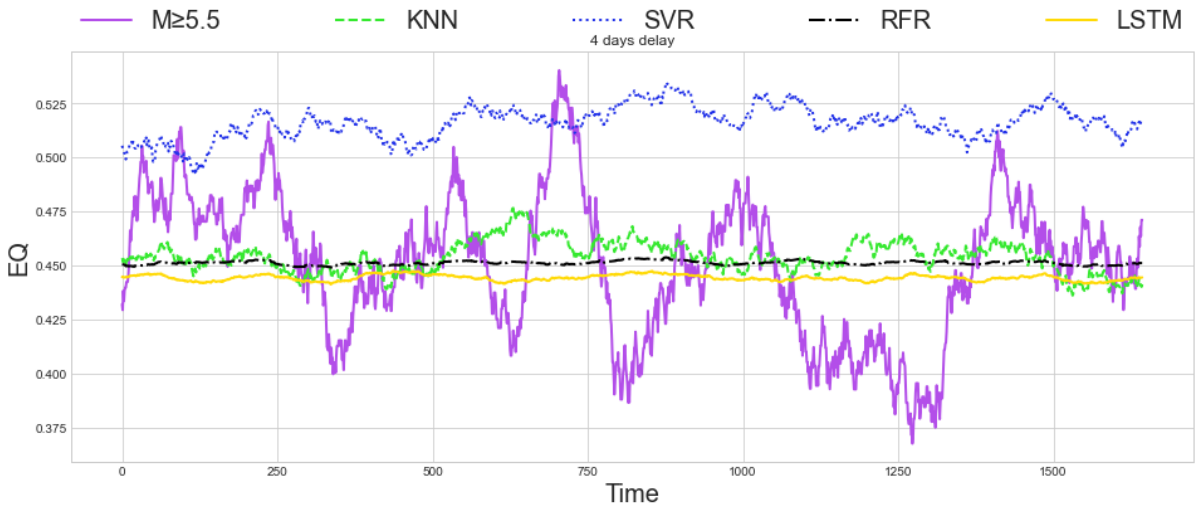


Figure 4.19 Global earthquakes M≥5.5: Compare actual and predicted values, Four Days Delay

For the five-day delay, part RFR had the highest accuracy in terms of NRMSE, while LSTM had the highest accuracy in terms of NMAE (Table 4.10 and Figure 4.20). Both values are close to each other. The situation with the lowest accuracy results is the same as it was before the four-day delay.SVR and KNN are located in the same position as they were in the four-day delay part of Figure 4.21. The LSTM prediction line is located above the RFR prediction line.

Table 4.10 Global earthquakes M≥5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3496	0.3532	0.34	0.3402
NRMSE by SD	1.0299	1.0404	1.0016	1.0021
NRMSE by mean	0.7813	0.7893	0.7598	0.7602
Mean absolute error				
MAE	0.2952	0.2922	0.2816	0.2787
NMAE by mean	0.6598	0.6529	0.6293	0.6227

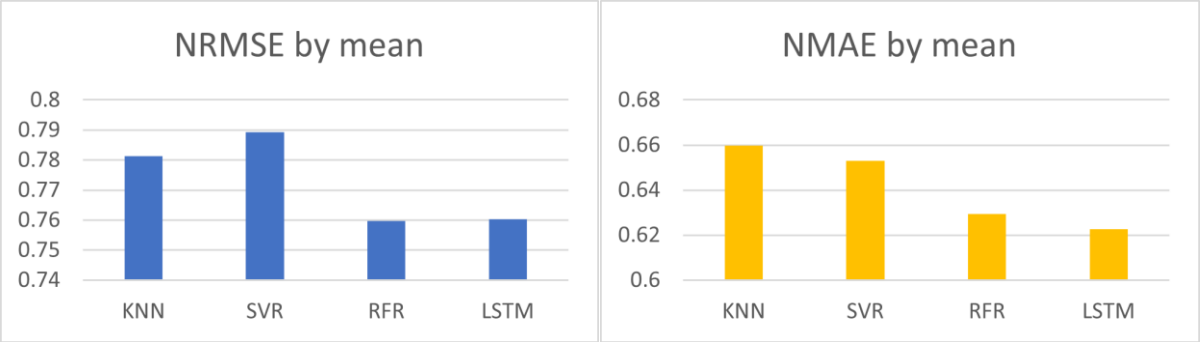


Figure 4.20 Errors: Global earthquakes M≥5.5, Five Days Delay

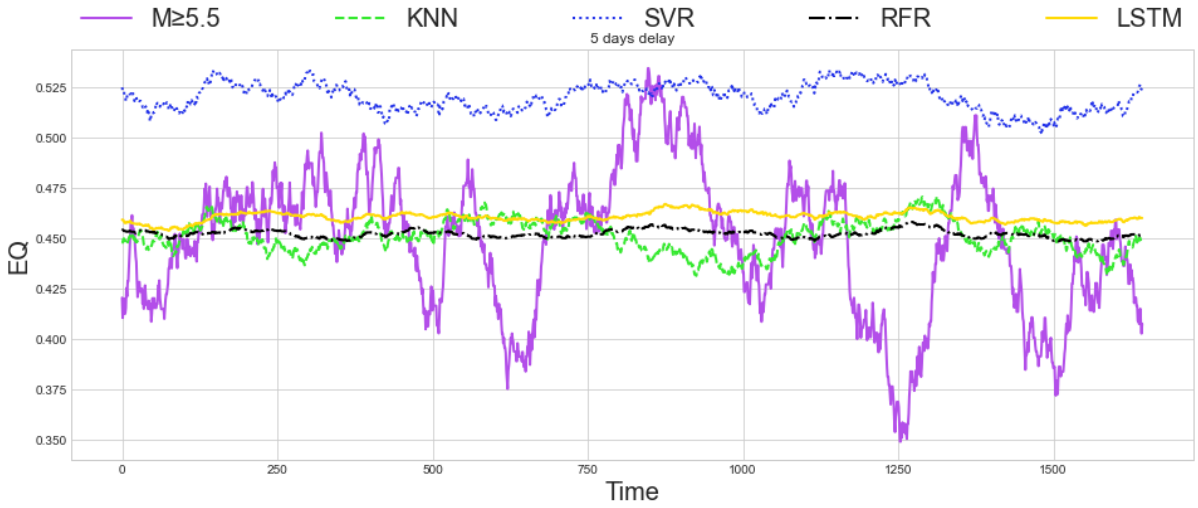


Figure 4.21 Global earthquakes: compare actual and predicted values, Five Days Delay

Table 4.11 and Figure 4.22 show that the results for the six-day delay part are identical to those for the five-day delays. The highest accuracy had LSTM and RFR, while the lowest accuracy had SVR and KNN. In addition, the prediction lines in Figure 4.23 have almost the same location as in the five-day delay part.

Table 4.11 Global earthquakes M≥5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3552	0.3588	0.3421	0.3429
NRMSE by SD	1.0384	1.0489	1.0003	1.0025
NRMSE by mean	0.8021	0.8103	0.7727	0.7744
Mean absolute error				
MAE	0.2998	0.2962	0.2826	0.2777
NMAE by mean	0.6772	0.6689	0.6383	0.6273

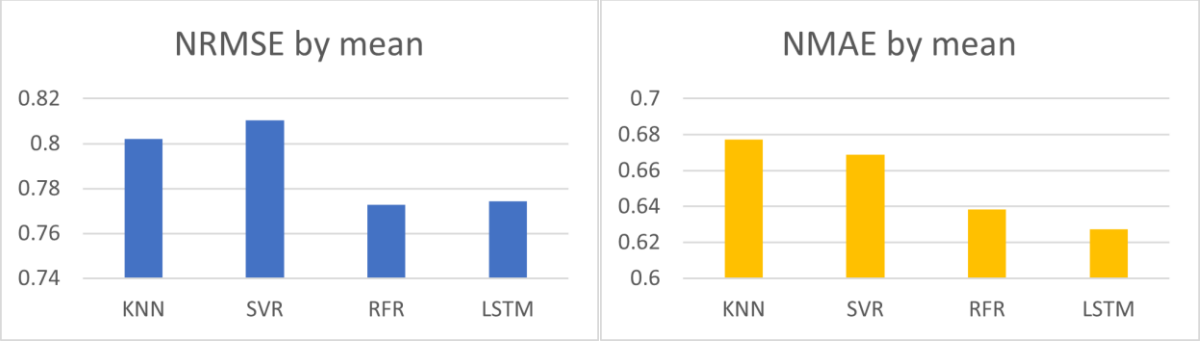


Figure 4.22 Errors: Global earthquakes M≥5.5, Six Days Delay

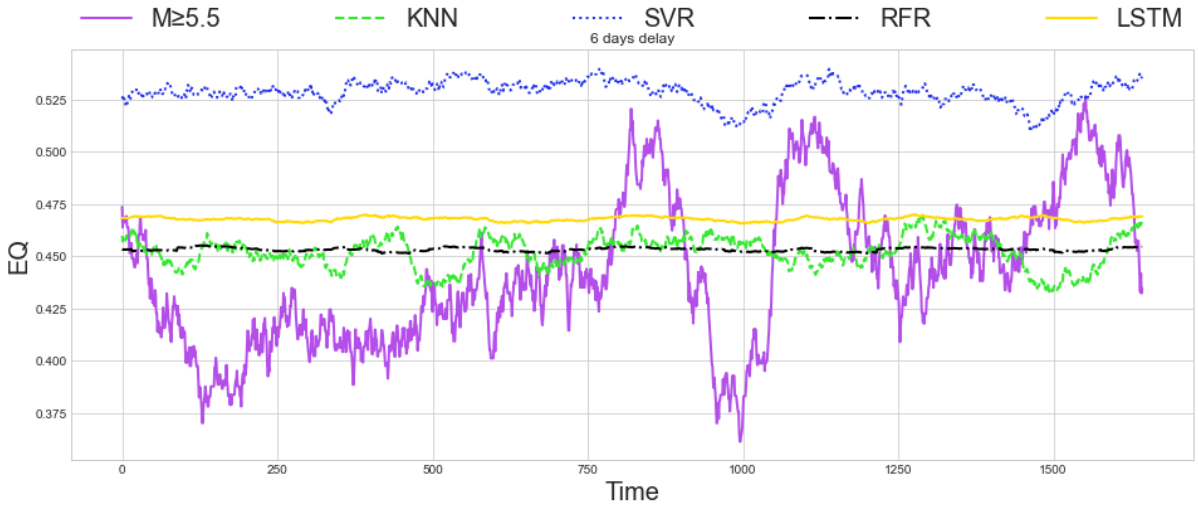


Figure 4.23 Global earthquakes M≥5.5: Compare actual and predicted values, Six Days Delay

Table 4.12 and Figure 4.24 show that for the seven-day delay part, LSTM has the highest accuracy in both metrics, followed by RFR. However, the errors are close to each other. The positions of SVR and KNN are the same as in the six-day delay part. Figure 4.25 demonstrates, that the positions, locations, and behaviorsof all prediction lines are the same as in the six-day delay part.

Table 4.12 Global earthquakes M≥5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3541	0.3616	0.3417	0.3415
NRMSE by SD	1.0375	1.0594	1.0011	1.0005
NRMSE by mean	0.7853	0.8019	0.7578	0.7574
Mean absolute error				
MAE	0.2996	0.2986	0.2832	0.2811
NMAE by mean	0.6645	0.6621	0.6281	0.6234

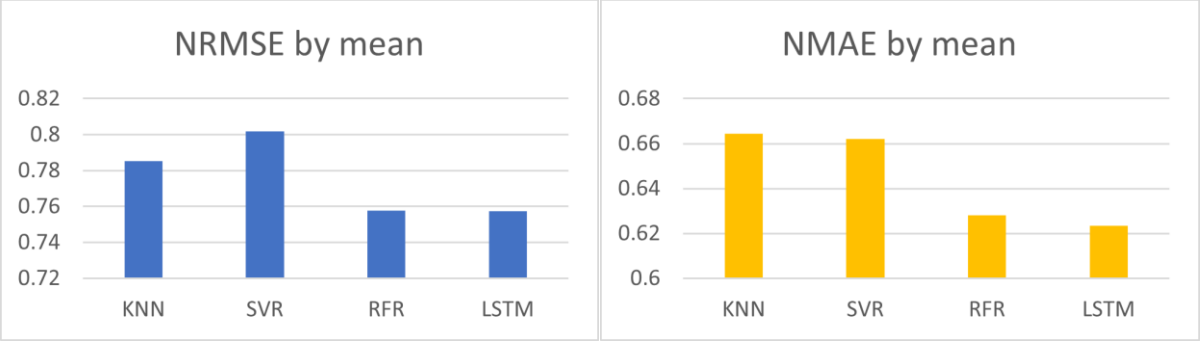


Figure 4.24 Errors: Global earthquakes M≥5.5, Seven Days Delay

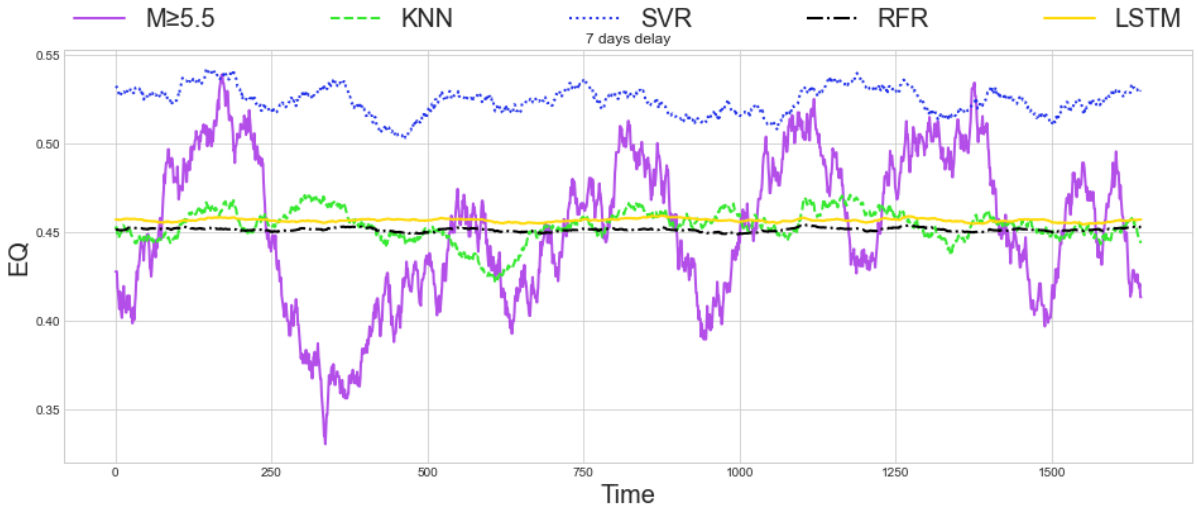


Figure 4.25 Global earthquakes M≥5.5: Compare actual and predicted values, Seven Days Delay

The above error results are quite similar. To find out if there is a difference or if it is a random fluctuation, the ANOVA test was used. The ANOVA test had been chosen because ANOVA is a statistical technique used to determine whether there are significant differences between the means of two or more groups, as in the current study. The null hypothesis for ANOVA is that there is no significant difference between the means of the groups. The alternative hypothesis is that there is a significant difference between the means of the groups. Before conducting the ANOVA test, a check for normality using a Shapiro-Wilk test was done. The null hypothesis for the Shapiro-Wilk test is that the data had a normal distribution (Spiegelhalter, 2019; Agresti, Franklin and Klingenberg, 2018; Mohd Razali and Yap, 2011).

To test if there is a difference between the error results, the RMSE normalised by standard deviation and MAE normalised by mean datasets were used. The level of significance was set at 0.05, which is a commonly used level (Spiegelhalter, 2019; Agresti, Franklin and Klingenberg, 2018). For the calculation of the results of ANOVA and Shapiro-Wilk tests, the statistical Python library was used (*Statistical functions (Scipy. Stats) — SciPy v1.7.1 Manual*,

2021). The results of the ANOVA and Shapiro-Wilk test (Appendix C) showed that there was a difference in the error results for both metrics of solar activity and global earthquake in both parts.

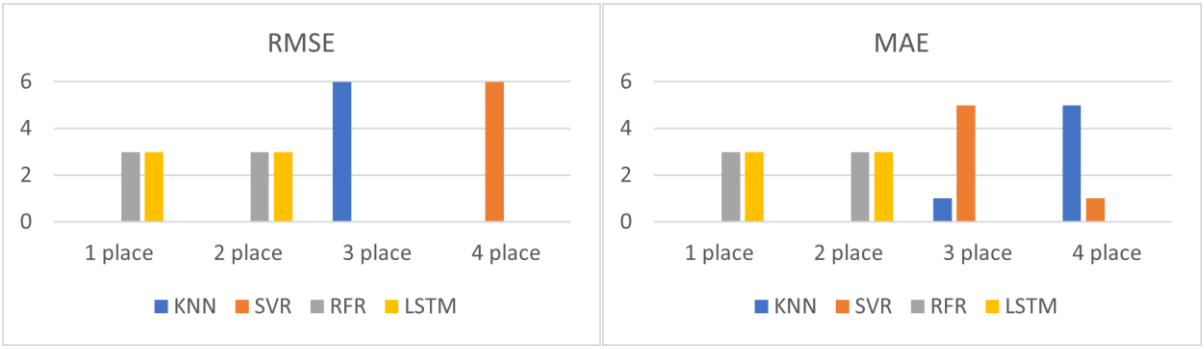


Figure 4.26 Global earthquakes M≥5.5, summarising results

Figure 4.26 shows the visualisation of the results. These graphs clearly show that LSTM and RFR have the highest accuracy. LSTM and RFR equally shared the first and second places in both metrics. Even though it was noted that the error numerical values are not very far from each other, comparing graphs (Figure 4.15 and similar graphs) show the prediction lines have different locations compared to each other and the actual values line. It was found that, in terms of normalising error values, the three-day delay part had the highest accuracy in both metrics.

As in the previous part of the experiment with earthquakes M<5.5, the earthquakes have few peaks with significant values. That is why, RMSE values are also preferable in this case. Additionally, it was observed that while SVR and RVR switched places, LSTM and KNN remained generally in the same positions as in the earlier section of the experiment. What is more, NRMSE and NMAE values in this section are higher than they were in the previous section, and the NRMSE by standard deviation, in some cases, had values greater than "1".

5 Chapter Five Influence of Solar activity on Shallow zone earthquakes, Intermediate zone earthquakes, and Deep zone earthquakes.

In the previous Chapter 4 the findings of a possible link between solar activity and global earthquakes were presented. However, in their study, Novikov *et al.* (2020) conducted an experiment that revealed that earthquakes may be influenced by the electric current generated by solar activity. The depth of the electric current's influence is determined by the Earth's crust. As a result, in the second segment of the study, the dependent variables, earthquakes, were divided by their depth, which was then divided by their magnitude. So, in the second segment of the study, there are three datasets: solar activity and earthquakes from the shallow zone (Chapter 2.2.1), intermediate zone solar activity and earthquakes, and deep zone solar activity and earthquakes.

The findings were divided into two categories, just as in Chapter 4. The first section shows how solar activity is linked to earthquakes with a Richter magnitude less than 5.5. The second section discusses solar activity and large earthquakes with a Richter magnitude of 5.5 or higher. The results are also presented in tables that include the RMSE and MSE as well as their normalised values (as in Chapter 4).

5.1 Solar activity and Shallow zone earthquakes with a Richter magnitude less than 5.5

The results of each section of the experiment with shallow zone earthquakes are presented in Table 5.1 through Table 5.6 The graphical interpretation of the above tables and the differences between actual and predicted earthquake values are presented in Figure 5.1 through Figure 5.12.

Table 5.1 Shallow zone earthquakes M<5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.279	0.2746	0.2769	0.2742
NRMSE by SD	0.97	0.955	0.963	0.9534
NRMSE by mean	0.5589	0.5502	0.5548	0.5493
Mean absolute error				
MAE	0.2358	0.2283	0.2362	0.2321
NMAE by mean	0.4723	0.4573	0.4732	0.4651

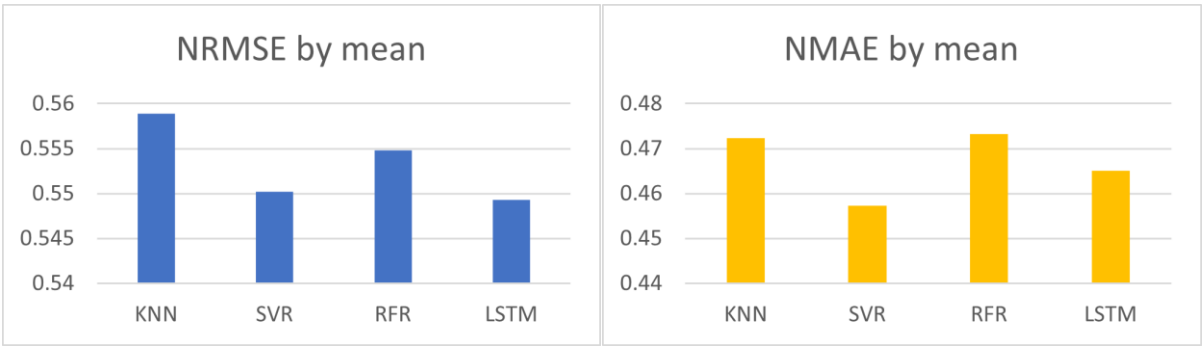


Figure 5.1 Errors: Shallow zone earthquakes M<5.5, Two Days Delay

Table 5.1 and Figure 5.1 show that the two-day delay outcomes are completely in the same range as in the experiment with the two-day delay of the previous dataset, solar activity, and global earthquake M<5.5 (Chapter 4.1). For the NRMSE metric, LSTM had the highest accuracy, and KNN had the lowest accuracy. In terms of NMAE, SVR had the highest accuracy, and RFR had the lowest accuracy. LSTM and SVR are the first two spots in both metrics, whereas RFR and KNN are the last two. Figure 5.2 shows that the actual values line has almost the same slope as in the previous dataset, Chapter 4.1. Moreover, the prediction lines of all algorithms also repeat their position and behaviour.

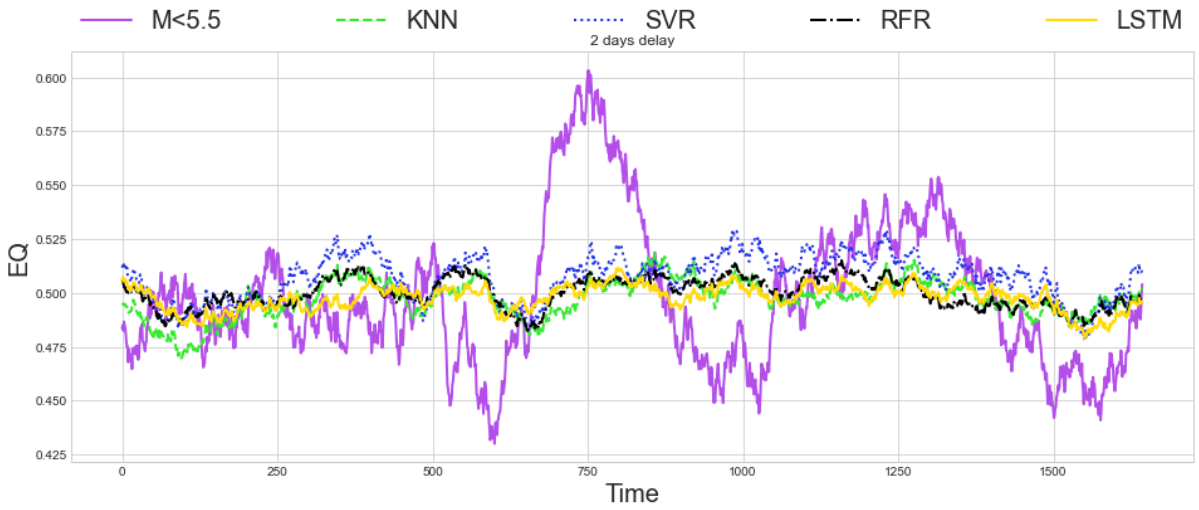


Figure 5.2 Shallow zone earthquakes M<5.5: Compare actual and predicted values, Two Days Delay

In the three-day delay part, the result is the same as in the two-day delay part. LSTM and SVR come in first and second, respectively, with KNN and RFR coming in third and fourth (Table 5.2 and Figure 5.3). The prediction lines of the algorithms, as shown in Figure 5.4, also repeat the location as in the two-day delay part.

Table 5.2 Shallow zone earthquakes M<5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2801	0.2761	0.2777	0.2735
NRMSE by SD	0.9724	0.9588	0.9643	0.9495
NRMSE by mean	0.5562	0.5484	0.5516	0.5431
Mean absolute error				
MAE	0.2371	0.2304	0.2377	0.2319
NMAE by mean	0.4708	0.4575	0.472	0.4606

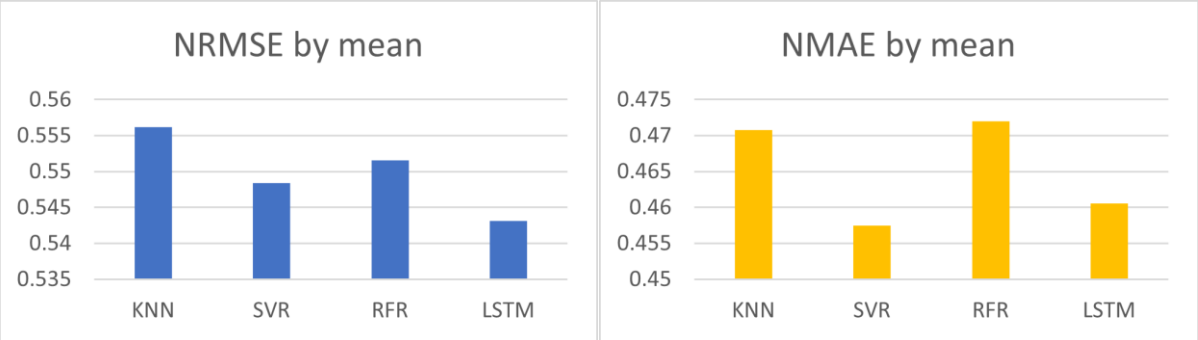


Figure 5.3 Errors: Shallow zone earthquakes M<5.5, Three Days Delay

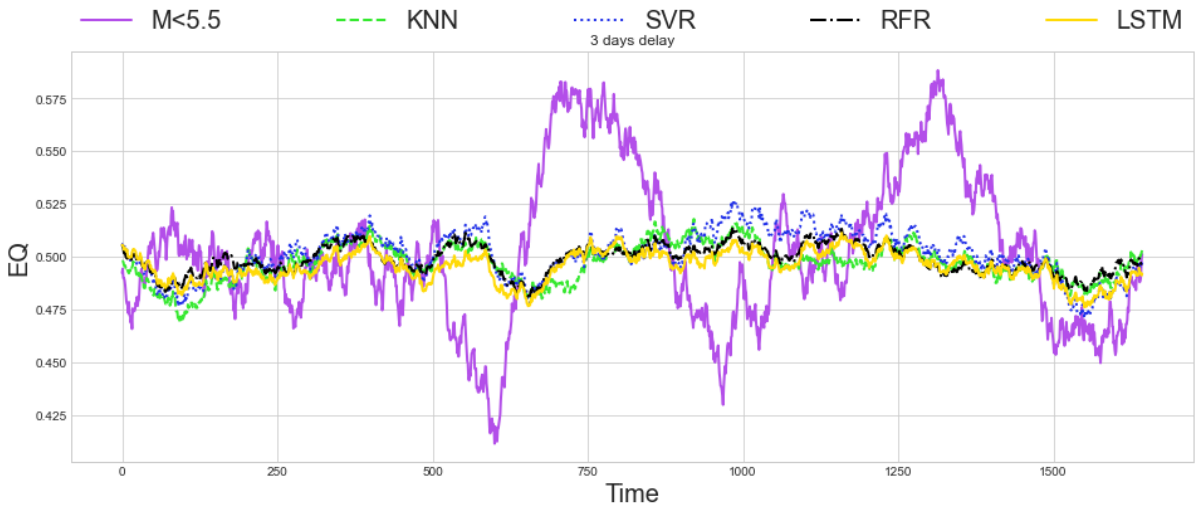


Figure 5.4 Shallow zone earthquakes M<5.5: Compare actual and predicted values, Three Days Delay

Table 5.3 and Figure 5.5 show that, as in the two previous parts, LSTM and SVR had the highest accuracy, while KNN and RFR had the lowest accuracy. The location of the algorithms' prediction lines in Figure 5.6 further supports this.

Table 5.3 Shallow zone earthquakes M<5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2818	0.2795	0.2806	0.2759
NRMSE by SD	0.9736	0.9658	0.9697	0.9533
NRMSE by mean	0.5592	0.5547	0.557	0.5476
Mean absolute error				
MAE	0.2381	0.2337	0.2401	0.2347
NMAE by mean	0.4725	0.4638	0.4765	0.4659

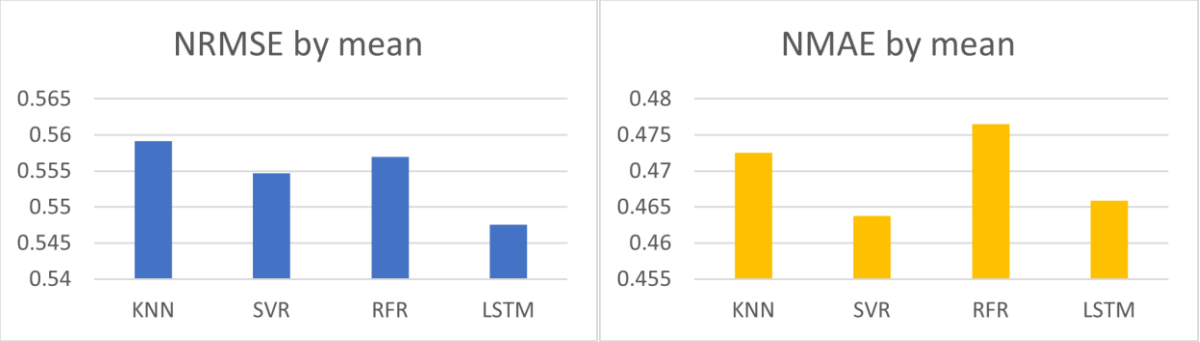


Figure 5.5 Errors: Shallow zone earthquakes M<5.5, Four Days Delay

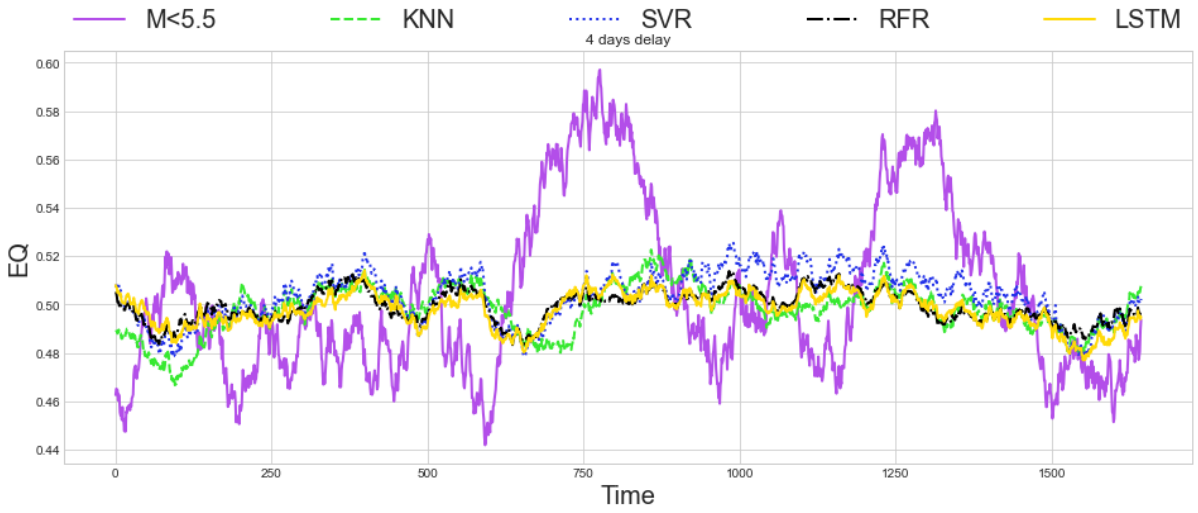


Figure 5.6 Shallow zone earthquakes M<5.5: compare actual and predicted values, Four Days Delay

Table 5.4 and Figure 5.7 demonstrate that, for the five-day delay part, in both metrics, LST showed the highest accuracy and KNN showed the lowest accuracy. As in previous sections, LSTM and SVR had the first two places, and RFR and KNN had the last two places. The prediction lines of the algorithms, Figure 5.8, have the same location as in the previous parts.

Table 5.4 Shallow zone earthquakes M<5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2812	0.2773	0.2766	0.2755
NRMSE by SD	0.9834	0.9695	0.967	0.9635
NRMSE by mean	0.56	0.552	0.5506	0.5486
Mean absolute error				
MAE	0.2384	0.2322	0.2356	0.231
NMAE by mean	0.4746	0.4624	0.4692	0.4599

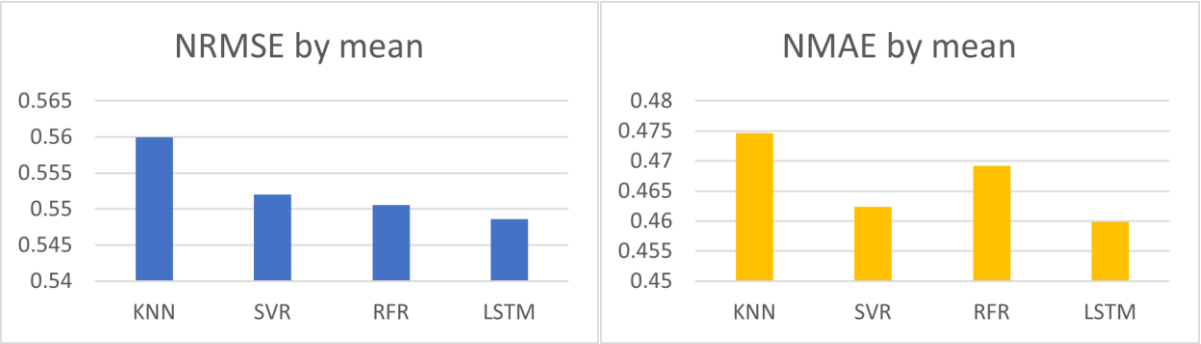


Figure 5.7 Errors: Shallow zone earthquakes M<5.5, Five Days Delay

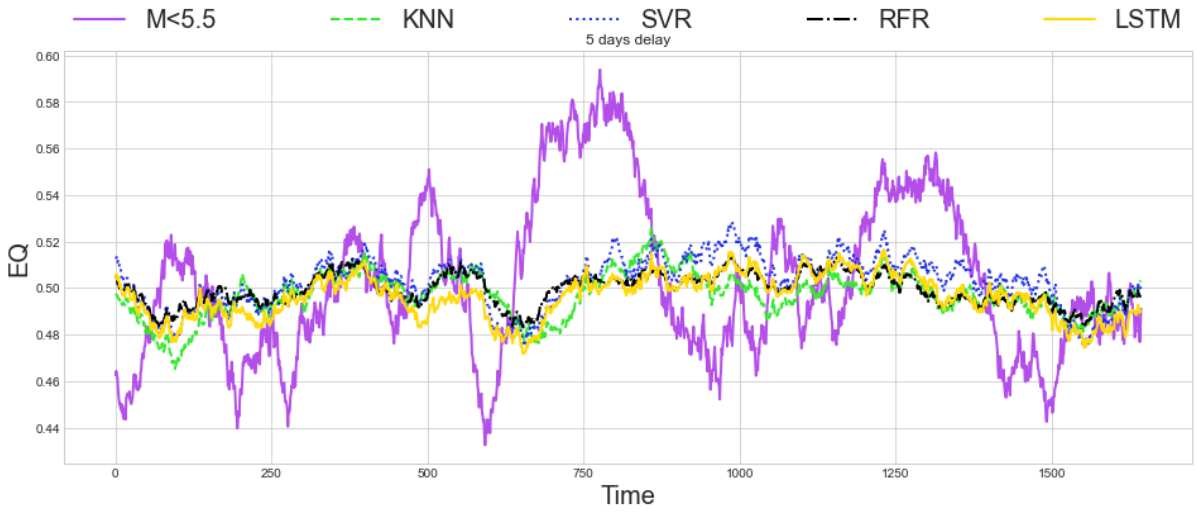


Figure 5.8 Shallow zone earthquakes M<5.5: compare actual and predicted values, Five Days Delay

In the six-day delay part, the outcomes showed the same result as in the first three parts for both metrics (Table 5.5 and Figure 5.9). In general, LSTM and SVR, as well as KNN and RFR, held the same positions in the preceding portions. However, Figure 5.10 demonstrates that the LSTM prediction line is located above the traditional ML algorithms' prediction lines.

Table 5.5 Shallow zone earthquakes M<5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2763	0.2731	0.2743	0.2713
NRMSE by SD	0.9748	0.9637	0.9678	0.9572
NRMSE by mean	0.5509	0.5446	0.5469	0.541
Mean absolute error				
MAE	0.2333	0.2278	0.2337	0.2296
NMAE by mean	0.4652	0.4542	0.4661	0.4579

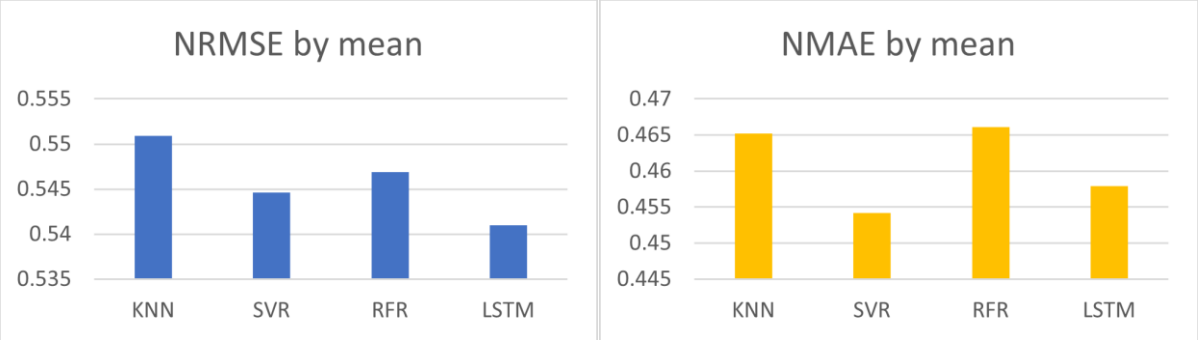


Figure 5.9 Errors: Shallow zone earthquakes M<5.5, Six Days Delay

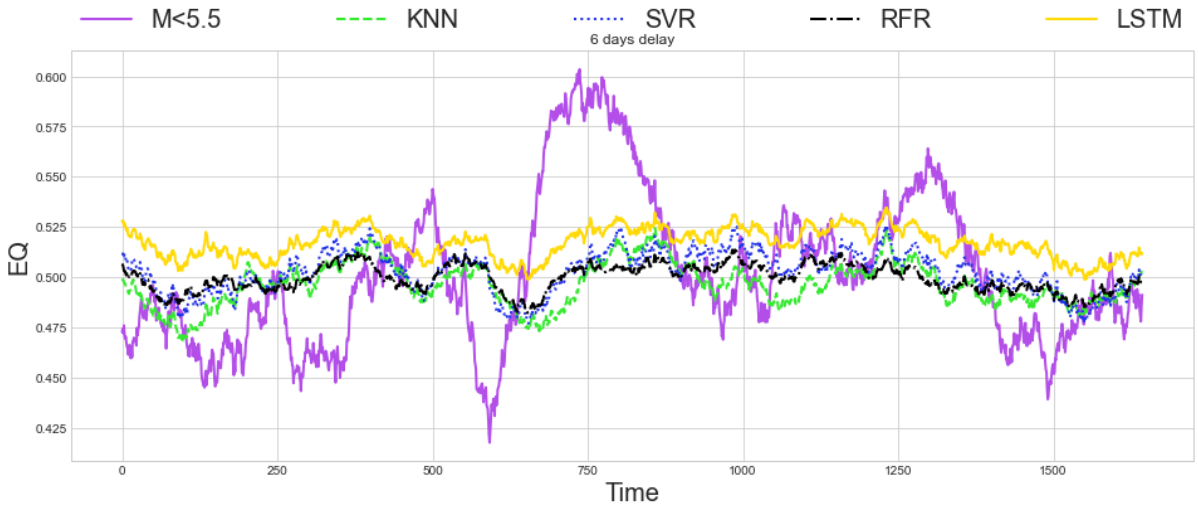


Figure 5.10 Shallow zone earthquakes M<5.5: Compare actual and predicted values, Six Days Delay

Table 5.6 and Figure 5.11 show that the algorithms' positions are almost the same as in the previous part, LSTM and SVR have the first positions in NRMSE and NMAE, respectively, while KNN has the last position in both metrics. Compared to other parts of the experiment, the LSTM prediction line is located below the traditional ML algorithms' prediction lines (Figure 5.12).

Table 5.6 Shallow zone earthquakes M<5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2798	0.275	0.2752	0.2732
NRMSE by SD	0.9844	0.9676	0.9683	0.961
NRMSE by mean	0.5597	0.5501	0.5505	0.5464
Mean absolute error				
MAE	0.2353	0.2294	0.2342	0.2295
NMAE by mean	0.4707	0.4587	0.4685	0.459

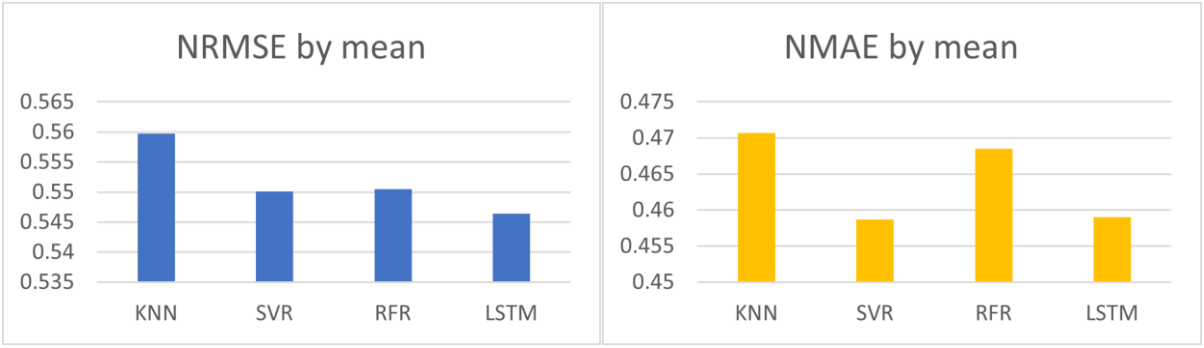


Figure 5.11 Errors: Shallow zone earthquakes M<5.5, Seven Days Delay

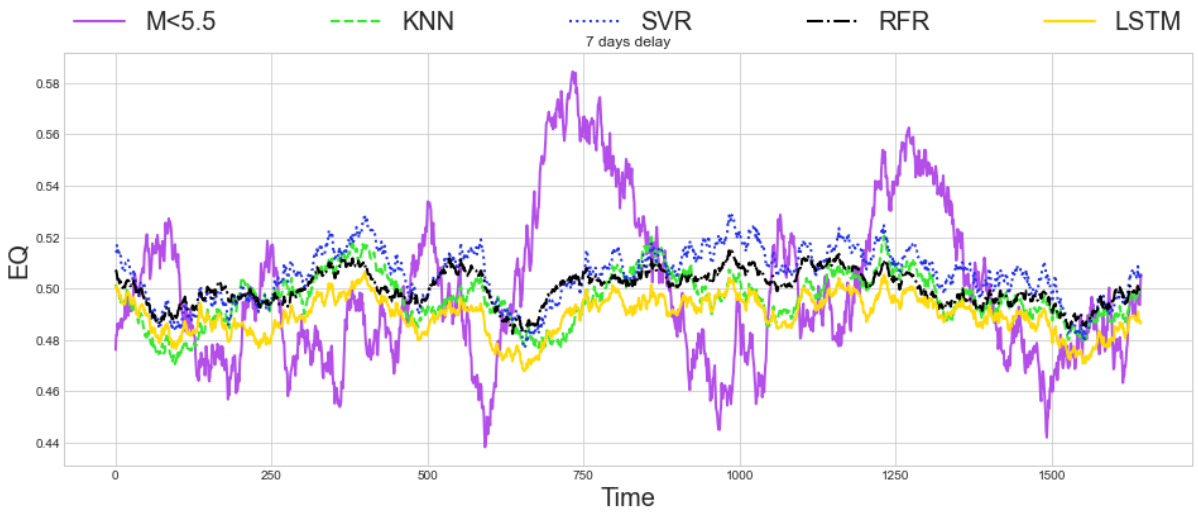


Figure 5.12 Shallow zone earthquakes M<5.5: Compare actual and predicted values, Seven Days Delay

The summary and visualisation of the results are presented in Figure 5.13. The graphs clearly show that LSTM and SVR have the highest performances. Also, it was noted, the highest and lowest accuracy of the algorithms are the same in all areas of the experiment. In both measures, LSTM and SVR provided the highest accuracy, while KNN and RFR had the lowest accuracy. As well as with the previous segments of the experiment, the earthquakes have peaks. Because of this, the values in the earthquake data that are further from the mean are important. As a result, as in the previous case, RMSE values are preferable in this case.

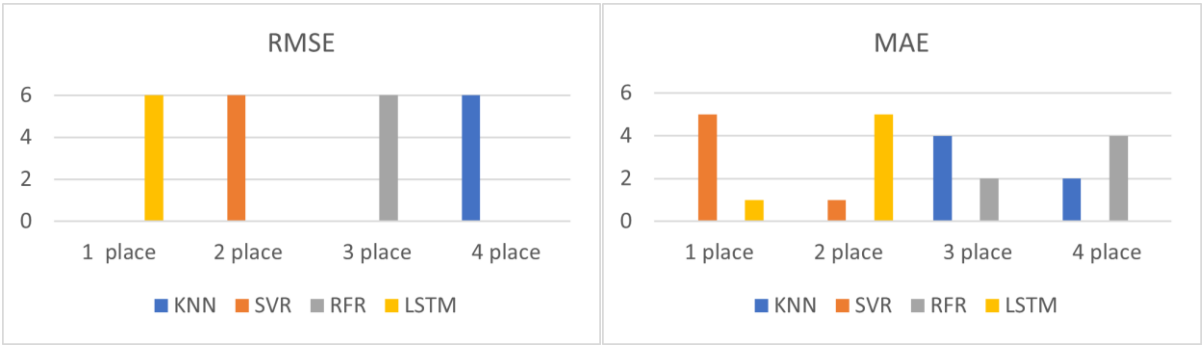


Figure 5.13 Shallow zone earthquakes M<5.5, summarising results

It was discovered that the three-day delay and six-day delay parts have the highest accuracy in terms of NRMSE, while the six-day delay part has the highest accuracy in terms of NMAE when it comes to normalising error values. It was also noted that the range of the results in the dataset for shallow earthquakes M<5.5 is completely the same as for global earthquakes M<5.5.

5.2 *Solar activity and Shallow zone earthquakes with a Richter magnitude equal to or greater than 5.5*

Table 5.7 through Table 5.12 show the outcomes of each component of the experiment. The graphical interpretation of the above tables and the disparities between actual and predicted values are shown in Figure 5.14 through Figure 5.25.

Table 5.7 Shallow zone earthquakes M≥5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3727	0.3688	0.3638	0.3637
NRMSE by SD	1.0233	1.0128	0.9991	0.9988
NRMSE by mean	0.8954	0.8862	0.8743	0.874
Mean absolute error				
MAE	0.3349	0.3267	0.3357	0.3356
NMAE by mean	0.8047	0.7851	0.8067	0.8064

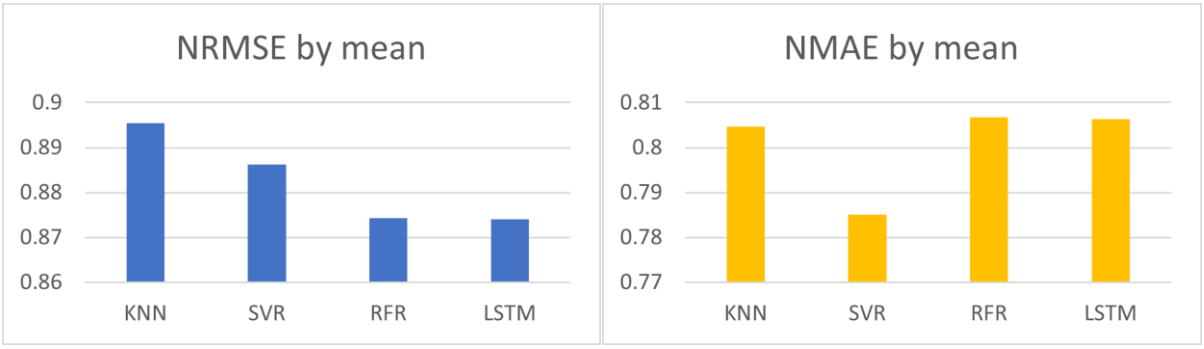


Figure 5.14 Errors: Shallow zone earthquakes M≥5.5, Two Days Delay

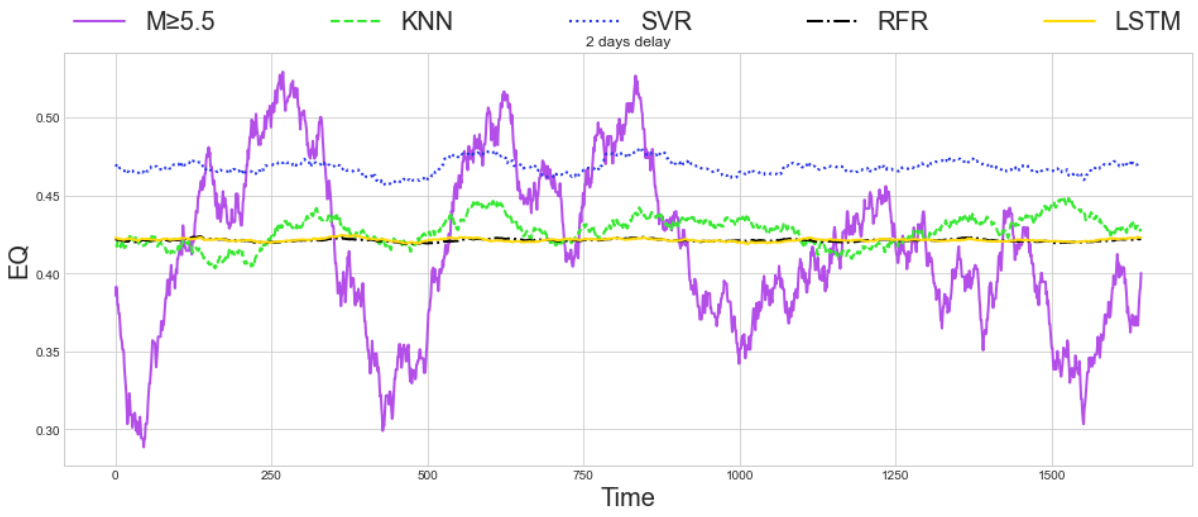


Figure 5.15 Shallow zone earthquakes M≥5.5: Compare actual and predicted values, Two Days Delay

Table 5.7 and Figure 5.14 show that, in terms of NRMSE, RFR, and LSTM, they had the highest accuracy, respectively. KNN had the lowest accuracy. However, in terms of NMAE, there is an opposite result: SVR and KNN had the highest accuracy, while LSTM and RFR had the lowest accuracy. As can be seen in Figure 5.15, the LSTM and RFR prediction lines are very close to each other and to the averages of the actual values. But the SVR prediction line is close to the upper limit of actual values. The KNN prediction line is also located near the average actual values, but it does not repeat the actual values line properly.

Table 5.8 Shallow zone earthquakes M≥5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3765	0.372	0.3649	0.365
NRMSE by SD	1.0305	1.0182	0.9986	0.999
NRMSE by mean	0.9043	0.8936	0.8764	0.8768
Mean absolute error				
MAE	0.3391	0.3293	0.3368	0.3371
NMAE by mean	0.8144	0.7909	0.8089	0.8089

The three-day delay is displayed in Table 5.8 and Figure 5.16. RFR and LSTM had the highest accuracy in terms of NRMSE, while KNN had the lowest accuracy. SVR and KNN had the highest accuracy in terms of NMAE. All of the algorithms' prediction lines are located almost exactly where they were in the two-delay part, Figure 5.17. Additionally, the RFR and LSTM values of NRMSE by standard deviation are extremely close to each other and to "1". The values of NRMSE by standard deviation for SVR and KNN are greater than "1". Similar circumstances exist for the two-day delay part.

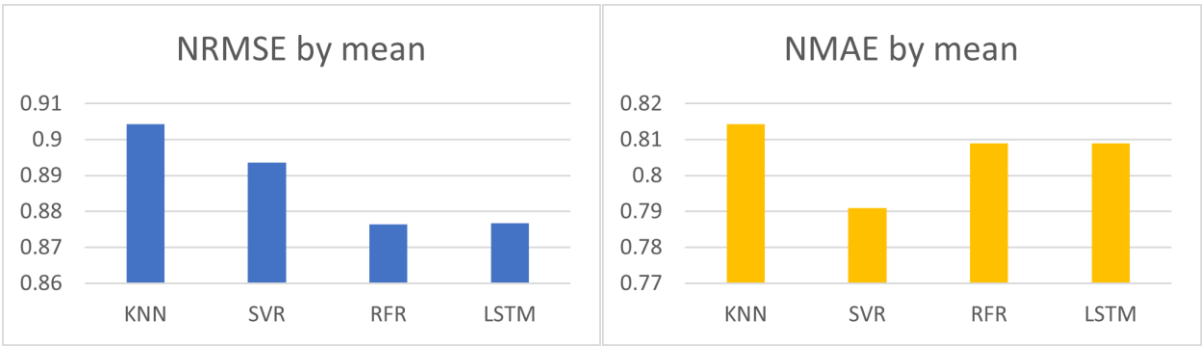


Figure 5.16 Errors: Shallow zone earthquakes M≥5.5, Three Days Delay

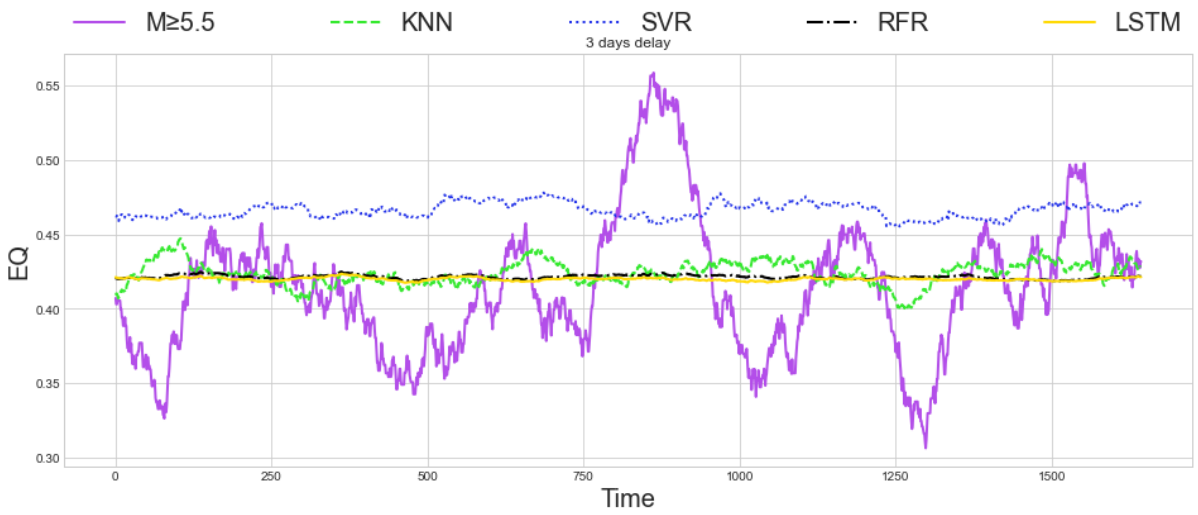


Figure 5.17 Shallow zone earthquakes M≥5.5: compare actual and predicted values, Three Days Delay

According to Table 5.9 and Figure 5.18, the four-day delay had the same results as the three-day delay in terms of NRMSE. The highest levels of accuracy were held by RFR and LSTM, with almost identical error values, followed by SVR and KNN. SVR had the highest accuracy in terms of NMAE, and LSTM had the lowest accuracy. It should be noted that the NRMSE by standard deviation values for SVR and KNN are greater than "1". Figure 5.19 shows that the prediction lines for traditional ML have almost the same location compared with the three-day delay part. The LSTM prediction line is located below its three-day delay position.

Table 5.9 Shallow zone earthquakes M≥5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.374	0.371	0.3646	0.3647
NRMSE by SD	1.0251	1.0169	0.9992	0.9996
NRMSE by mean	0.8974	0.8902	0.8748	0.8751
Mean absolute error				
MAE	0.3376	0.3289	0.3363	0.338
NMAE by mean	0.81	0.7893	0.8069	0.811

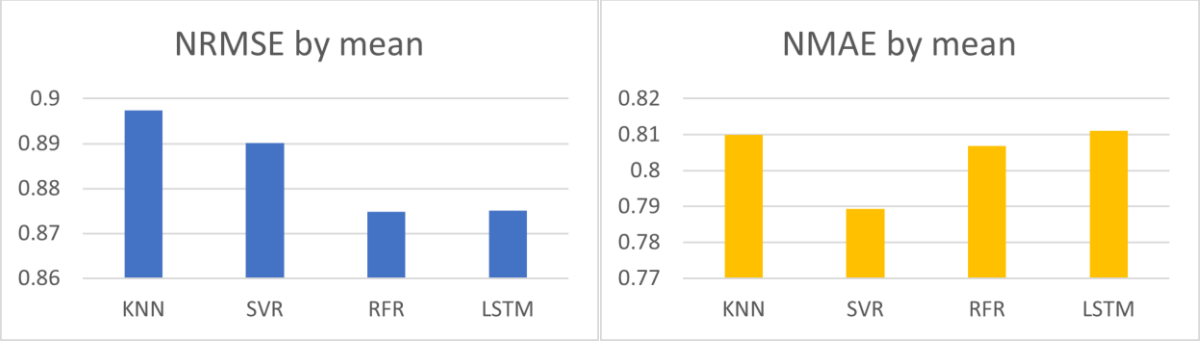


Figure 5.18 Errors: Shallow zone earthquakes M≥5.5, Four Days Delay

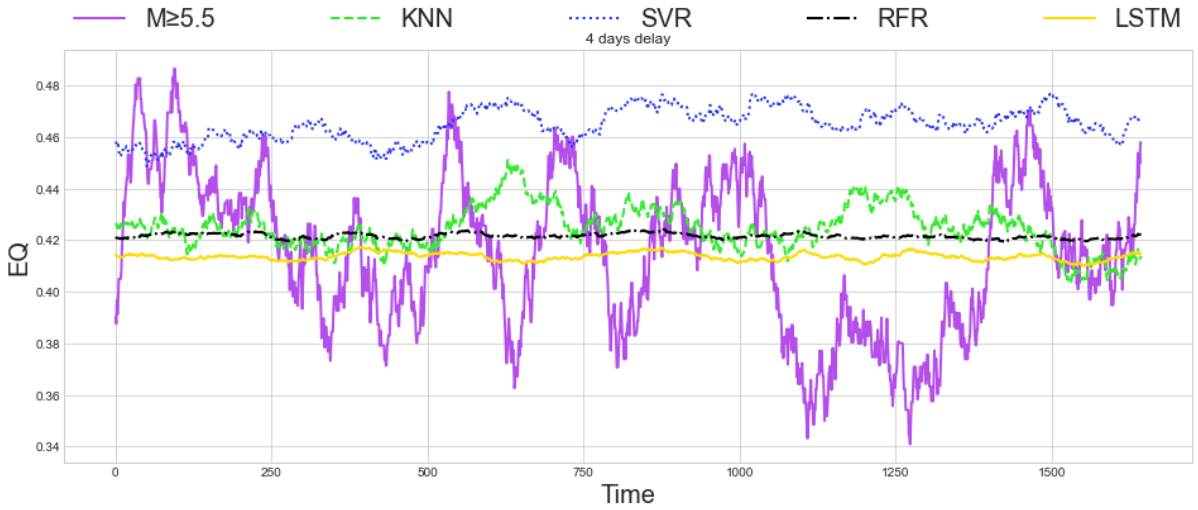


Figure 5.19 Shallow zone earthquakes M≥5.5: compare actual and predicted values, Four Days Delay

RFR and LSTM had the highest accuracy with a five-day delay, while SVR and KNN had the lowest accuracy (Table 5.10 and Figure 5.20). SVR and LSTM had the highest accuracy in terms of NMAE, while KNN and RFR had the lowest accuracy. Figure 5.21 illustrates that the traditional ML algorithms' prediction lines remained in place while the LSTM prediction line moved.

Table 5.10 Shallow zone earthquakes M≥5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3693	0.3667	0.3593	0.3612
NRMSE by SD	1.0282	1.0208	1.0003	1.0057
NRMSE by mean	0.8722	0.8659	0.8485	0.8531
Mean absolute error				
MAE	0.3298	0.3195	0.3307	0.3259
NMAE by mean	0.7788	0.7545	0.781	0.7696

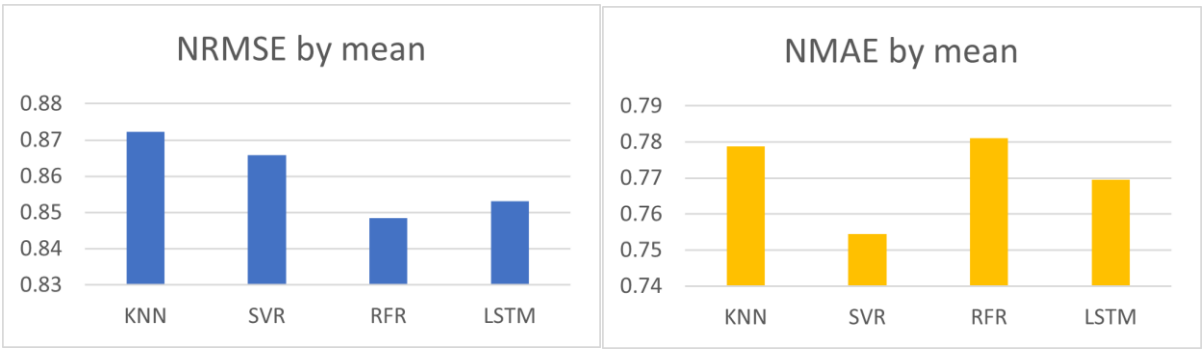


Figure 5.20 Errors: Shallow zone earthquakes M≥5.5, Five Days Delay

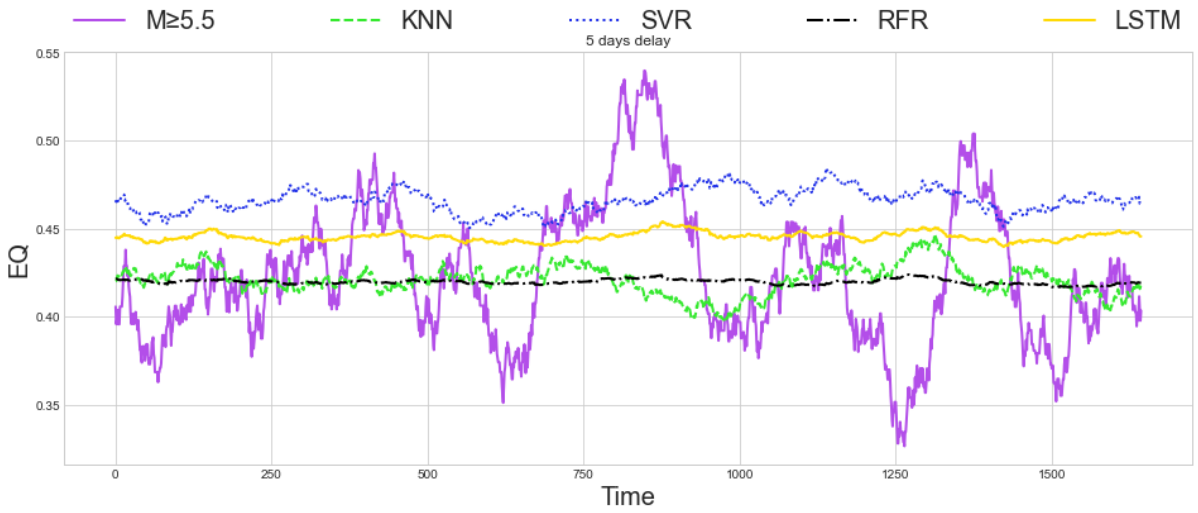


Figure 5.21 Shallow zone earthquakes M≥5.5: compare actual and predicted values, Five Days Delay

Table 5.11 together with Figure 5.22 and Figure 5.23 show that in both metrics, the six-day delay had the same results as the three-day delay part.

Table 5.11 Shallow zone earthquakes M≥5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3748	0.3714	0.3633	0.3638
NRMSE by SD	1.0321	1.0227	1.0004	1.0018
NRMSE by mean	0.9064	0.8981	0.8786	0.8798
Mean absolute error				
MAE	0.3364	0.3275	0.3351	0.3333
NMAE by mean	0.8135	0.7921	0.8104	0.806

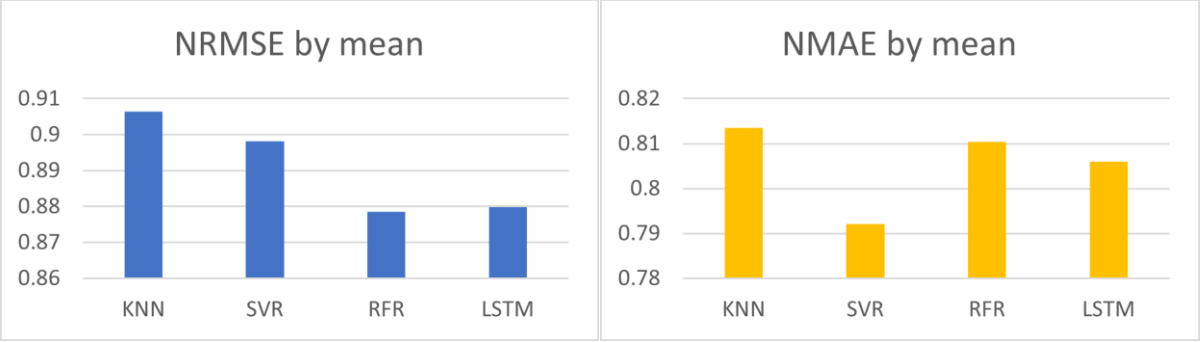


Figure 5.22 Shallow zone earthquakes M≥5.5, Six Days Delay

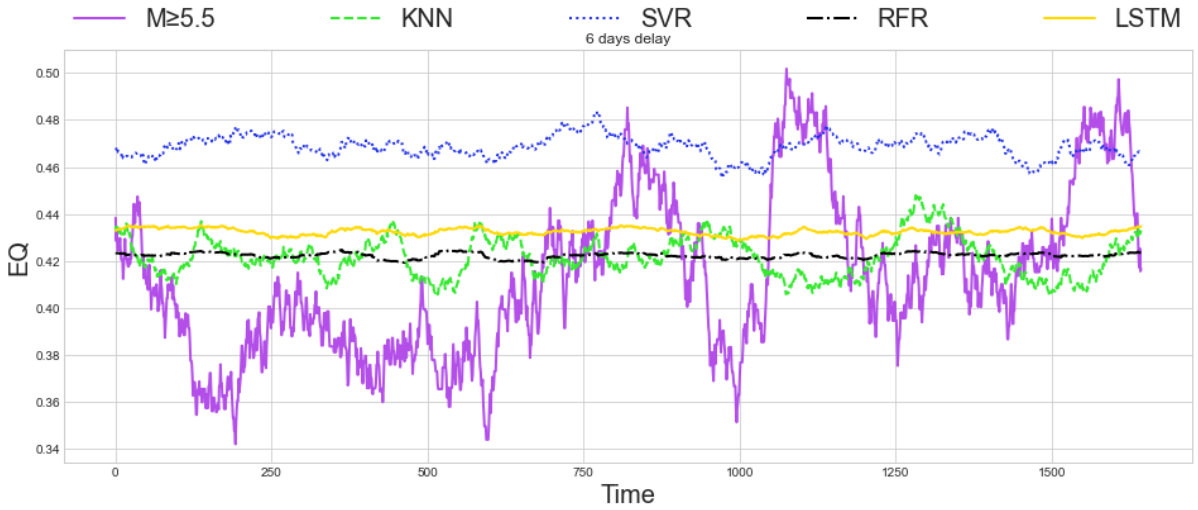


Figure 5.23 Shallow zone earthquakes M≥5.5: compare actual and predicted values, Six Days Delay

Table 5.12 and Figure 5.24 and Figure 5.25 demonstrate that, for both metrics, the seven-day delay produced the same outcomes as the three- and six-day delay parts.

Table 5.12 Shallow zone earthquakes M≥5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3787	0.3731	0.3653	0.366
NRMSE by SD	1.0372	1.022	1.0006	1.0024
NRMSE by mean	0.9042	0.891	0.8723	0.8738
Mean absolute error				
MAE	0.3411	0.3308	0.3372	0.3327
NMAE by mean	0.8144	0.7898	0.8051	0.7949

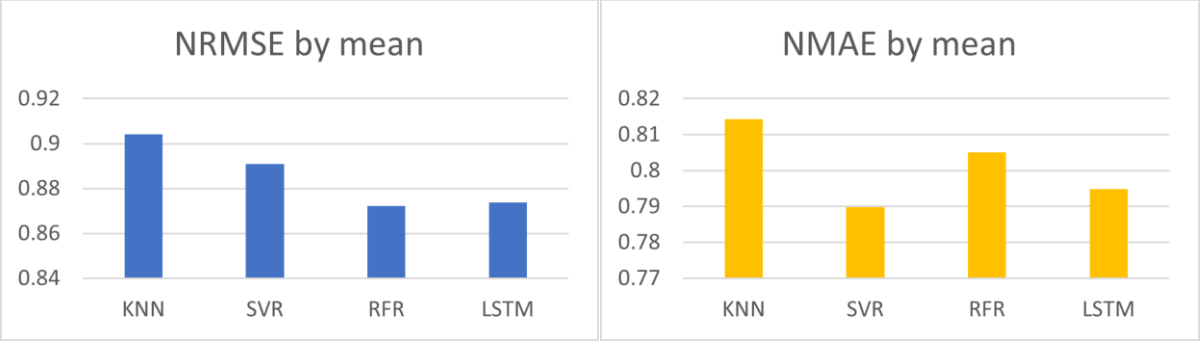


Figure 5.24 Errors: Shallow zone earthquakes M≥5.5, Seven Days Delay

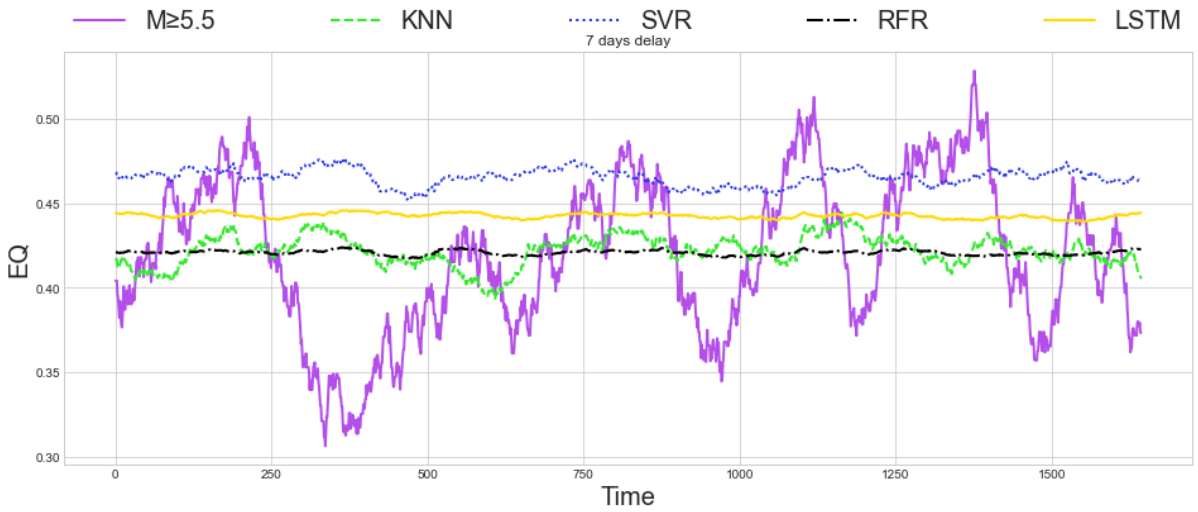


Figure 5.25 Shallow zone earthquakes M≥5.5: compare actual and predicted values, Seven Days Delay

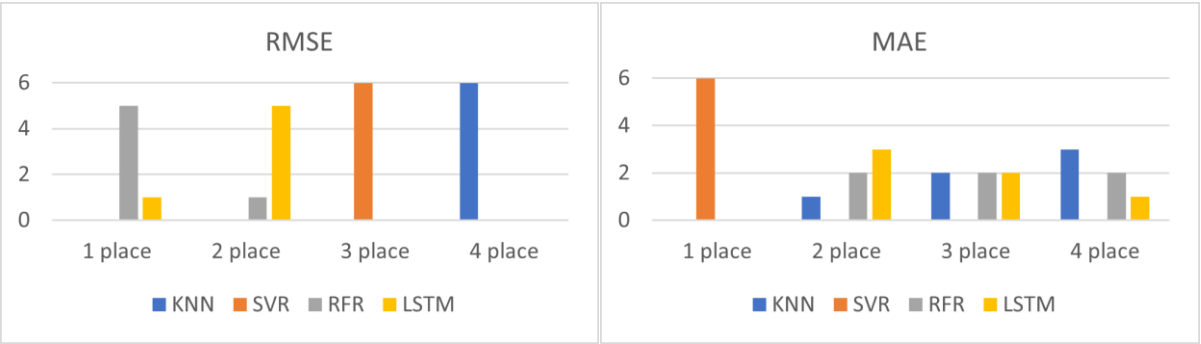


Figure 5.26 Shallow zone earthquakes M≥5.5, summarising results

Figure 5.26 demonstrates that RFR and LSTM had the highest accuracy in terms of NRMSE. Whereas, in terms of NMAE, SVR had the highest accuracy, followed by LSTM. It was found that the three-day delay part had the highest accuracy in terms of NRMSE and the five-day delay part had the highest accuracy in terms of NMAE. Furthermore, both metrics values in

this dataset are greater than the metrics values in the dataset for shallow zone earthquakes with a Richter magnitude less than 5. Furthermore, it was noted that NRMSE values normalised by standard deviation were very close to "1" or even greater than "1". What is more, the earthquakes have peaks, just like in the earlier segments of the experiment. The values in the earthquake data that are further from the mean are significant. That is why RMSE values are preferred in this instance as well as the one before it.

5.3 *Solar activity and Intermediate zone earthquakes with a Richter magnitude less than 5.5*

Table 5.13 through Table 5.18 summarise the results of each section of the intermediate zone earthquake experiment. Figure 5.27 through Figure 5.38 show the graphical interpretation of the above tables and the disparities between earthquakes' actual and projected values.

Table 5.13 Intermediate zone earthquakes M<5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2778	0.2751	0.2778	0.2742
NRMSE by SD	0.9674	0.9581	0.9674	0.9547
NRMSE by mean	0.5491	0.5438	0.5491	0.542
Mean absolute error				
MAE	0.2347	0.2306	0.2372	0.2325
NMAE by mean	0.4639	0.4558	0.4688	0.4597

Table 5.13 and Figure 5.27 demonstrate that the first place, in the two-day delay part, was LSTM, whereas RFR and KNN shared the last place. In terms of NMAE, the first place had SVR, the second had LSTM, and the last had RFR. Figure 5.28 shows that the prediction lines of the algorithms are located almost together, while the LSTM prediction line is located slightly separated from the others and, in some cases, follows the actual values lines better than the others.

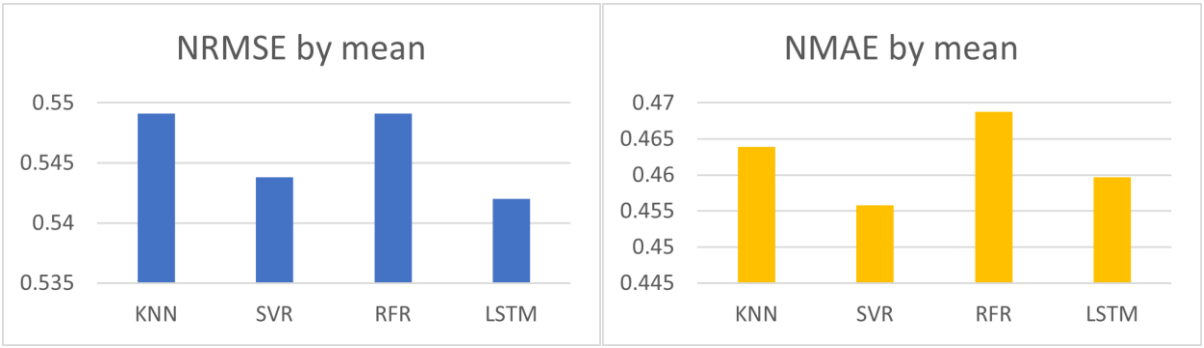


Figure 5.27 Errors: Intermediate zone earthquakes M<5.5, Two Days Delay

Table 5.14 and Figure 5.29 illustrate the three-day delay part. In terms of NRMSE, the first place went to LSTM, and the last place went to KNN, with SVR and RFR in second and third

place, respectively. The first two places in terms of NMAE were SVR and LSTM, and the last two places were RFR and KNN, respectively. Figure 5.30 demonstrates that all algorithms' prediction lines, as well as those for the two-day delay, go together. However, in some cases, LSTM and KNN diverge from the other algorithms.

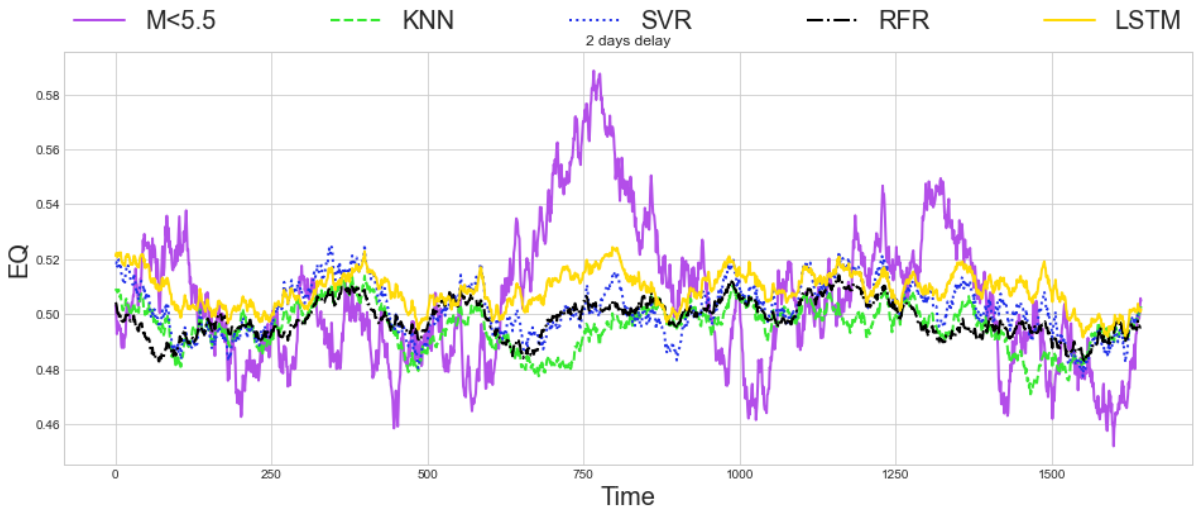


Figure 5.28 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Two Days Delay

Table 5.14 Intermediate zone earthquakes M<5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2777	0.276	0.2764	0.2737
NRMSE by SD	0.9715	0.9656	0.967	0.9574
NRMSE by mean	0.5442	0.5409	0.5417	0.5363
Mean absolute error				
MAE	0.2355	0.2314	0.2354	0.2317
NMAE by mean	0.4615	0.4535	0.4614	0.4542

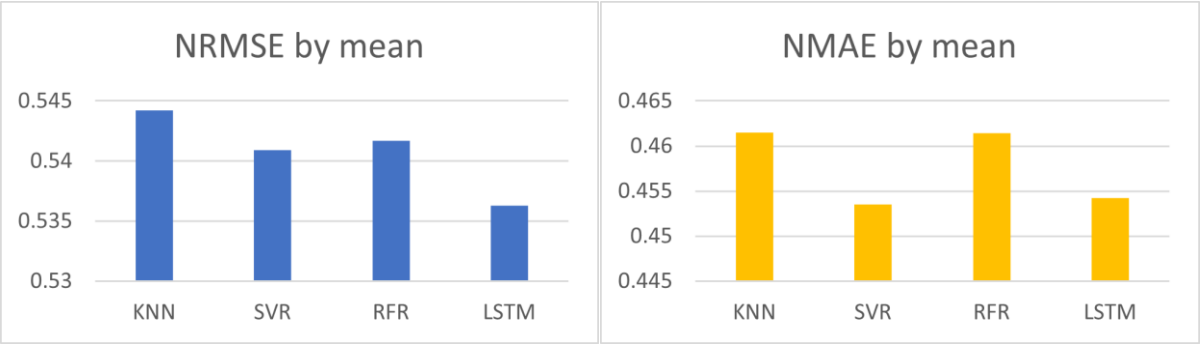


Figure 5.29 Errors: Intermediate zone earthquakes M<5.5, Three Days Delay

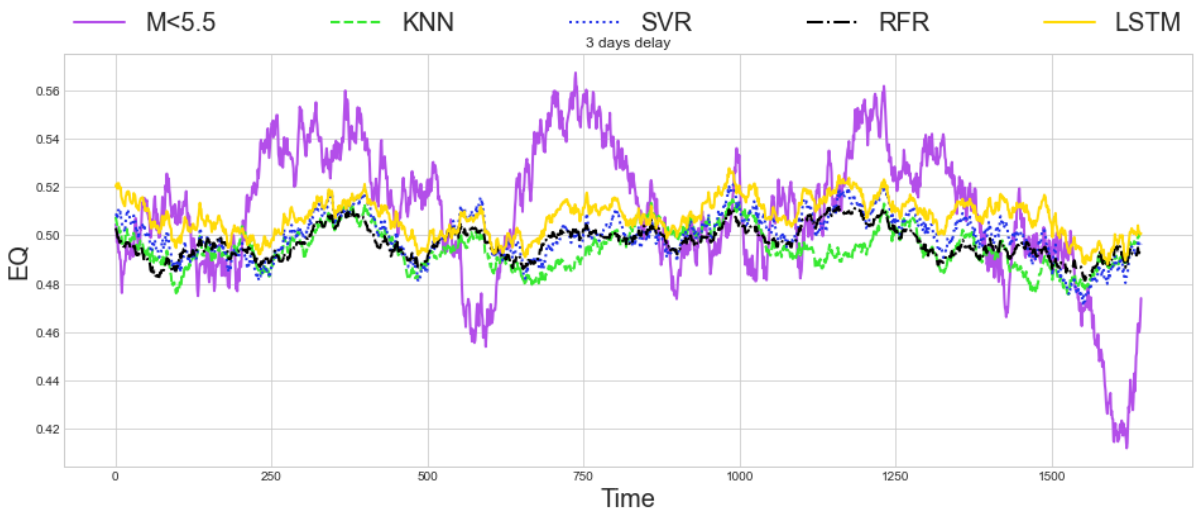


Figure 5.30 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Three Days Delay

Table 5.15, as well as Figure 5.31 and Figure 5.32, demonstrate that the components of the four-day delay were identical to those of the three-day delay.

Table 5.15 Intermediate zone earthquakes with Richter magnitude less than 5.5 (M < 5.5) Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2754	0.273	0.2733	0.272
NRMSE by SD	0.974	0.9652	0.9663	0.9619
NRMSE by mean	0.5442	0.5393	0.5399	0.5374
Mean absolute error				
MAE	0.231	0.2266	0.2321	0.2297
NMAE by mean	0.4563	0.4477	0.4586	0.4539

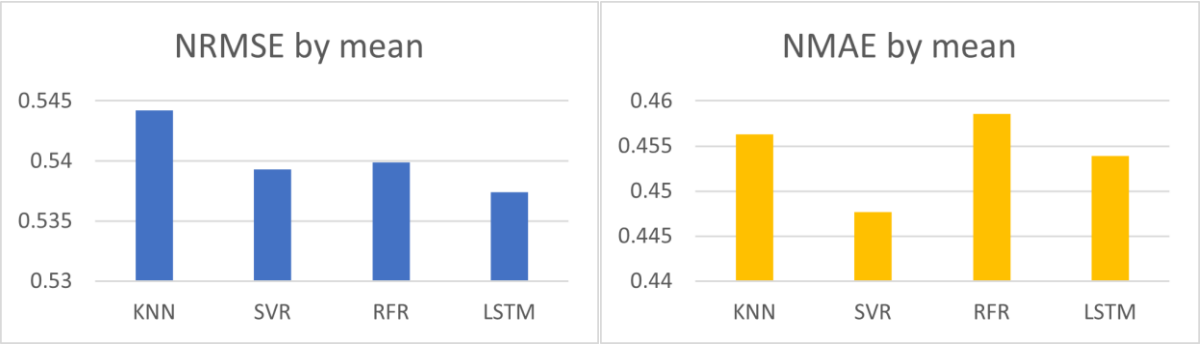


Figure 5.31 Errors: Intermediate zone earthquakes M<5.5, Four Days Delay

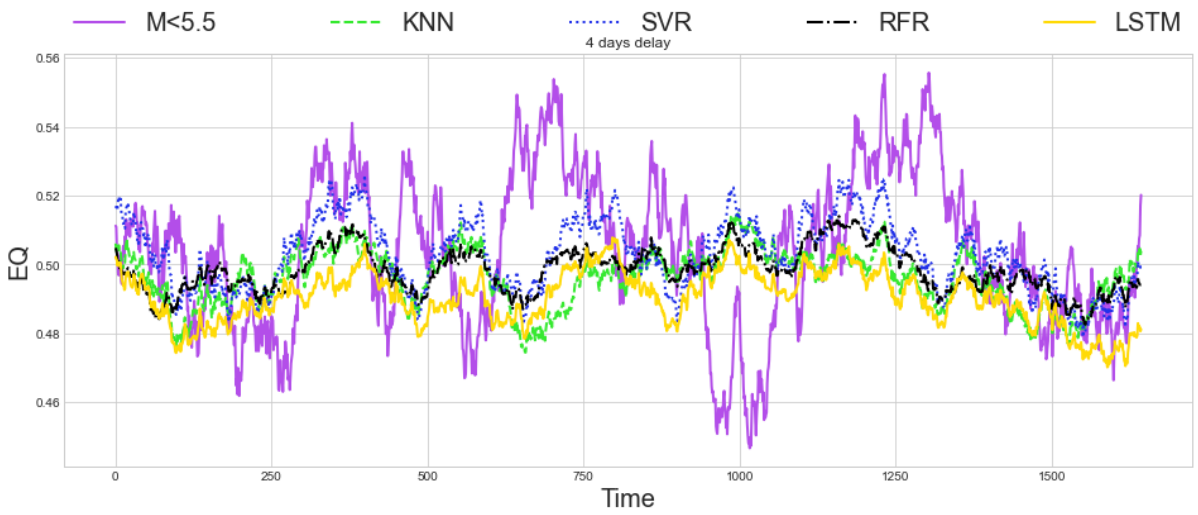


Figure 5.32 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Four Days Delay

Table 5.16 and Figure 5.33 and Figure 5.34 demonstrate that the results for the five-day delay part were the same the those for the three- and four-day delay parts.

Table 5.16 Intermediate zone earthquakes M<5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2789	0.276	0.2762	0.2723
NRMSE by SD	0.9752	0.965	0.9658	0.9522
NRMSE by mean	0.5516	0.5459	0.5464	0.5386
Mean absolute error				
MAE	0.2361	0.2299	0.2347	0.2301
NMAE by mean	0.467	0.4547	0.4642	0.4552

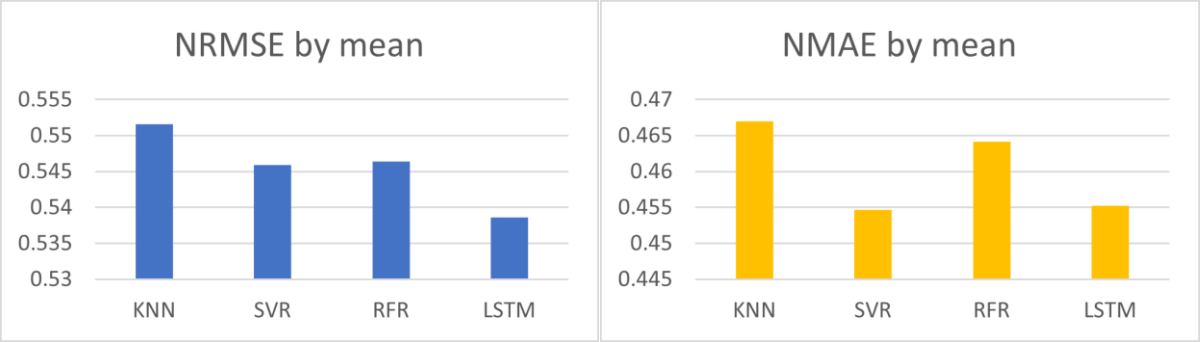


Figure 5.33 Errors: Intermediate zone earthquakes M<5.5, Five Days Delay

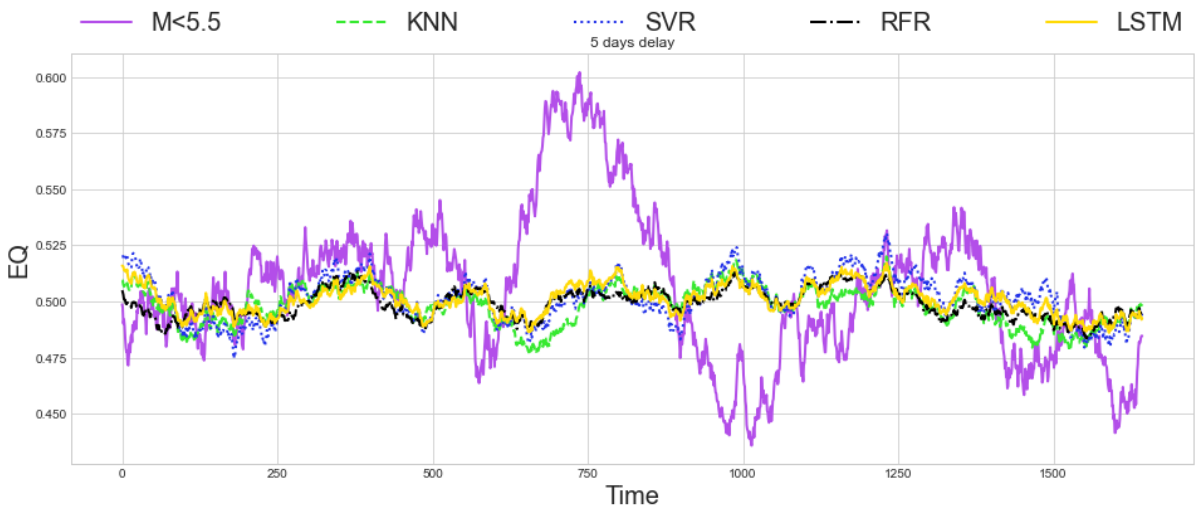


Figure 5.34 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Five Days Delay

Table 5.17 and Figure 5.35 illustrate that in the six-day delay part, the results were slightly different. In terms of NRMSE, SVR and KNN took the first two positions, while LSTM and RFR took the last two positions. Figure 5.36 supports this and demonstrates that the LSTM prediction line is located below the other algorithm prediction lines and below its previous locations.

Table 5.17 Intermediate zone earthquakes M<5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2695	0.2687	0.2715	0.27
NRMSE by SD	0.9568	0.9538	0.9638	0.9585
NRMSE by mean	0.5314	0.5297	0.5353	0.5323
Mean absolute error				
MAE	0.2253	0.2224	0.23	0.2261
NMAE by mean	0.4441	0.4384	0.4534	0.4458

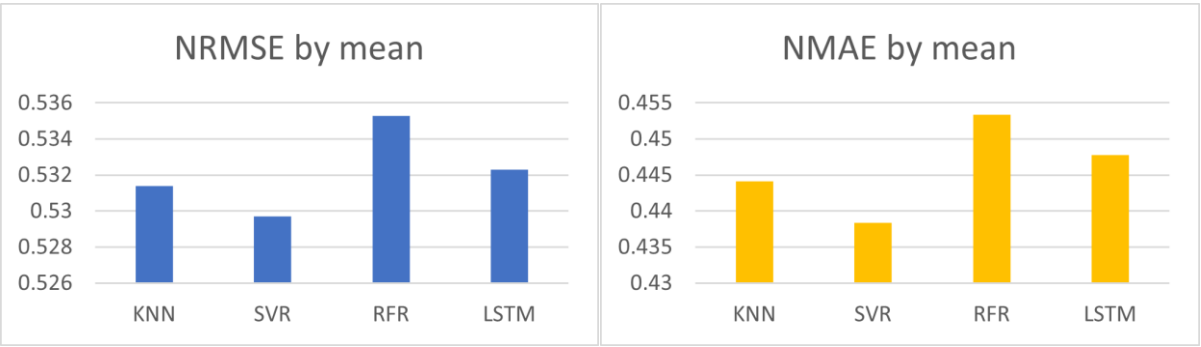


Figure 5.35 Errors: Intermediate zone earthquakes M<5.5, Six Days Delay

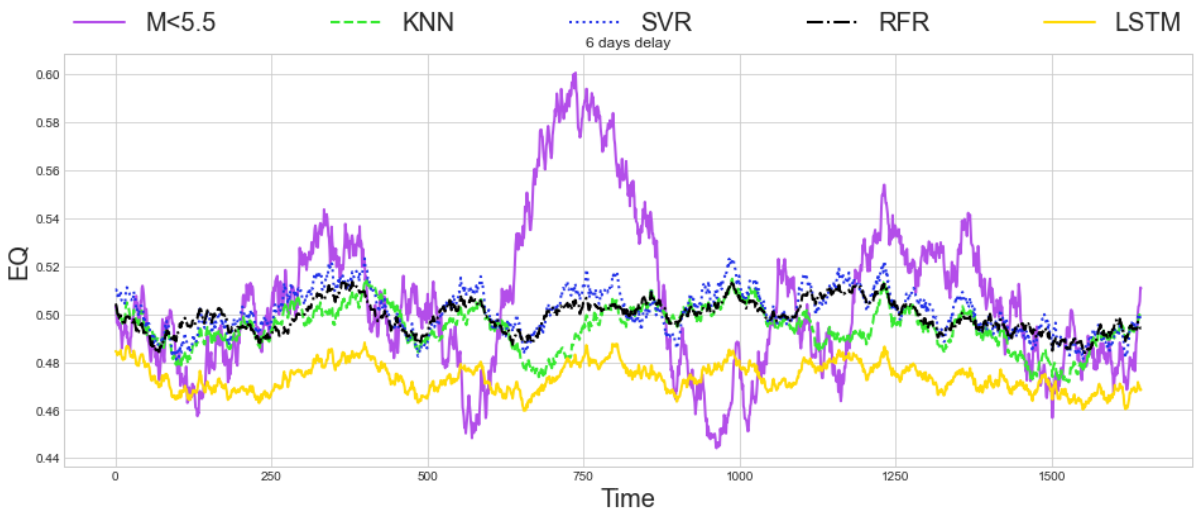


Figure 5.36 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Six Days Delay

SVR and LSTM, respectively, showed the highest accuracy in both metrics with the seven-day delay, according to Table 5.18 and Figure 5.37, while RFR showed the lowest accuracy. Figure 5.38 shows that the LSTM prediction line is currently above the other prediction lines, in contrast to the almost constant convergence of the traditional ML algorithms' prediction lines.

Table 5.18 Intermediate zone earthquakes M<5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2758	0.2732	0.2759	0.274
NRMSE by SD	0.9661	0.9569	0.9666	0.96
NRMSE by mean	0.5457	0.5406	0.5461	0.5423
Mean absolute error				
MAE	0.2324	0.2288	0.2351	0.231
NMAE by mean	0.46	0.4528	0.4653	0.4571

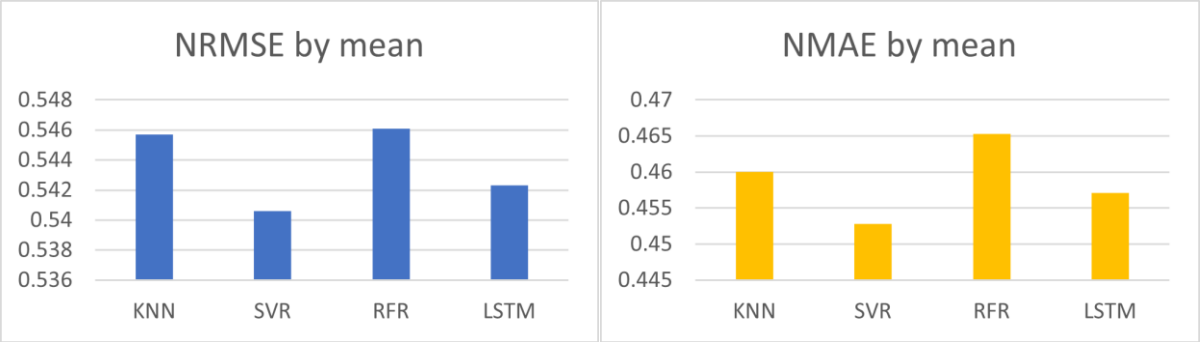


Figure 5.37 Errors: Intermediate zone earthquakes M<5.5, Seven Days Delay

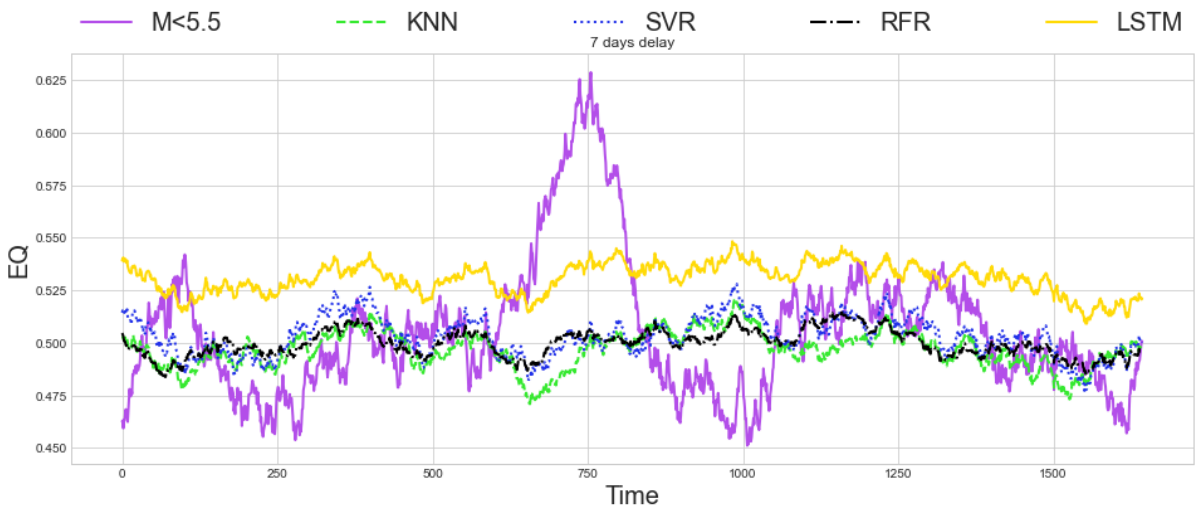


Figure 5.38 Intermediate zone earthquakes M<5.5: compare actual and predicted values, Seven Days Delay

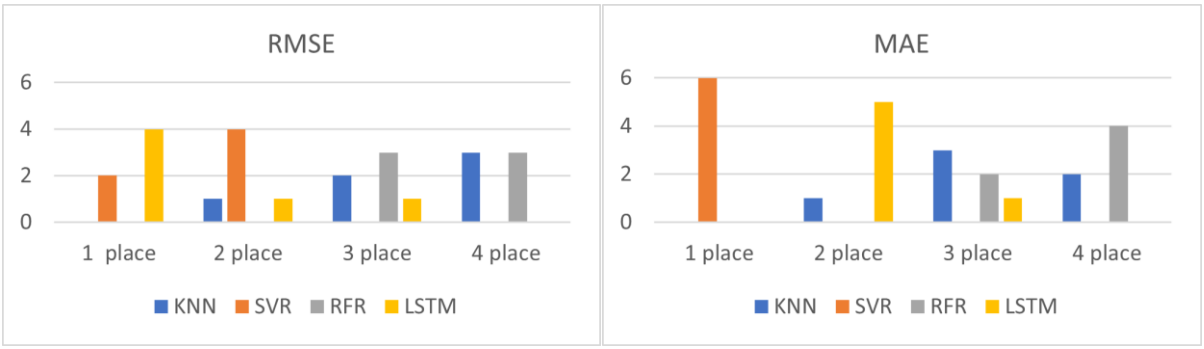


Figure 5.39 Intermediate zone earthquakes M<5.5, summarising results

The results were visualised in Figure 5.39. In both metrics, it was discovered that LSTM and SVR had the highest accuracy. In RNMSE, LSTM was the first, while in NMAE, SVR was the first. In both metrics, RFR and KNN showed the lowest accuracy. In terms of NRMSE, it was discovered that the five-day delay part has the highest accuracy, while the six-day delay part has the highest accuracy in terms of NMAE. What is more, it was noted that intermediate depth data errors are higher than shallow depth data errors. Additionally, just like in the earlier parts of the experiment, the earthquakes have peaks. Significant values can be found in the earthquake data that deviate more from the mean. Because of this, RMSE values are preferred in both this situation and the one before it.

5.4 *Solar activity and Intermediate zone earthquakes with a Richter magnitude equal to or greater than 5.5*

Table 5.19 through Table 5.24 summarise the results of each section of the intermediate zone earthquake with a Richter magnitude of 5.5 or greater. Figure 5.40 through Figure 5.51 demonstrate the graphical interpretation of the above tables and the differences between actual and forecast earthquake magnitudes.

Table 5.19 and Figure 5.40 demonstrate that in the two-day delay in terms of NRMSE, LSTM had the highest accuracy, while KNN had the lowest accuracy. In terms of NMAE, SVR had the highest accuracy, while KNN had the lowest accuracy. It was also found that the normalised errors in both metrics are quite large, which shows that the prediction is not accurate enough. The line graphs in Figure 5.41 also show that the results are far from ideal. While the LSTM, RFR, and KNN prediction lines are located near the averages of the actual values, the SVR prediction line is located at the lower range of the actual values line.

Table 5.19 Intermediate zone earthquakes M≥5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.344	0.3323	0.3306	0.33
NRMSE by SD	1.0423	1.0068	1.0016	0.9998
NRMSE by mean	2.4741	2.3899	2.3775	2.3732
Mean absolute error				
MAE	0.2433	0.2088	0.241	0.2352
NMAE by mean	1.75	1.502	1.7332	1.6914

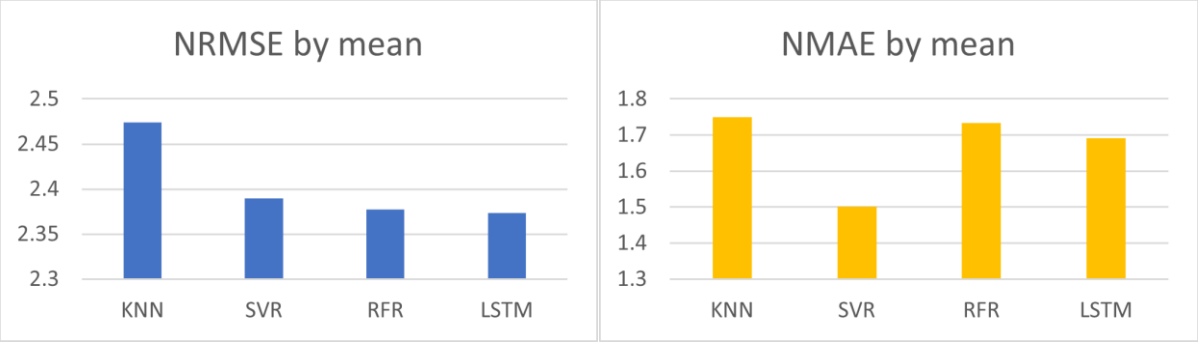


Figure 5.40 Errors: Intermediate zone earthquakes M≥5.5, Two Days Delay

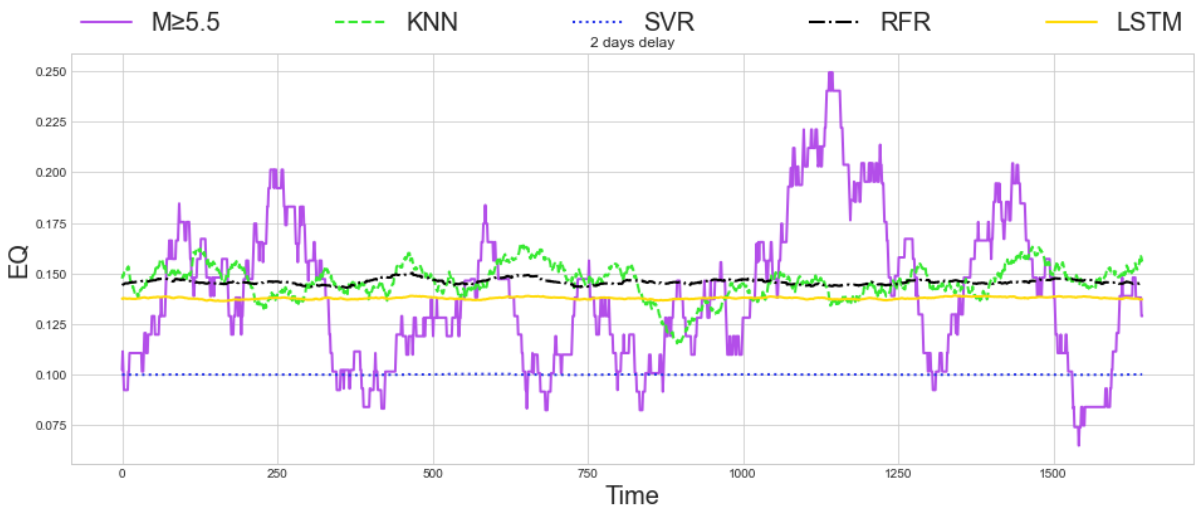


Figure 5.41 Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Two Days Delay

The three-day delay results, Table 5.20 and Figure 5.42, showed that, in terms of NRMSE, LSTM had the highest accuracy and KNN had the lowest accuracy. In terms of NMAE, SVR had the highest accuracy, and RFR had the lowest accuracy. The metric values were similarly high as in the prior part of the experiment. The prediction lines of all algorithms (Figure 5.43) repeat the location as it was in the two-day delay part.

Table 5.20 Intermediate zone earthquakes M≥5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3399	0.3325	0.3306	0.3302
NRMSE by SD	1.0293	1.0069	1.0011	1.0001
NRMSE by mean	2.4378	2.3847	2.3708	2.3684
Mean absolute error				
MAE	0.2391	0.209	0.241	0.2343
NMAE by mean	1.7152	1.4991	1.7282	1.6802

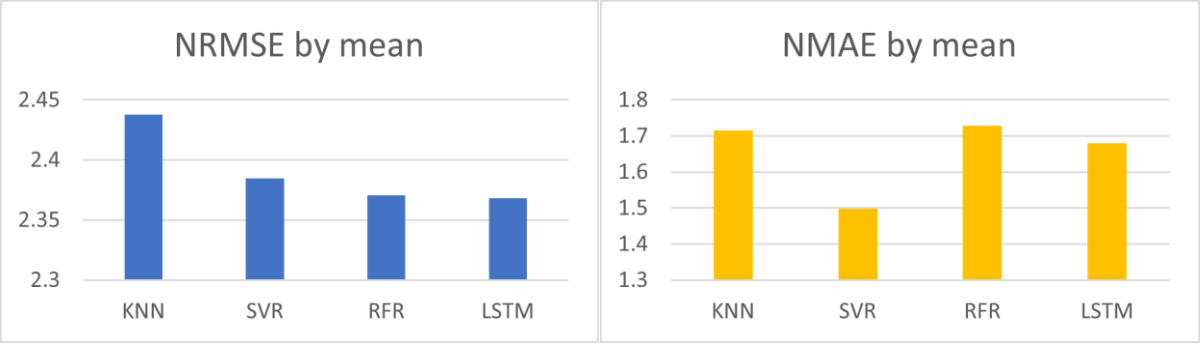


Figure 5.42 Errors: Intermediate zone earthquakes M≥5.5, Three Days Delay

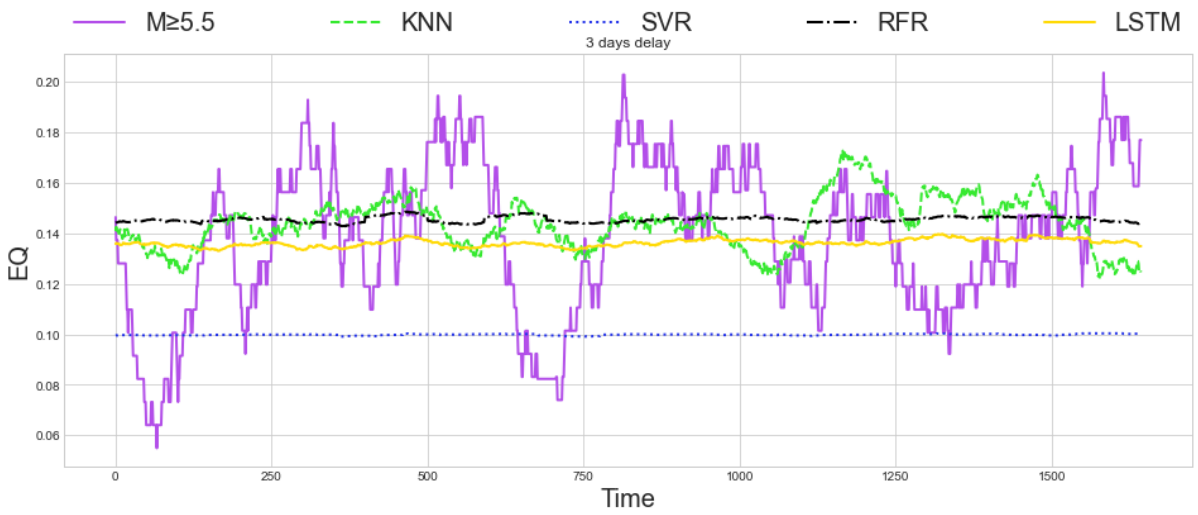


Figure 5.43 : Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Three Days Delay

Table 5.21 and Figure 5.44 illustrate that, as with previous parts of the experiment, in the four-day delay part, the metrics values were rather high. However, in terms of NRMSE, LSTM had the highest accuracy. In terms of NMAE, SVR had the highest accuracy. KNN had the lowest accuracy in both metrics. The line graphs in Figure 5.45 are repeated at the same location as in the previous two- and three-day delay parts.

Table 5.21 Intermediate zone earthquakes M≥5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	2.4378	0.3469	0.3429	0.3425
NRMSE by SD	1.0341	1.0114	0.9998	0.9987
NRMSE by mean	2.3258	2.2748	2.2487	2.2464
Mean absolute error				
MAE	0.2519	0.2192	0.2473	0.2496
NMAE by mean	1.6521	1.4377	1.6217	1.6371

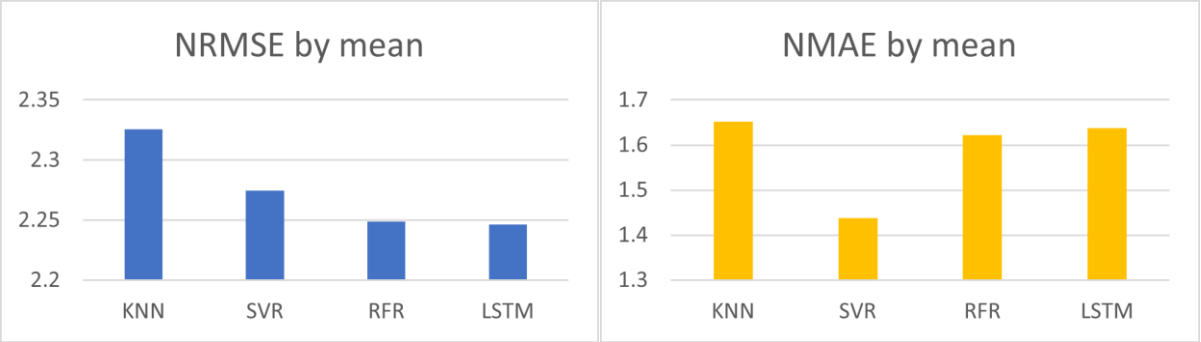


Figure 5.44 Errors: Intermediate zone earthquakes M≥5.5, Four Days Delay

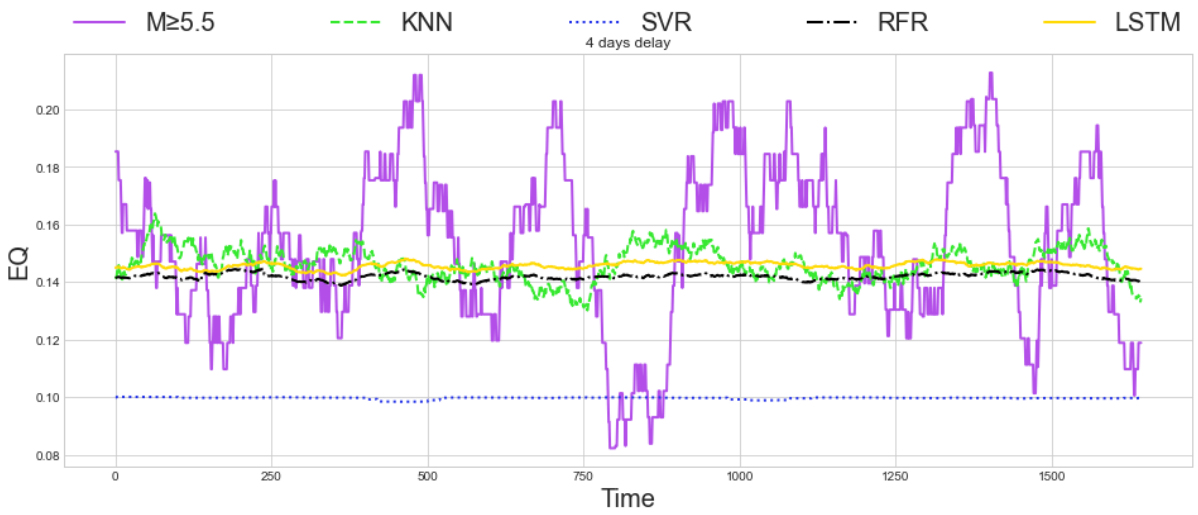


Figure 5.45 Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Four Days Delay

Table 5.22 and Figure 5.46 show that in the five-day delay part, in terms of NRMSE, the results repeat the position as in the previous parts. However, in terms of NMAE, LSTM has the lowest accuracy. Also, as in the previous parts, the error values are too high. The line graphs in Figure 5.47 are repeated at the same location as in the previous parts. Also, it was noted that the prediction lines of all algorithms did not follow the actual values line.

Table 5.22 Intermediate zone earthquakes M≥5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3386	0.3327	0.3303	0.3302
NRMSE by SD	1.0244	1.0066	0.9995	0.9992
NRMSE by mean	2.4317	2.3895	2.3726	2.3719
Mean absolute error				
MAE	0.2353	0.2089	0.2405	0.2433
NMAE by mean	1.6903	1.5004	1.7272	1.7478

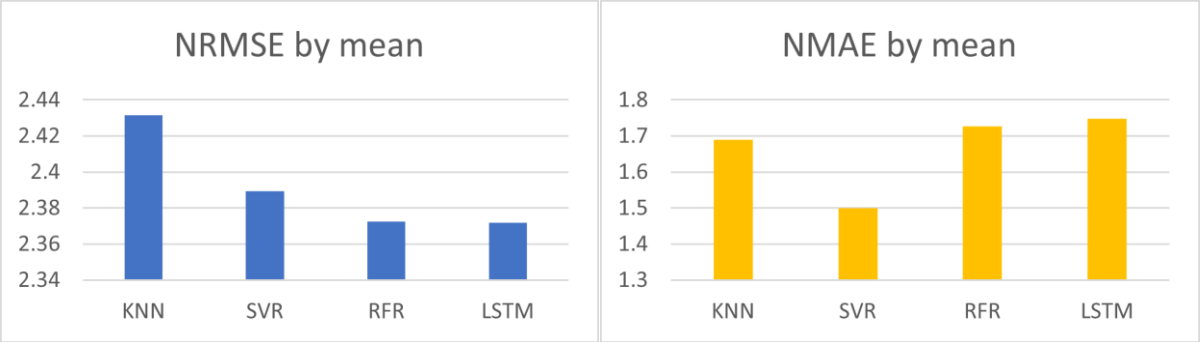


Figure 5.46 Errors: Intermediate zone earthquakes M≥5.5, Five Days Delay

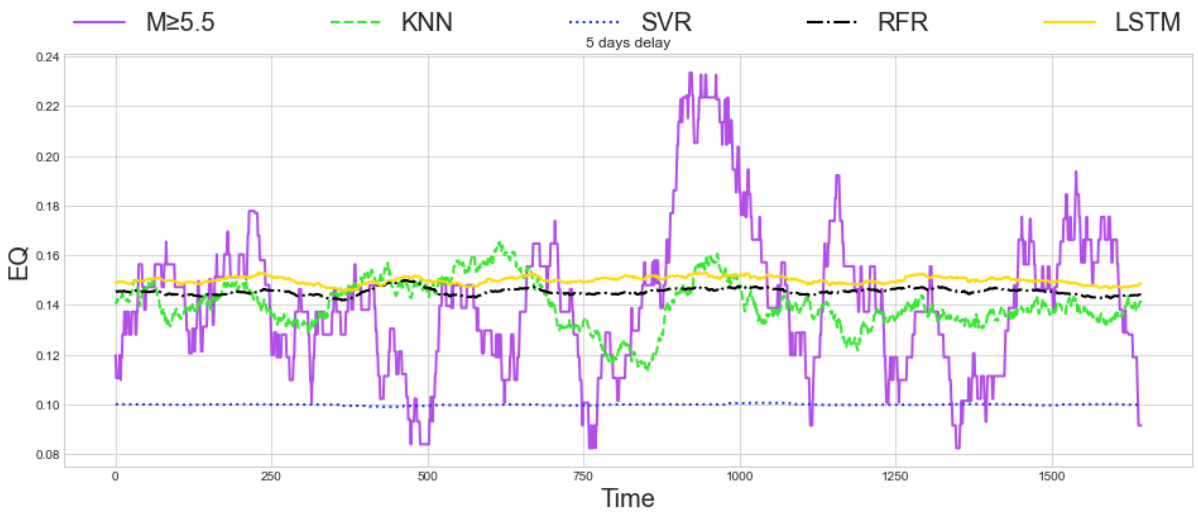


Figure 5.47 Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Five Days Delay

Table 5.23, Figure 5.48, and Figure 5.49 demonstrate that in the six-day delay part, the results and locations of the prediction lines were the same as in the previous parts of the experiment: the two-day delay and four-day delay parts. The outcome is quite high, as it was in the previous parts.

Table 5.23 Intermediate zone earthquakes M≥5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3419	0.3337	0.3319	0.3311
NRMSE by SD	1.0316	1.007	1.0016	0.9992
NRMSE by mean	2.4267	2.3688	2.3561	2.3505
Mean absolute error				
MAE	0.2422	0.21	0.2418	0.2355
NMAE by mean	1.7192	1.4909	1.7167	1.6714

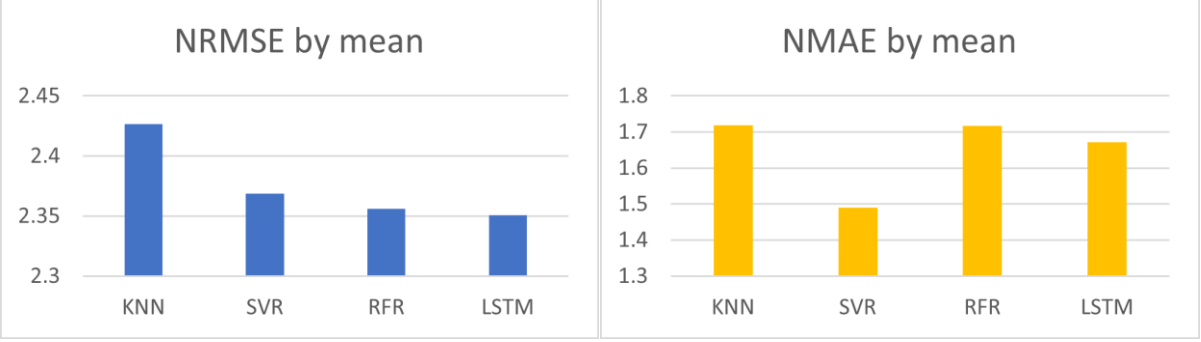


Figure 5.48 Errors: Intermediate zone earthquakes M≥5.5, Six Days Delay

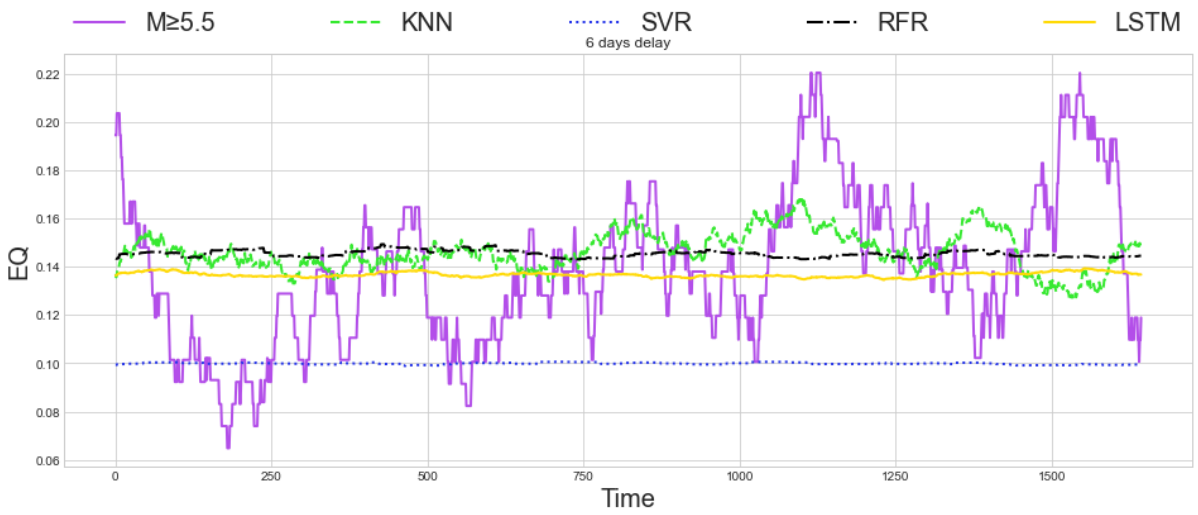


Figure 5.49 Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Six Days Delay

Table 5.24, Figure 5.50, and Figure 5.51 show that the outcomes and locations of the prediction lines in the seven-day delay part were the same as in the five-day delay part of the experiment. Like in the earlier parts, the result is quite high.

Table 5.24 Intermediate zone earthquakes M≥5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3483	0.3441	0.3418	0.3403
NRMSE by SD	1.023	1.0108	1.004	0.9996
NRMSE by mean	2.3101	2.2827	2.2671	2.2573
Mean absolute error				
MAE	0.2474	0.2177	0.2477	0.2482
NMAE by mean	1.6408	1.4442	1.643	1.6462

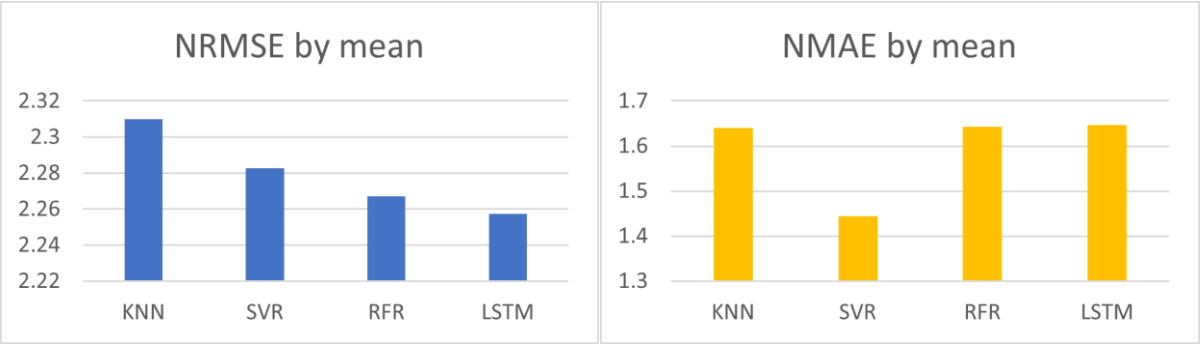


Figure 5.50 Errors: Intermediate zone earthquakes M≥5.5, Seven Days Delay

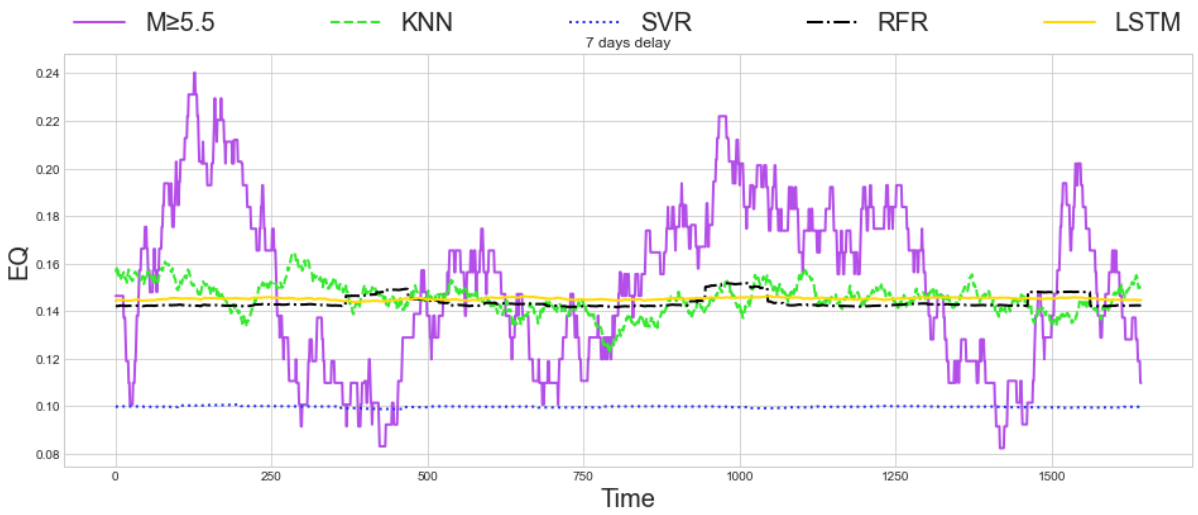


Figure 5.51 Intermediate zone earthquakes M≥5.5: compare actual and predicted values, Seven Days Delay

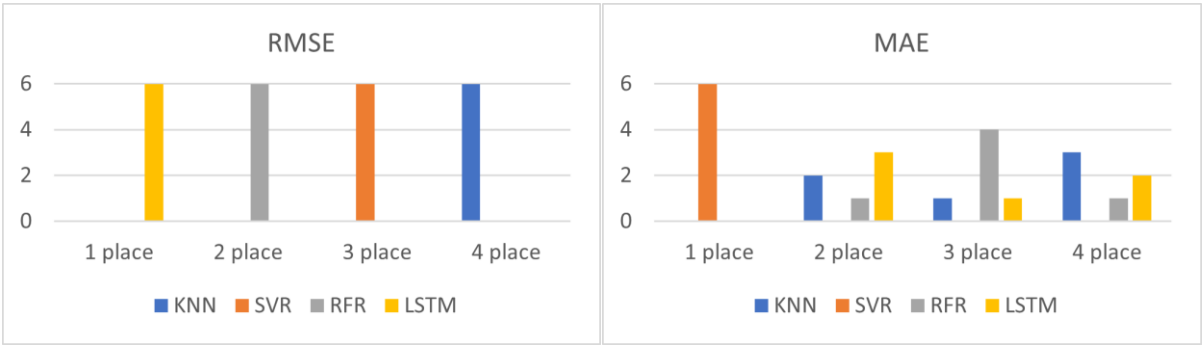


Figure 5.52 Intermediate zone earthquakes M≥5.5, summarising results

The visualisation of the findings can be seen in Figure 5.52. Everything is clear in terms of RMSE; all techniques had the same positions in the every-day delay parts. In terms of NMAE, however, only SVR had the highest accuracy; with all other techniques, the accuracy was changing. However, the normalised error values show that the prediction is not perfect. Also, it was noted that the error values are greater than in the previous segments of the experiment with global and shallow earthquakes. It was discovered that the four-day delay part produced the highest accuracy in both metrics. What is more, the earthquake data contains significant values that deviate more from the mean. As a result, in both this scenario and the one before it, RMSE values are preferred.

5.5 *Solar activity and Deep zone earthquakes with a Richter magnitudes less than 5.5*

The results of each part of the deep zone earthquake (M<5.5) experiment are summarised in Table 5.25 through Table 5.30. Figure 5.53 through Figure 5.64 represent the graphical interpretation of the above tables and the differences between actual and predicted earthquake values.

Table 5.25 and Figure 5.53 demonstrate that in the two-day delay part, LSTM and RFR showed the highest accuracy in both metrics, while KNN had the lowest accuracy in both metrics. However, it was noted that the SVR and KNN values of NRMSE by standard deviation are greater than "1" and the LSTM and RFR values are very close to "1". Figure 5.54 shows that none of the algorithms' prediction lines perfectly match the line representing actual values. The SVR prediction line is located above the others. The LSTM and RFR prediction lines are located around the averages of the actual values.

Table 5.25 Deep zone earthquakes M<5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3078	0.3049	0.3008	0.3
NRMSE by SD	1.0167	1.0069	0.9934	0.9908
NRMSE by mean	0.6417	0.6355	0.6269	0.6253
Mean absolute error				
MAE	0.2611	0.2598	0.2548	0.2554
NMAE by mean	0.5443	0.5416	0.5311	0.5323

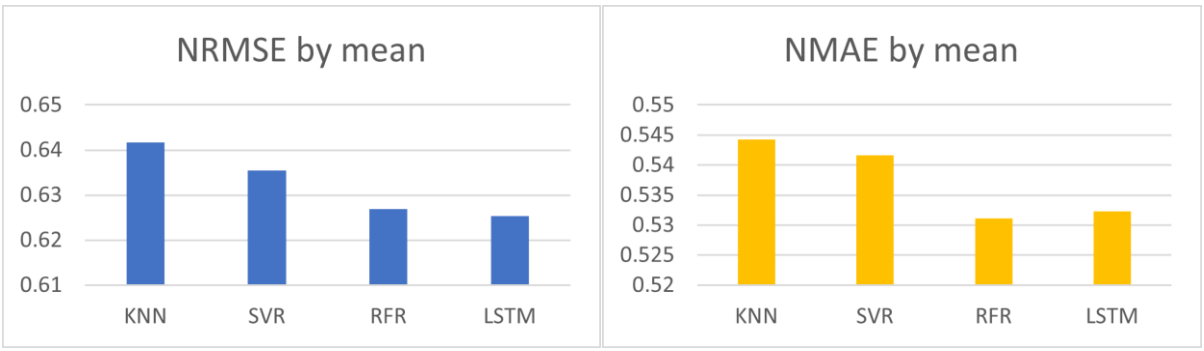


Figure 5.53 Errors: Deep zone earthquakes M<5.5, Two Days Delay

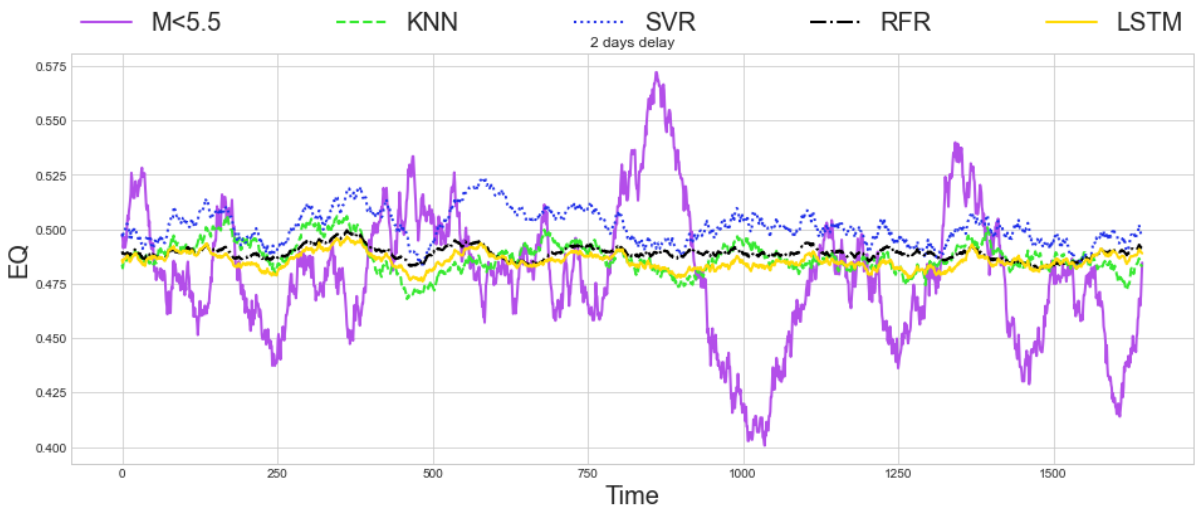


Figure 5.54 Deep zone earthquakes M<5.5: compare actual and predicted values, Two Days Delay

Table 5.26 together with Figure 5.55 and Figure 5.56 illustrate that in the three-day delay part, the range of the results and the prediction lines are repeated from the two-day delay part. Furthermore, NRMSE by standard deviation values are very close to or greater than "1".

Table 5.26 Deep zone earthquakes M<5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3145	0.3122	0.3081	0.3071
NRMSE by SD	1.0159	1.0085	0.9953	0.9921
NRMSE by mean	0.6442	0.6395	0.6312	0.6292
Mean absolute error				
MAE	0.266	0.265	0.2602	0.2603
NMAE by mean	0.5449	0.5428	0.5331	0.5333

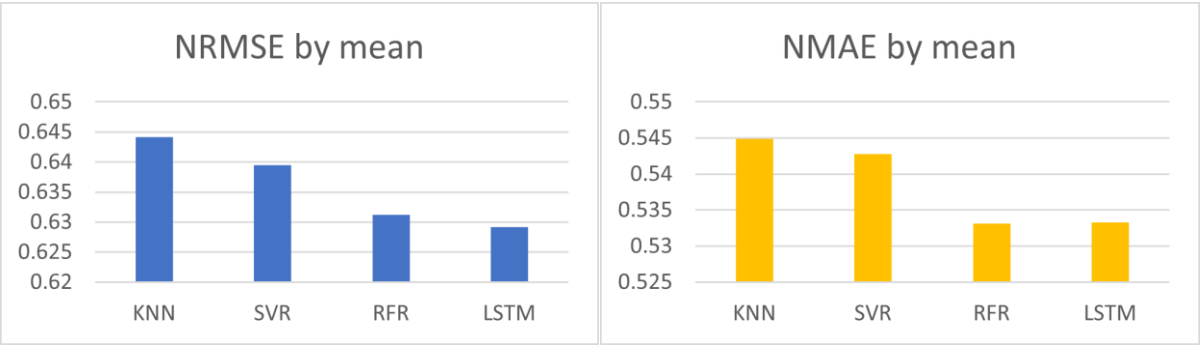


Figure 5.55 Errors: Deep zone earthquakes M<5.5, Three Days Delay

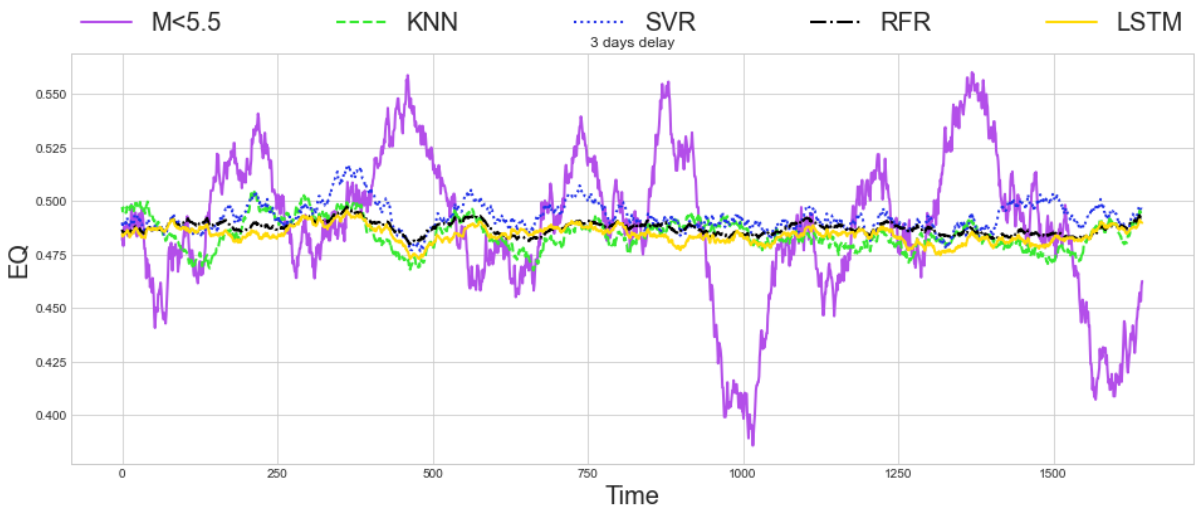


Figure 5.56 Deep zone earthquakes M<5.5: compare actual and predicted values, Three Days Delay

Table 5.27 and Figure 5.57 how that in the four-day delay part, in both metrics, RFR and LSTM had the highest accuracy, respectively. The lowest accuracy got KNN in NRMSE and SVR in NMAE. Figure 5.58 demonstrates that the RFR, SVR, and KNN prediction lines are located approximately at the same location as they were in the previous parts, while the LSTM prediction line is located below the others. As well as with the previous parts, the SVR and KNN values of NRMSE by standard deviation are greater than "1" and the RFR and LSTM are very close to "1".

Table 5.27 Deep zone earthquakes M<5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3095	0.3078	0.3017	0.3034
NRMSE by SD	1.0169	1.0114	0.9913	0.9967
NRMSE by mean	0.636	0.6325	0.62	0.6234
Mean absolute error				
MAE	0.2611	0.2619	0.2559	0.2571
NMAE by mean	0.5365	0.5381	0.5257	0.5284

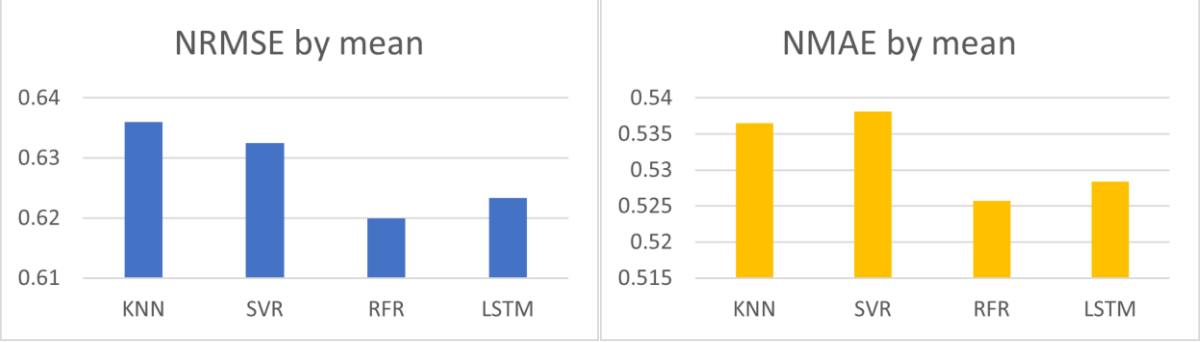


Figure 5.57 Errors: Deep zone earthquakes M< 5.5, Four Days Delay

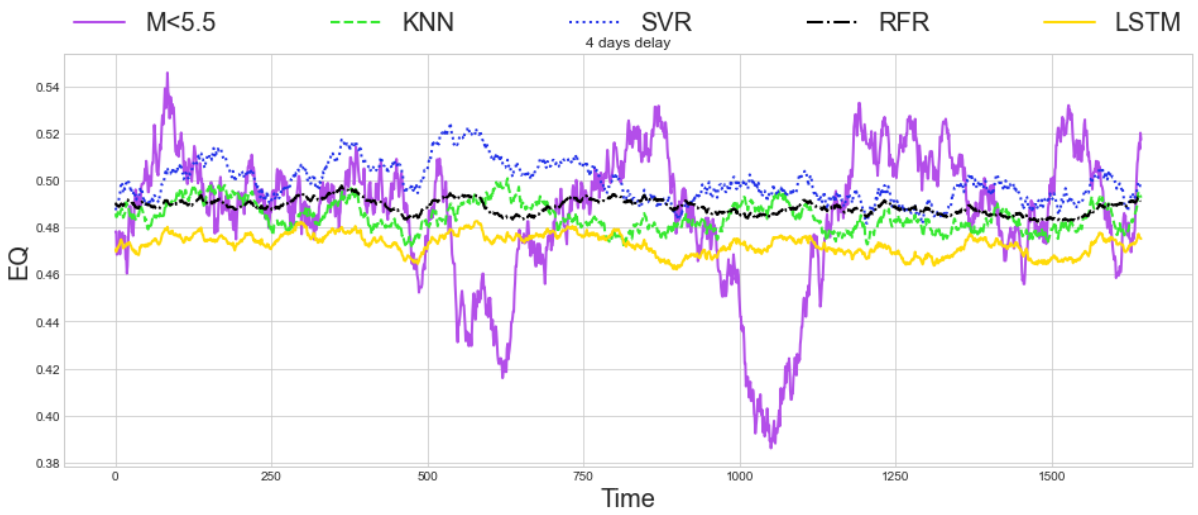


Figure 5.58 Deep zone earthquakes M< 5.5: compare actual and predicted values, Four Days Delay

Table 5.28 and Figure 5.59 show that in the five-day delay, RFR had the highest accuracy in both metrics, followed by LSTM, while KNN had the lowest accuracy in both metrics. Figure 5.60 shows that the LSTM prediction line is located above the previous positions, while the others are located where they were in the previous parts. The NRMSE by standard deviation values are greater or very close to "1" as in the previous parts.

Table 5.28 Deep zone earthquakes M<5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3082	0.3043	0.3009	0.3018
NRMSE by SD	1.0158	1.0031	0.9916	0.9947
NRMSE by mean	0.6386	0.6306	0.6234	0.6253
Mean absolute error				
MAE	0.2609	0.2594	0.2557	0.2569
NMAE by mean	0.5405	0.5376	0.5297	0.5322

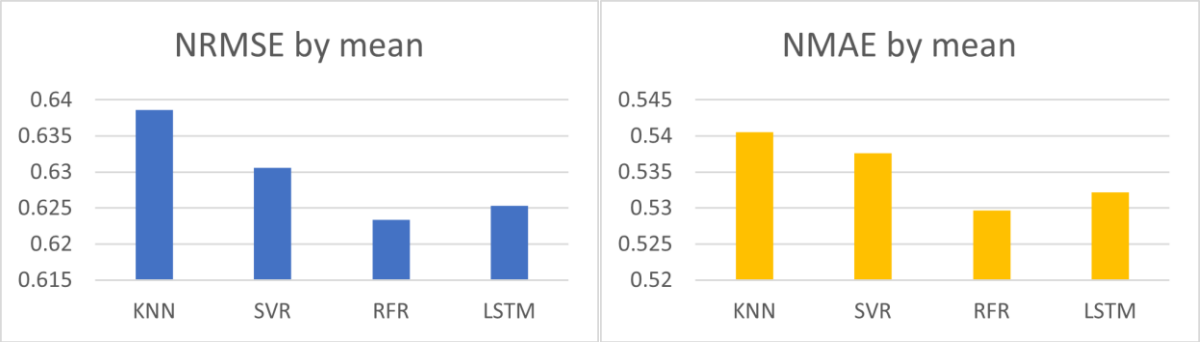


Figure 5.59 Errors: Deep zone earthquakes M<5.5, Five Days Delay

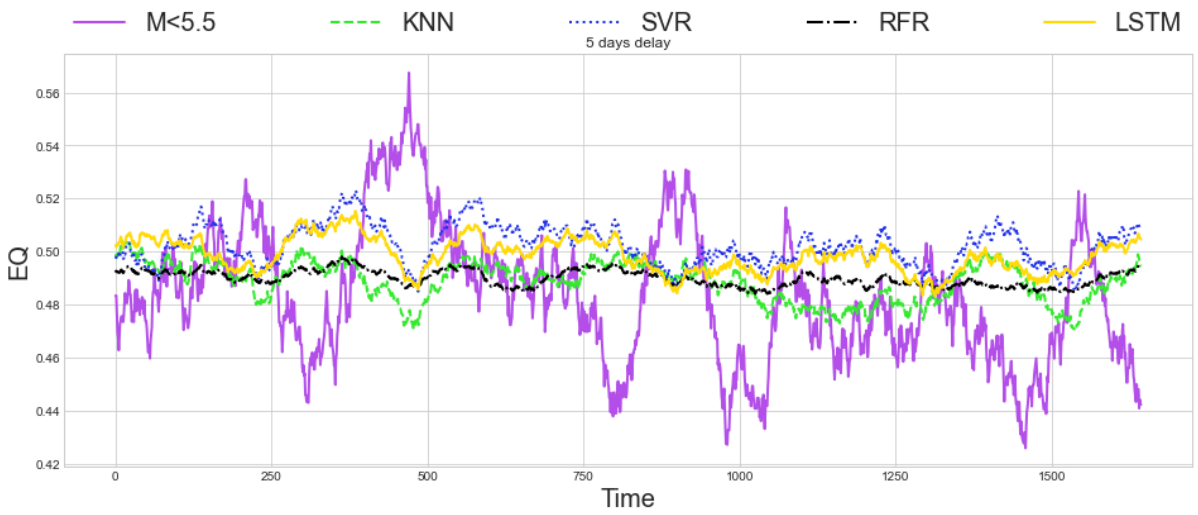


Figure 5.60 Deep zone earthquakes M<5.5: compare actual and predicted values, Five Days Delay

Table 5.29 and Figure 5.61 demonstrate that in the six-day delay in both metrics, LSTM had the highest accuracy, followed by RFR, while SVR and KNN had the lowest accuracy. As in the earlier parts, the NRMSE by standard deviation values are greater than or very close to "1". Figure 5.62 shows the previous locations of the KNN, SVR, and RFR prediction lines, while the LSTM prediction line is located below the others and its previous locations.

Table 5.29 Deep zone earthquakes M<5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3103	0.3056	0.3042	0.3034
NRMSE by SD	1.0149	0.9995	0.9951	0.9923
NRMSE by mean	0.6303	0.6207	0.618	0.6163
Mean absolute error				
MAE	0.2624	0.2595	0.258	0.2574
NMAE by mean	0.5331	0.5271	0.5241	0.5228

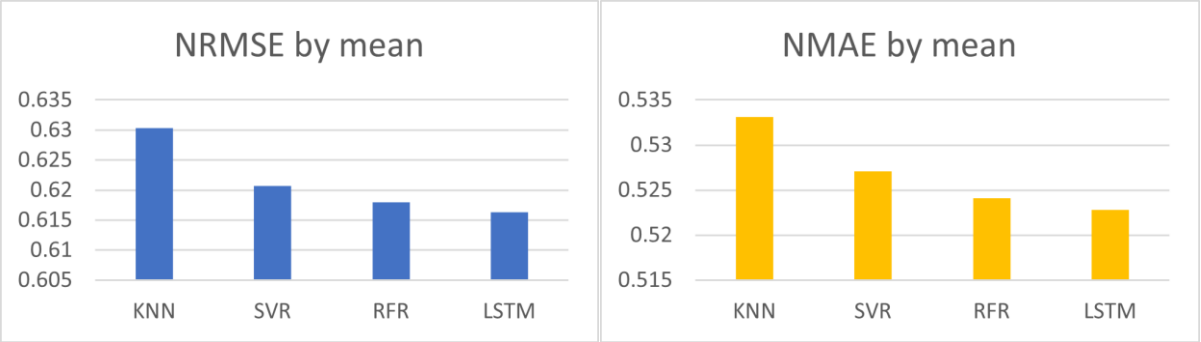


Figure 5.61 Errors: Deep zone earthquakes M<5.5, Six Days Delay

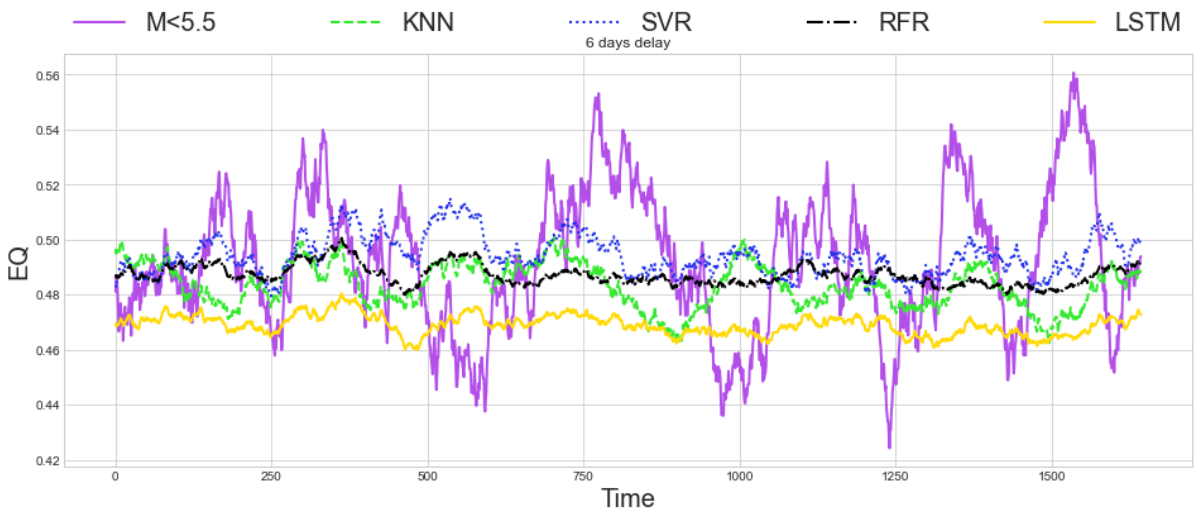


Figure 5.62 Deep zone earthquakes M<5.5: compare actual and predicted values, Six Days Delay

Table 5.30, Figure 5.63, and Figure 5.64 illustrate that in the seven-day delay part, the range of the accuracy and the positions of the prediction lines are the same as in the four-day delay part.

Table 5.30 Deep zone earthquakes M<5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.3048	0.2997	0.2972	0.2991
NRMSE by SD	1.017	0.9999	0.9915	0.998
NRMSE by mean	0.6314	0.6208	0.6156	0.6196
Mean absolute error				
MAE	0.2564	0.2524	0.2499	0.2526
NMAE by mean	0.531	0.5229	0.5177	0.5234

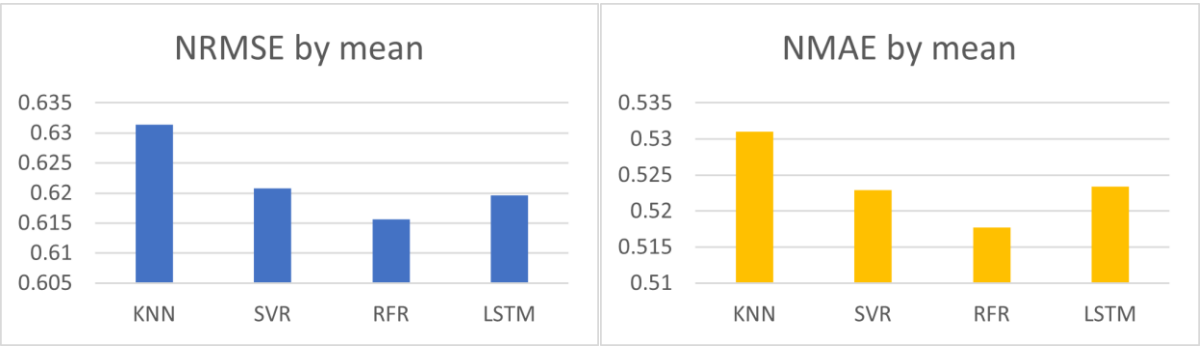


Figure 5.63 Errors: Deep zone earthquakes M<5.5, Seven Days Delay

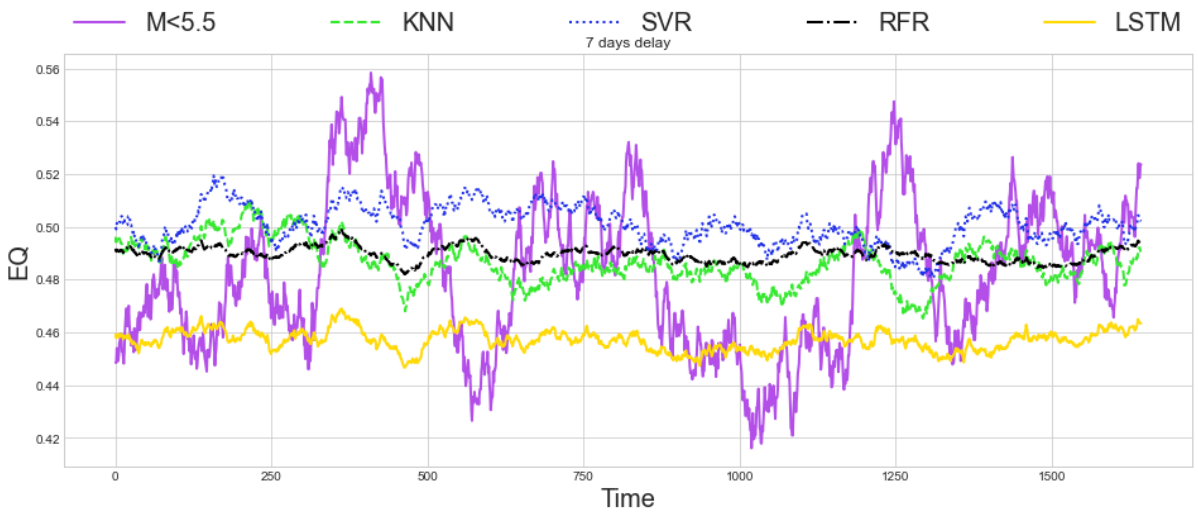


Figure 5.64 Deep zone earthquakes M<5.5: compare actual and predicted values, Seven Days Delay

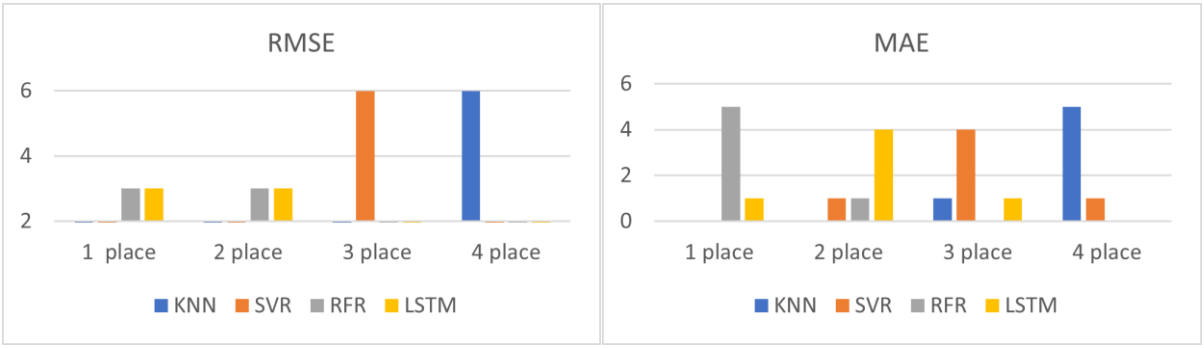


Figure 5.65 Deep zone earthquakes M<5.5, summarising results

In this part of the experiment, RFR and LSTM had the highest accuracy in both metrics, as shown in Figure 5.65. It was also noted that the values of the two highest accuracy results in both metrics are really close to each other. It was also found that the errors were quite near "1" or even higher. Furthermore, while the location of the LSTM prediction line changed, the location of the traditional ML algorithms' prediction lines remained roughly the same. It was found that the two-day delay part had the highest accuracy in terms of NRMSE, while the seven-day delay part had the highest accuracy in terms of NMAE. The data from the earthquakes also includes notable values that deviate more from the mean. As a result, RMSE values are also preferred here.

5.6 *Solar activity and Deep zone earthquakes with Richter magnitude equal to or greater than 5.5*

Table 5.31 through Table 5.36 summarise the results of each part of the deep zone earthquake (M≥5.5) experiment. The graphical interpretation of the above tables and the differences between actual and predicted earthquake values are displayed in Figure 5.66 through Figure 5.77.

Table 5.31 and Figure 5.66 show that in the two-day delay part, LSTM had the highest accuracy and KNN had the lowest accuracy in terms of NRMSE. In terms of NMAE, RFR and LSTM had the highest accuracy, while KNN and SVR had the lowest accuracy. Also, it was noted that the normalised error values in both metrics are too high, which causes the poor prediction. Figure 5.67 demonstrates that the prediction lines of the algorithms do not follow the actual line correctly. LSTM and RFR prediction lines are located around averages of the actual values. The KNN prediction line is also located near averages but with peaks, and the SVR prediction line is above the others.

Table 5.31 Deep zone earthquakes M≥5.5, Two Days Delay

Two Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2552	0.2505	0.2491	0.2486
NRMSE by SD	1.0261	1.0075	1.0015	0.9998
NRMSE by mean	3.6931	3.6259	3.6044	3.5985
Mean absolute error				
MAE	0.1334	0.1545	0.1308	0.131
NMAE by mean	1.9302	2.2366	1.8926	1.8958

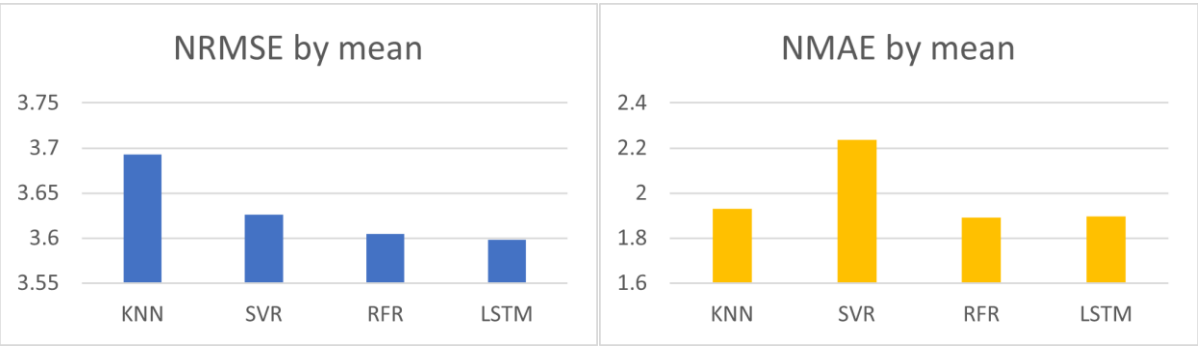


Figure 5.66 Errors: Deep zone earthquakes M≥5.5, Two Days Delay

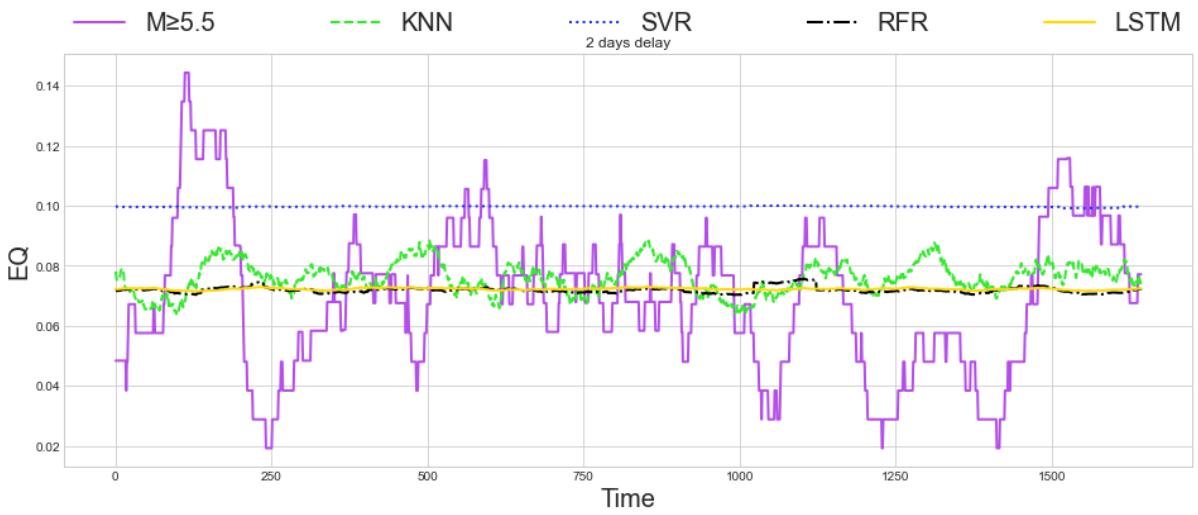


Figure 5.67 Deep zone earthquakes M≥5.5: compare actual and predicted values, Two Days Delay

Table 5.32 and Figure 5.68 illustrate that, in the three day-delay part, both metrics' error values are too high. In terms of NRMSE, SVR had the highest accuracy, and KNN had the lowest accuracy. In terms of NMAE, LSTM had the highest accuracy, while SVR had the lowest accuracy. Figure 5.69 shows that the prediction lines' locations in relation to one another were identical to those in the two-day part. However, the algorithms' prediction lines shifted with respect to the line representing actual values, which explains why the SVR achieved the highest accuracy in NRMSE.

Table 5.32 Deep zone earthquakes M≥5.5, Three Days Delay

Three Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2847	0.2769	0.2776	0.2781
NRMSE by SD	1.0292	1.0008	1.0032	1.0056
NRMSE by mean	3.261	3.171	3.1786	3.1851
MAE	0.1438	0.1689	0.1427	0.1397
NMAE by mean	1.647	1.9345	1.634	1.6002

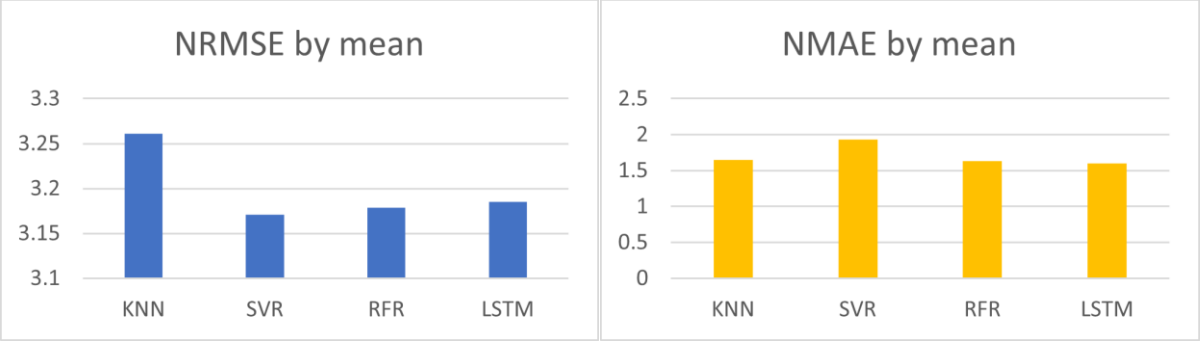


Figure 5.68 Errors: Deep zone earthquakes M≥5.5, Three Days Delay

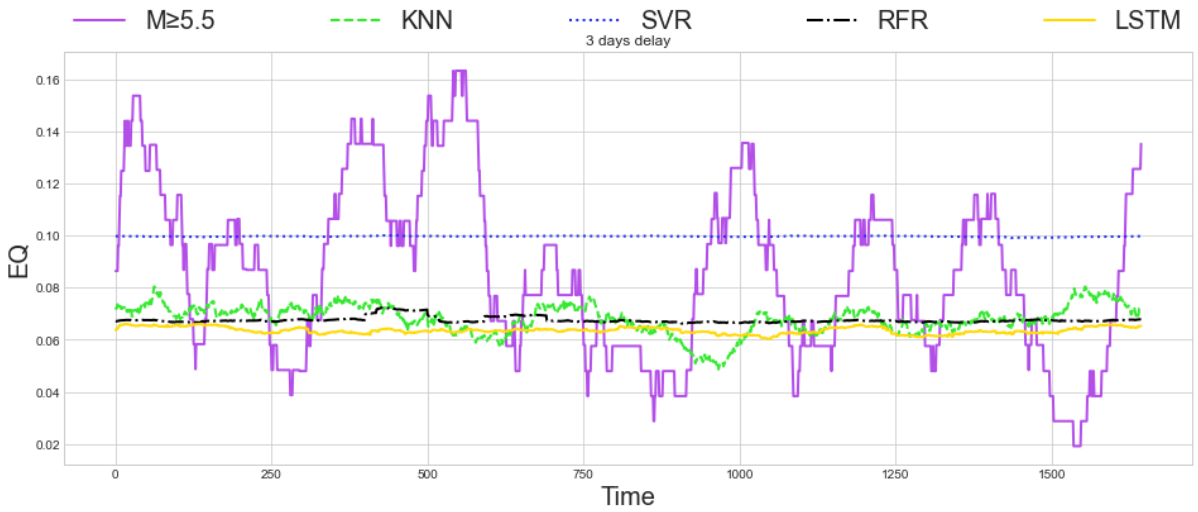


Figure 5.69 Deep zone earthquakes M≥5.5: compare actual and predicted values, Three Days Delay

Table 5.33, Figure 5.70, and Figure 5.71 demonstrate that the four-day delay part had the same range of results and prediction line location as the two-day delay part, with error values in both measures being too high. This also shows the poor accuracy of the prediction.

Table 5.33 Deep zone earthquakes M≥5.5, Four Days Delay

Four Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2658	0.2598	0.259	0.2588
NRMSE by SD	1.0277	1.0045	1.0014	1.0006
NRMSE by mean	3.5338	3.4542	3.4435	3.441
Mean absolute error				
MAE	0.135	0.1593	0.1351	0.1327
NMAE by mean	1.7946	2.1172	1.7956	1.7637

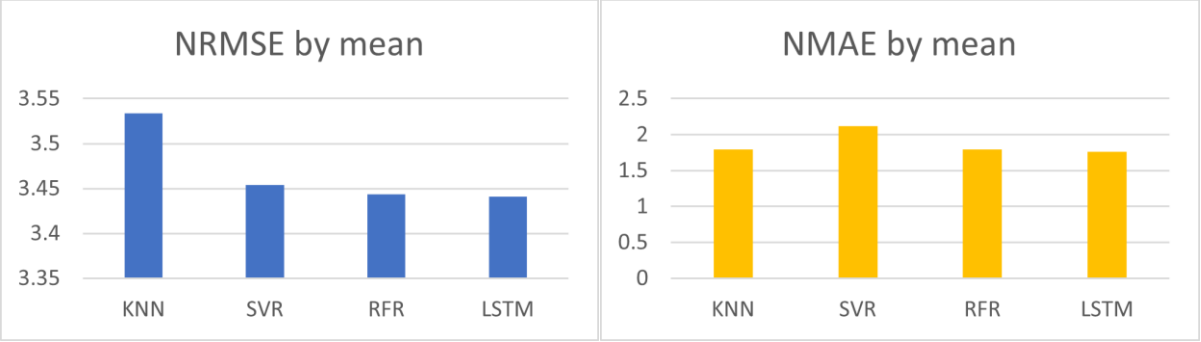


Figure 5.70 Errors: Deep zone earthquakes M≥5.5, Four Days Delay

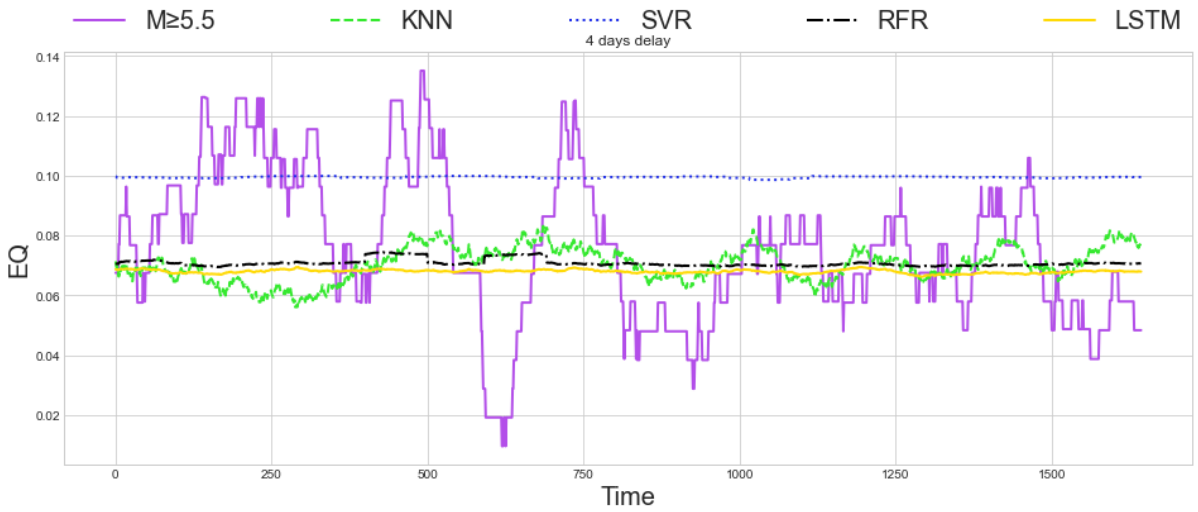


Figure 5.71 Deep zone earthquakes M≥5.5: compare actual and predicted values, Four Days Delay

The high value of errors and the same location of the prediction lines in the five-day delay part are shown in Table 5.34, Figure 5.72, and Figure 5.73

Table 5.34 Deep zone earthquakes M≥5.5, Five Days Delay

Five Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2543	0.2481	0.2463	0.2461
NRMSE by SD	1.0333	1.0082	1.001	1.0002
NRMSE by mean	3.7678	3.6765	3.6501	3.6472
Mean absolute error				
MAE	0.1282	0.1532	0.1295	0.1353
NMAE by mean	1.8992	2.2708	1.9196	2.0045

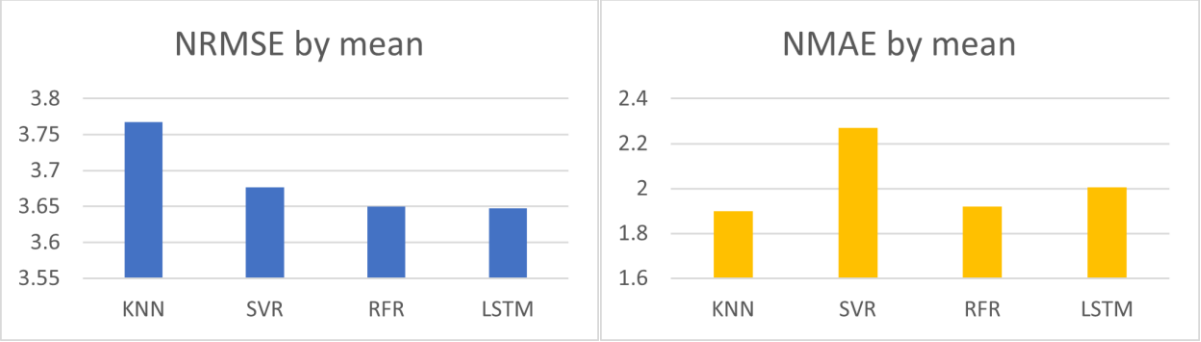


Figure 5.72 Errors: Deep zone earthquakes M≥5.5, Five Days Delay

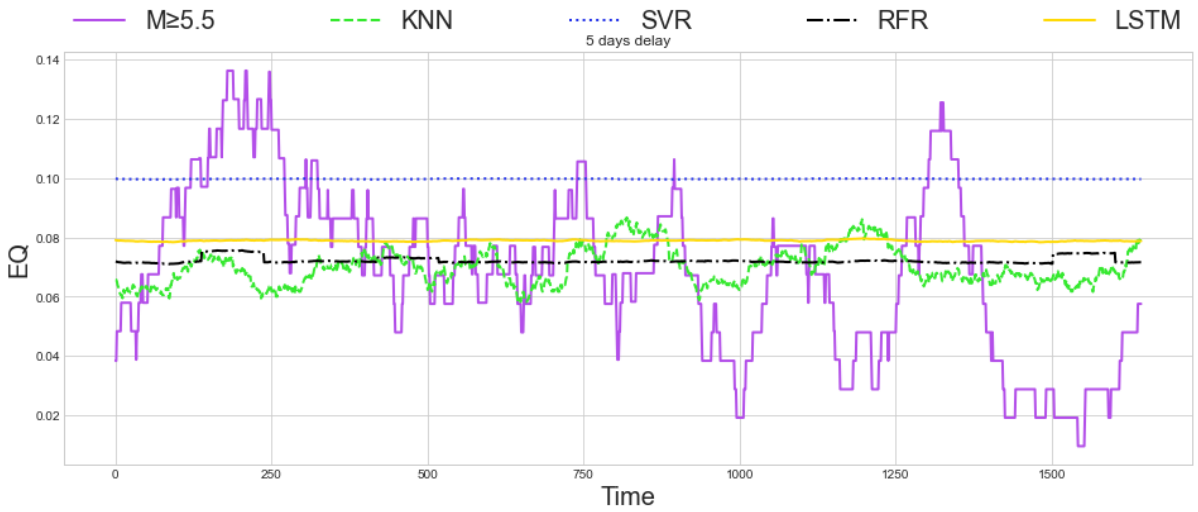


Figure 5.73 Deep zone earthquakes M≥5.5: compare actual and predicted values, Five Days Delay

Table 5.35, Figure 5.74, and Figure 5.75 show the range of results, and the prediction line locations in the six-day delay were similar to those in the five-day delay, with high normalised error values.

Table 5.35 Deep zone earthquakes M≥5.5, Six Days Delay

Six Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.2697	0.2621	0.2614	0.2611
NRMSE by SD	1.0326	1.0036	1.0011	0.9997
NRMSE by mean	3.509	3.4106	3.4021	3.3972
Mean absolute error				
MAE	0.1358	0.1606	0.1361	0.1404
NMAE by mean	1.7674	2.0904	1.7712	1.8272

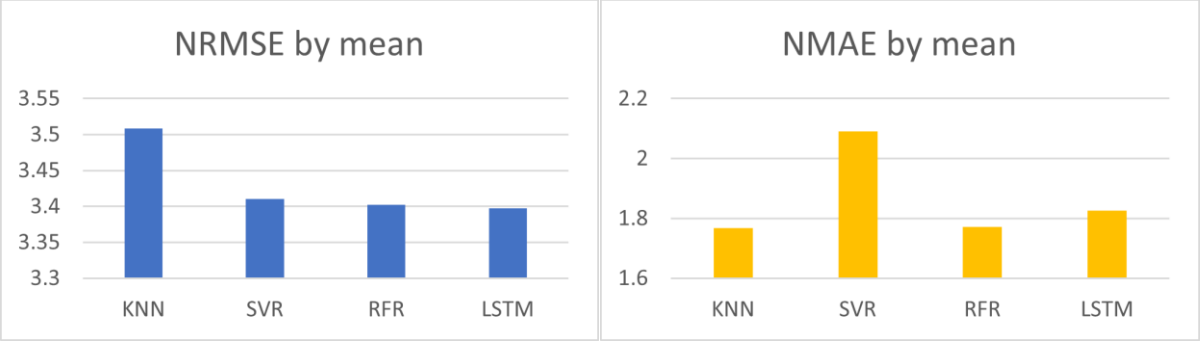


Figure 5.74 Errors: Deep zone earthquakes M≥5.5, Six Days Delay

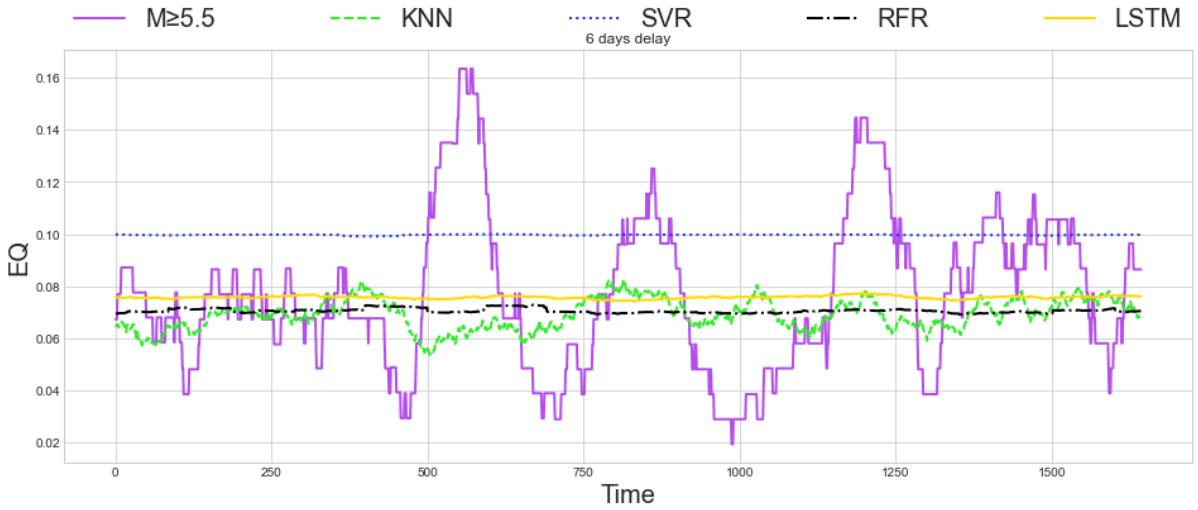


Figure 5.75 Deep zone earthquakes M≥5.5: compare actual and predicted values, Six Days Delay

The range of results and the locations of the prediction lines in the seven-day delay part are similar to those in the four-day delay part, with high normalised error values, as shown in Table 5.36, Figure 5.76, and Figure 5.77.

Table 5.36 Deep zone earthquakes M≥5.5, Seven Days Delay

Seven Days Delay				
Algorithm / Error	KNN	SVR	RFR	LSTM
Root mean squared error				
RMSE	0.256	0.2511	0.2493	0.2492
NRMSE by SD	1.0268	1.0073	1.0002	0.9998
NRMSE by mean	3.6794	3.6097	3.5842	3.5827
Mean absolute error				
MAE	0.131	0.1548	0.1312	0.1265
NMAE by mean	1.8835	2.2258	1.8853	1.819

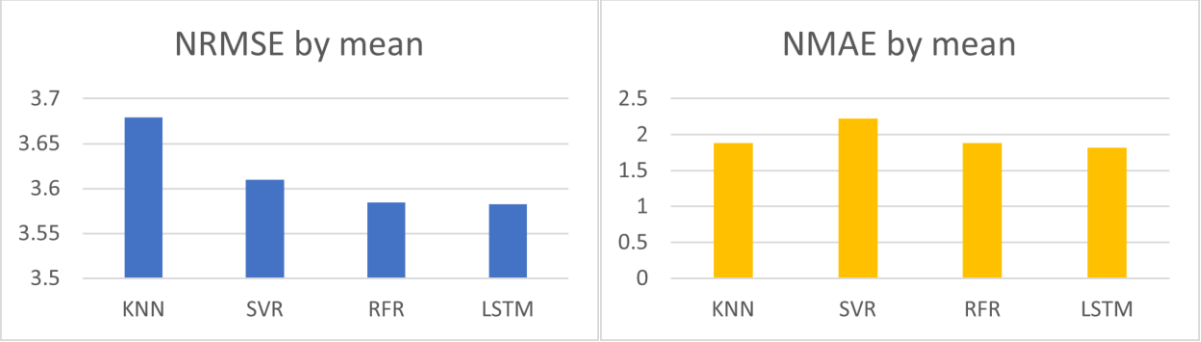


Figure 5.76 Errors: Deep zone earthquakes M≥5.5, Seven Days Delay

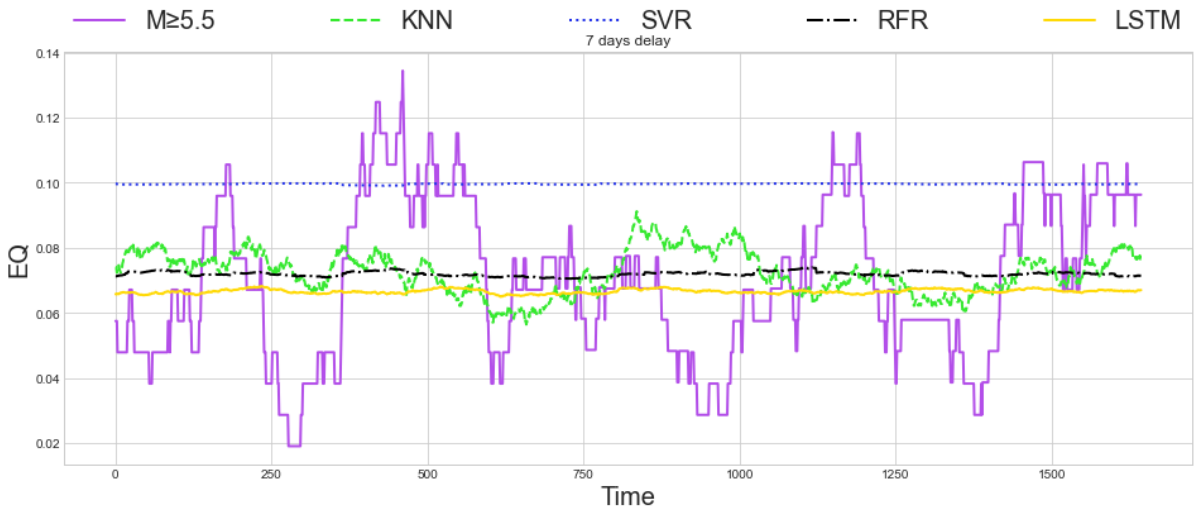


Figure 5.77 Deep zone earthquakes M≥5.5: compare actual and predicted values, Seven Days Delay

Similar to the previous section, with solar activity and the global earthquake, the results are quite similar. The same as with the previous section (Chapter 4), the ANOVA test was done using the RMSE normalised by standard deviation and MAE normalised by mean datasets. The results of the ANOVA and Shapiro-Wilk tests in the Appendix C. In the solar activity and shallow zone earthquakes, there is a difference in the error results in both metrics in the parts with a Richter magnitude less than 5.5. However, in the part where solar activity and shallow zone earthquakes with a Richter magnitude equal to or greater than 5.5 RMSE have different results, there was not able to implement the ANOVA for the MAE as the MAE results do not have a normal distribution. In the solar activity and intermediate zone earthquakes with a Richter magnitude less than 5.5, both error results had a difference, but in the second part, with a Richter magnitude greater than 5.5, not every error result had a normal distribution, which is why it was not possible to implement ANOVA. The solar activity and deep zone earthquakes with a Richter magnitude less than 5.5 have a difference in both error results. However, in the other part, with a Richter magnitude greater than 5.5, the RMSE data do not

have a normal distribution, and there is a difference in the MAE results. So, it can be concluded that there is a difference in the errors results.

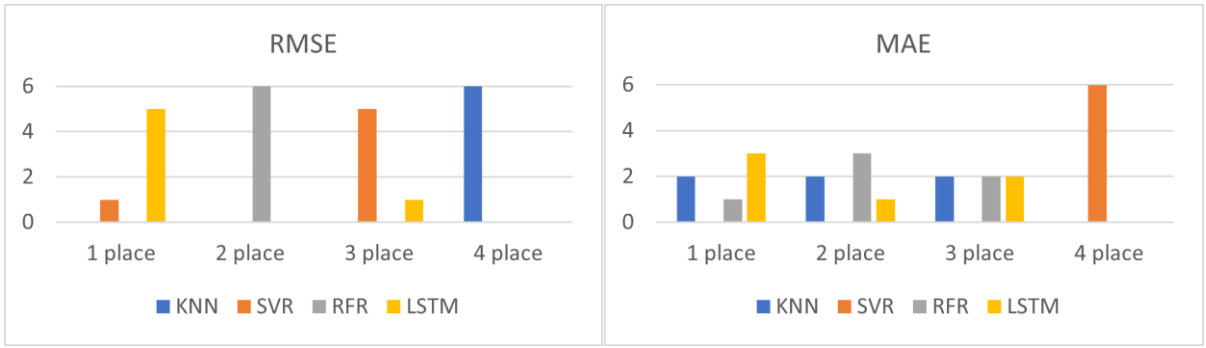


Figure 5.78 Deep zone earthquakes M≥5.5, summarising results

Figure 5.78 shows that the range variations between two metrics were rather large in this segment of the experiment. In terms of NRMSE, the first two positions were occupied by LSTM, whereas in terms of NMAE, the first three positions were occupied by LSTM, KNN, and RFR. However, the normalised error values are too high in and the accuracy of the prediction is poor. For NRMSE, it was discovered that the six-day delay part produced the highest accuracy, whereas for NMAE, the three-day delay part produced the highest accuracy. There are also notable values that are located far from the mean in the earthquake data. That is why RMSE values are preferred.

6 Chapter Six Evaluation, Conclusion, and Future work

6.1 Evaluation

The main research question of this study is: “How effective is machine learning in predicting of earthquakes based on solar activity?” To attempt this question, the following sub questions have been asked: "What characteristics and types of earthquakes and solar activity should be used?"

This study looked at the relationship between solar activity and earthquakes. Earthquakes occur globally, which is why the total number of earthquakes that occurred globally was selected. Based on the literature (Novikov *et al.*, 2017; Odintsov, Ivanov-Kholodnyi and Georgieva, 2007; Novikov *et al.*, 2020) and experiments conducted in this study, it was discovered that earthquake magnitude and depth are the most important characteristics of earthquakes that can be used in earthquake prediction. That is why, two different experimental designs were used in this study. In the first experimental design, earthquakes were divided into two groups: global earthquakes with a Richter magnitude of less than 5.5 and global earthquakes with a Richter magnitude of 5.5 or higher (refer to Chapter 4). The second experimental design used the earthquake's zones (the shallow zone, intermediate zone, and deep zone) to sort global earthquakes by depth in addition to their Richter magnitude (refer to Chapter 5).

With regards to solar activity, there are numerous solar activity events. Based on the literature and the results of previous seismological and space studies (Nishii, Qin and Kikuyama, 2020; Novikov *et al.*, 2017; Odintsov, Ivanov-Kholodnyi and Georgieva, 2007; Novikov *et al.*, 2020), the appropriate solar activity events, that can influence earthquakes were selected. Sunspot number, solar wind characteristics (solar wind speed, proton density, and proton temperature), and solar flares were chosen for solar activity, and used in both experimental designs (refer to Chapters 4 and 5). Thus, the first research sub question was answered.

The second research sub question is “How to evaluate the efficacy and effectiveness of the machine learning algorithms used in the study to ensure the efficacy of the analysis?” The analysis of the characteristics of the collected data for this study's experiment revealed that the data points' relationships are nonlinear (refer to Chapter 3.7) and there is a possibility of having a division by zero in the analysis (Chapter 2.5.2, *equation (2)*). Thus, the R^2 error and MAPE were discovered to be unsuitable data evaluation metrics. Following that, RMSE and MAE were used to assess the algorithm's efficacy and accuracy. To compare the results and determine their significance, the normalised values of the RMSE and MAE were used.

However, it was discovered during the experiment (refer to Chapter 4 and 5) that earthquakes have upper and lower peaks. In the earthquake data, the values that are far from the mean are important. Moreover, MAE is the average of absolute error values, whereas RMSE is the square root of the average of squared errors. That's why the RMSE gives large errors a lot of weight. Because RMSE gives more weight to observations that are farther from the mean, its values are more desirable in this situation.

The third research sub question is “Which machine learning algorithms should be chosen to answer the research question, and which one would give the highest accuracy among those chosen?” Two requirements were considered to determine which of the ML algorithms should be used. The first was to choose ML algorithms with different approaches (Chapter 2.5.3). The second was the non-linear relationship between earthquakes and solar activity (Chapter 3.7). As a result, four algorithms were selected: K-nearest neighbour, support vector regression, random forest regression, and long short-term memory neural network. KNN is one of the simplest and fastest Euclidean distance algorithms (Alpaydin, 2014). The SVR algorithm is based on the kernel (Smola and Schölkopf, 2004). RFR is an ensemble learning algorithm (Alpaydin, 2014). LSTM is a neural network algorithm (Hochreiter and Schmidhuber, 1997).

The results of the study are presented in Chapters 4 and 5. In the first segment of the experiment, the relationships between solar activity and the global earthquake were studied. In the first section of this experiment’s segment, the NRMSE analysis indicates that the LSTM model has a higher accuracy in predicting earthquakes than other models (Figure 4.13). In the second section of this experiment’s segment, the NRMSE analysis indicates that the LSTM and RFR models have a higher accuracy in predicting earthquakes than other models (Figure 4.26). Summarizing the results of all the segments of the experiment with shallow zone earthquakes (Figure 5.13 and Figure 5.26), intermediate zone earthquakes (Figure 5.39 and Figure 5.52), and deep zone earthquakes (Figure 5.65 and Figure 5.78), the general result is visualised in Figure 6.1.

According to the results summarised in Figure 6.1, LSTM outperforms other models in terms of EQ prediction accuracy, followed by RFR. However, the error values were close to each other. But the basic algorithm settings were used in the study, and to find out which algorithm will give the highest accuracy, more experiments with changing parameters need to be done. KNN is typically applied to related neighbours of data points. SVR considers each row to be a separate training sample and attempts to predict value based on the information gathered. RFR does not have an overfitting problem and uses an ensemble technique. LSTM attempts to analyse the entire dataset before predicting the next number. What is more, it should be mentioned that while traditional learning algorithms, during one segment of the experiment,

showed the constant location of the prediction lines in relation to the averages, minimums, and maximums of the actual values, in different time-delay parts of the experiment, the LSTM prediction lines changed their locations.

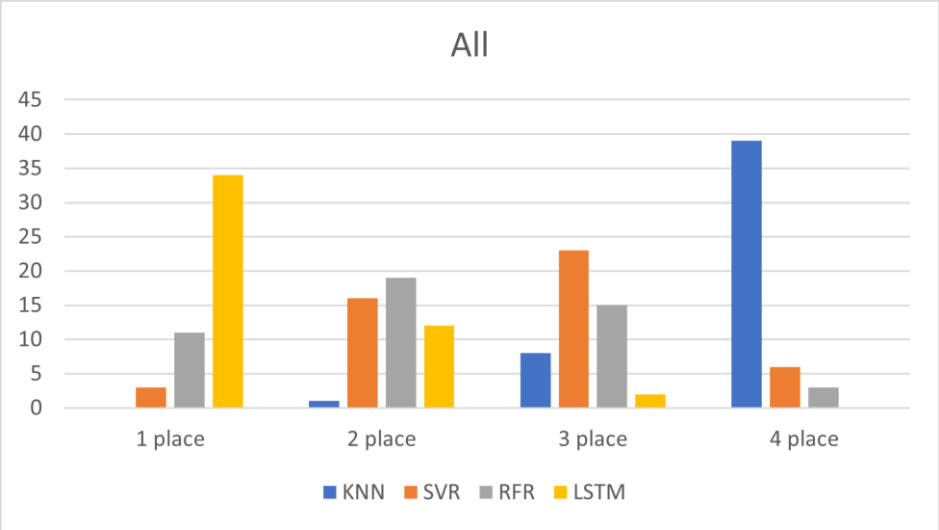


Figure 6.1 NRMSE: Summing up all the result ranges

The fourth research sub question is “Do solar activity events have the same impact on different types of earthquakes, and what other factors are important, such as time delay?” Based on the findings in all segments of the experiments (Chapters 4 and 5), solar activity and global earthquake (Table 4.1 through Table 4.12), shallow zone earthquakes (Table 5.1 through Table 5.12), intermediate zone earthquakes (Table 5.13 through Table 5.24), and deep zone earthquakes (Table 5.25 through Table 5.36), it can be discovered that both error values (used in the study) in the relationships between solar activity and earthquakes M<5 are less than both error values in the relationship between solar activity and earthquakes M≥5.

Considering these results, it can be concluded that solar activity has a stronger effect on earthquakes M<5.5 than on earthquakes M≥5. This conclusion is backed up by a study by Nishii, Qin and Kikuyama (2020), who found that SA has the greatest impact on earthquakes with a Richter magnitude less than 4. However, this conclusion differs from that of Odintsov, Ivanov-Kholodnyi and Georgieva (2007), who concluded that solar activity has a greater effect on earthquakes M>5.5. This can be explained by the fact that in their study, Odintsov, Ivanov-Kholodnyi and Georgieva (2007) used earthquake data with a minimum magnitude of five. They also focused primarily on high-speed solar wind. In contrast, Nishii, Qin, and Kikuyama (2020) had nine solar activity events and an earthquake M>3.

Based on the segments of the experiment where earthquakes were divided by their depth (Table 5.1 through Table 5.36), it was also discovered that as the depth of earthquakes increases, so do the error values. As a result, a suggestion can be made that solar activity has the greatest impact on earthquakes in the shallow zone. Moreover, in cases with greater depth and higher magnitude, RMSE and MAE values, normalised by mean, are bigger than "1". This leads to the assumption that solar activity may not have an influence on earthquakes with great depth and high magnitude. The study by Novikov *et al.* (2020) corroborated the findings of the study. Novikov *et al.* (2020) carried out an experiment that revealed that the electric current generated by solar activity can influence earthquakes. The higher the electrical conductivity, the higher the current density in the lower Earth crust levels, according to Novikov *et al.* (2020).

During the experiment, it was noted, that in the segments of the experiment with global earthquakes and shallow zone earthquakes, the lowest NRMSE values were in the three-day delay part. The other parts of the experiment with intermediate zone and deep zone earthquakes showed the lowest RMSE values in the different-day delay parts of the experiment. However, it should be noted that segments of the experiment had the highest error values (Table 5.13 through Table 5.36). This finding is supported by Sytinskii, (1973) who stated that earthquakes occur mainly 2-3 days after solar activity passes the central solar meridian. Also, Odintsov *et al.* (2006) and Odintsov, Ivanov-Kholodnyi and Georgieva (2007) found that solar wind with a high velocity had an influence on earthquakes.

6.2 Conclusion

The data were collected between 1996 and 2020, throughout the 23rd and 24th solar cycles. There were over 8,000 time-series data collections in the data. An 80/20 proportion for testing and training sets was employed in the study. Four ML algorithms were used in the study: KNN, SVR, RFR, and LSTM.

During the implementation of the machine learning models, several issues were encountered. There are two types of data in the study: earthquake data and solar activity data. The earthquake data were of rather high quality, while the solar activity data contained errors, missing values, and negative values. Also, both earthquake and solar activity data had a lot of outliers. This involved tasks such as data cleaning, normalization, and dimensionality reduction. There are many different types of machine learning models, each with its own strengths and weaknesses. Choosing the right model for a particular problem can be challenging and may require experimentation and domain expertise. Furthermore, machine learning models have hyperparameters that need to be tuned to achieve optimal performance to avoid underfitting or overfitting. To avoid these problems, previous study results were taken into account.

As can be seen from Figure 6.1, the LSTM model showed the highest accuracy in prediction compared with the other algorithms. Though there isn't a massive difference between the accuracy values of each algorithm. However, the accuracy of the prediction is far from being perfect. Moreover, the lowest error values, normalised by mean, are around 0.5, which shows that there are still a lot of issues to work out. For example, increase the size of a dataset by adding new variables or improving algorithms settings.

The only changeable attribute in KNN is the "K"-value. That is why, to improve the accuracy of KNN, the most obvious solution is to change the value of "K". However, refer to Figure 3.43 – Figure 3.50 where the increasing of the "K" value to greater than 17 does not significantly change the error values, so this will not improve the accuracy of the prediction. On the other hand, some studies (AL Kafaf, Kim and Lu, 2017; Wang, Xu and Zhao, 2021), to improve KNN, used a variation of KNN and proposed starting from a pre-processing stage, which classified the training dataset by categories. However, this method needs further study.

The kernel function is the main key to SVR. There are four kernels that are widely used in SVR (refer to Chapter 2.5.3). Besides changing kernels, the corresponding kernels' parameters also need to be set. Changing these kernels and comparing the error values can help find the most accurate solution.

In the case of RFR, changing the tree numbers is one of the first steps towards improving accuracy. Probst and Boulesteix, (2018) demonstrated the positive influence of increasing the number of trees on accuracy. Also, changing the parameters of RFR (such as the maximum depth of the tree) will change the accuracy of the prediction.

As for the LSTM model, changing network topology settings, adding more neurons to hidden layers (making LSTM wider), or making the LSTM model deeper by adding hidden layers (and trying a combination of these methods) will help find high prediction accuracy in LSTM. Moreover, changing the number of nodes and epochs will influence accuracy. Also, increasing the volume of the training dataset will change the accuracy of prediction, which is not the appropriate option for traditional ML algorithms. However, increasing the above parameters will require more training. So, it can be suggested that the LSTM model has more potential for the prediction of earthquakes based on solar activity events because it has more parameters that can be changed to improve the final accuracy of the prediction. However, it should be mentioned that the LSTM model is more expensive compared to the traditional ML algorithms used in this study.

6.3 Future Work

The study reported here is only a basic step towards earthquake prediction, but it has suggestions for future research. There are still a lot of issues to work out. The finding showed that an artificial neural network has more potential than traditional machine learning algorithms, even with basic settings. Although LSTM is far from being successful in predicting earthquakes, the finding indicates that the level of the prediction might be increased further by improving the algorithm and taking into account additional variables (solar activity and earthquakes) than those included in the study.

Here are a few ways to improve the performance of neural networks. Increasing the number of hidden layers, as a consequence, it appears that the more layers, the better the outcomes. However, it only requires a variable number of layers to be tested. Weights: first-time weights are set at random when training neural networks. Although weights updating occurs, neural networks can sometimes converge in local minima. Also, random weights do not function effectively when using a multi-layered design. It can be provided with the most appropriate starting weights. Also, the quantity of data used to train a neural network should be increased because the amount of data required varies greatly depending on the challenge. Further experiments need to be done using different neural network settings to compare the accuracy of the predictions.

However, neural networks are too time-consuming, energy-intensive, and require more expensive resources. That is why additional experiments should be done using traditional machine learning algorithms, changing their parameters, and adding additional algorithms. Additionally, ensemble learning must be used to determine whether the prediction's accuracy can be improved.

One of the goals here was to see if solar activity events influenced earthquakes. Analysing their relationships, using machine learning techniques for two solar cycles, showed the possibility of this impact. However, since earthquakes are currently unpredictable, the various machine learning models should be used to uncover relationships between solar activity and earthquakes.

Daily solar activity and earthquake data were examined for two solar cycles (24 years). The relationship between solar activity and earthquakes was positively demonstrated during this period. The next step is to test the methods over a longer period of time than two solar cycles (24 years). Furthermore, various solar activity events can be included in the research to improve the outcomes. The magnetic field, dynamic pressure, Earth's distance from the sun during the event, and its degree of tilt are only a few examples. Moreover, photos of solar activity events, such as solar wind or solar flares, might be used as an alternative.

However, it should be noted that not all solar activity events have the same effect on earthquakes; Nishii, Qin, and Kikuyama (2020) also suggested this. Finding the most effective solar activity events and the least effective solar activity events should be necessary at this point. This can be accomplished by reducing solar activity events one at a time and conducting an experiment. According to the study's findings and those of Nishii, Qin, and Kikuyama (2020), the impact of solar activity on earthquakes varies depending on the Richter magnitude of the earthquakes. Earthquakes should be ranged by an order of magnitude more frequently in future studies than was done in the current study.

According to Novikov et al. (2020), the electric current created by solar activity can impact earthquakes, which is also corroborated by the study's findings. In addition, the electrical conductivity of different Earth surfaces varies. As a result, one of the earthquake variables should be earthquake coordinates, ranged by tectonic plates.

References

ACE real-time solar wind | NOAA / NWS space weather prediction center (2021). Available at: https://www.swpc.noaa.gov/products/ace-real-time-solar-wind (Accessed: 7 June 2021).

Aguilar-Martinez, S. and Hsieh, W.W. (2009) 'Forecasts of tropical pacific sea surface temperatures by neural networks and support vector regression', *International Journal of Oceanography*, 2009, pp. 1–13. Available at: https://doi.org/10.1155/2009/167239.

Agresti, A., Franklin, C. A. and Klingenberg, B. (2018) *Statistics the art and science of learning from data*. 4th ed., global ed. Harlow [etc.]: Pearson.

AL Kafaf, D., Kim, D.-K. and Lu, L. (2017) 'B-kNN to Improve the Efficiency of kNN':, in *Proceedings of the 6th International Conference on Data Science, Technology and Applications*. *6th International Conference on Data Science, Technology and Applications*, Madrid, Spain: SCITEPRESS - Science and Technology Publications, pp. 126–132. Available at: https://doi.org/10.5220/0006393301260132.

Allan, R. (2006) 'Oscillations and teleconnections', in H.A. Bridgman and J.E. Oliver (eds) *The Global Climate System: Patterns, Processes, and Teleconnections*. Cambridge: Cambridge University Press, pp. 25–58. Available at: https://doi.org/10.1017/CBO9780511817984.003.

Alpaydin, E. (2014) *Introduction to machine learning*. Third edition. Cambridge, Massachusetts: The MIT Press (Adaptive computation and machine learning).

Amezquita Sanchez, J.P. *et al.* (2017) 'Detection of ULF Geomagnetic Anomalies Associated to Seismic Activity Using EMD Method and Fractal Dimension Theory', *IEEE Latin America Transactions*, 15(2), pp. 197–205. Available at: https://doi.org/10.1109/TLA.2017.7854612.

Andrade, A.T.C. *et al.* (2016) 'Outlier detection using k-means clustering and lightweight methods for Wireless Sensor Networks', in *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*. *IECON 2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society*, Florence, Italy: IEEE, pp. 4683–4688. Available at: https://doi.org/10.1109/IECON.2016.7794093.

Andrews, E.D. *et al.* (2004) 'Influence of ENSO on flood frequency along the California Coast', *Journal of Climate*, 17(2), pp. 337–348. Available at: https://doi.org/10.1175/1520-0442(2004)017<0337:IOEOFF>2.0.CO;2.

Asaly, S., Gottlieb, L.-A. and Reuveni, Y. (2021) 'Using support vector machine (SVM) and Ionospheric Total Electron Content (TEC) data for solar flare predictions', *IEEE Journal of*

Selected Topics in Applied Earth Observations and Remote Sensing, 14, pp. 1469–1481. Available at: https://doi.org/10.1109/JSTARS.2020.3044470.

Asim, K.M. *et al.* (2017) 'Earthquake magnitude prediction in Hindukush region using machine learning techniques', *Natural Hazards*, 85(1), pp. 471–486. Available at: https://doi.org/10.1007/s11069-016-2579-3.

Barnard, P.L. *et al.* (2015) 'Coastal vulnerability across the pacific dominated by el niño/southern oscillation', *Nature Geoscience*, 8(10), pp. 801–807. Available at: https://doi.org/10.1038/ngeo2539.

Benkedjouh, T. *et al.* (2015) 'Health assessment and life prediction of cutting tools based on support vector regression', *Journal of Intelligent Manufacturing*, 26(2), pp. 213–223. Available at: https://doi.org/10.1007/s10845-013-0774-6.

Bijan, N., Saied, P. and Somayeh, M. (2013) 'The effect of solar cycle's activities on earthquake: a conceptual idea for forecasting', *Disaster Advances*, 6, p. 8.

Bobra, M.G. and Couvidat, S. (2015) 'Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm', *The Astrophysical Journal*, 798(2), p. 135. Available at: https://doi.org/10.1088/0004-637X/798/2/135.

Breiman, L. (2001) 'Random forests', *Machine Learning*, 45(1), pp. 5–32. Available at: https://doi.org/10.1023/A:1010933404324.

Calvo, P. *et al.* (2020) 'Plants are intelligent, here's how', *Annals of Botany*, 125(1), pp. 11–28. Available at: https://doi.org/10.1093/aob/mcz155.

Camporeale, E., Carè, A. and Borovsky, J.E. (2017) 'Classification of solar wind with machine learning', *Journal of Geophysical Research: Space Physics*, 122(11). Available at: https://doi.org/10.1002/2017JA024383.

Chai, T. and Draxler, R.R. (2014) 'Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature', *Geoscientific Model Development*, 7(3), pp. 1247–1250. Available at: https://doi.org/10.5194/gmd-7-1247-2014.

Chambers, J.M. (ed.) (1983) *Graphical methods for data analysis*. Belmont, Calif. : Boston: Wadsworth International Group ; Duxbury Press (The Wadsworth statistics/probability series).

Chander, R. (1999) 'Can dams and reservoirs cause earthquakes?', *Resonance*, 4(11), pp. 4–13. Available at: https://doi.org/10.1007/BF02837323.

Composable cycles — cycler 0.11.0 documentation (2021). Available at: https://matplotlib.org/cycler/ (Accessed: 1 February 2023).

Dani, T. and Sulistiani, S. (2019) 'Prediction of maximum amplitude of solar cycle 25 using machine learning', *Journal of Physics: Conference Series*, 1231, p. 012022. Available at: https://doi.org/10.1088/1742-6596/1231/1/012022.

Daniell, J.E., Khazai, B. and Wenzel, F. (2013) 'Uncovering the 2010 Haiti earthquake death toll', *Natural Hazards and Earth System Sciences Discussions*, 1(3), pp. 1913–1942. Available at: https://doi.org/10.5194/nhessd-1-1913-2013.

Dao, D.V. *et al.* (2020) 'A sensitivity and robustness analysis of GPR and ANN for high-performance concrete compressive strength prediction using a monte carlo simulation', *Sustainability*, 12(3), p. 830. Available at: https://doi.org/10.3390/su12030830.

Das, S. *et al.* (2011) 'Machine learning techniques applied to prediction of residual strength of clay', *Open Geosciences*, 3(4). Available at: https://doi.org/10.2478/s13533-011-0043-1.

Dätwyler, C. *et al.* (2019) 'El Niño–Southern Oscillation variability, teleconnection changes and responses to large volcanic eruptions since AD 1000', *International Journal of Climatology*, 39(5), pp. 2711–2724. Available at: https://doi.org/10.1002/joc.5983.

Dewitte, B. *et al.* (2012) 'Change in El Niño flavours over 1958–2008: Implications for the long-term trend of the upwelling off Peru', *Deep Sea Research Part II: Topical Studies in Oceanography*, 77–80, pp. 143–156. Available at: https://doi.org/10.1016/j.dsr2.2012.04.011.

Donat, M.G. *et al.* (2010) 'Examination of wind storms over Central Europe with respect to circulation weather types and NAO phases', *International Journal of Climatology*, 30(9), pp. 1289–1300. Available at: https://doi.org/10.1002/joc.1982.

Dong, L. and Shan, J. (2013) 'A comprehensive review of earthquake-induced building damage detection with remote sensing techniques', *ISPRS Journal of Photogrammetry and Remote Sensing*, 84, pp. 85–99. Available at: https://doi.org/10.1016/j.isprsjprs.2013.06.011.

Draper, N.R. and Smith, H. (1998) *Applied regression analysis*. 3rd ed. New York: Wiley (Wiley series in probability and statistics. Texts and references section).

Dresch, A., Lacerda, D.P. and Antunes, J.A.V. (2015) 'Design science research', in Dresch, A., Lacerda, D. P., and Antunes Jr, J. A. V., *Design Science Research*. Cham: Springer

International Publishing, pp. 67–102. Available at: https://doi.org/10.1007/978-3-319-07374-3_4.

Dunjko, V. and Briegel, H.J. (2018) 'Machine learning & artificial intelligence in the quantum domain: a review of recent progress', *Reports on Progress in Physics*, 81(7), p. 074001. Available at: https://doi.org/10.1088/1361-6633/aab406.

Earthquakes (2021). Available at: https://www.usgs.gov/natural-hazards/earthquake-hazards/earthquakes (Accessed: 29 May 2021).

Earthquakes magnitude scale and classes (2021). Available at: http://www.geo.mtu.edu/UPSeis/magnitude.html (Accessed: 19 February 2021).

Feldman, U. (2005) 'On the sources of fast and slow solar wind', *Journal of Geophysical Research*, 110(A7), p. A07109. Available at: https://doi.org/10.1029/2004JA010918.

Feldstein, S.B. and Franzke, C.L.E. (2017) 'Atmospheric teleconnection patterns', in C.L.E. Franzke and T.J. OKane (eds) *Nonlinear and Stochastic Climate Dynamics*. Cambridge: Cambridge University Press, pp. 54–104. Available at: https://doi.org/10.1017/9781316339251.004.

Fidani, C. (2010) 'The earthquake lights (EQL) of the 6 April 2009 Aquila earthquake, in Central Italy', *Natural Hazards and Earth System Sciences*, 10(5), pp. 967–978. Available at: https://doi.org/10.5194/nhess-10-967-2010.

Frick, R.W. (1996) 'The appropriate use of null hypothesis testing.', *Psychological Methods*, 1(4), pp. 379–390. Available at: https://doi.org/10.1037/1082-989X.1.4.379.

Ghaedi, M. *et al.* (2016) 'Application of least squares support vector regression and linear multiple regression for modeling removal of methyl orange onto tin oxide nanoparticles loaded on activated carbon and activated carbon prepared from Pistacia atlantica wood', *Journal of Colloid and Interface Science*, 461, pp. 425–434. Available at: https://doi.org/10.1016/j.jcis.2015.09.024.

Ghahramani, Z. (2004) 'Unsupervised learning', in O. Bousquet, U. von Luxburg, and G. Rätsch (eds) *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures*. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 72–112. Available at: https://doi.org/10.1007/978-3-540-28650-9_5.

GOES-R Space Weather | NCEI, N.C. for E. (2021) *GOES-R Series | NCEI*. U.S. Department of Commerce. Available at: https://www.ngdc.noaa.gov/stp/satellite/goes-r.html (Accessed: 30 April 2021).

Gribbin, J. (1971) 'Relation of sunspot and earthquake activity', *Science*, 173(3996), pp. 558–558. Available at: https://doi.org/10.1126/science.173.3996.558.b.

Gruzdev, A. and Bezverkhnii, V. (2018) 'Analysis of relation of Central England surface air temperature to the 11-year solar cycle', in O.A. Romanovskii and G.G. Matvienko (eds) *24th International Symposium on Atmospheric and Ocean Optics: Atmospheric Physics*. *XXIV International Symposium, Atmospheric and Ocean Optics, Atmospheric Physics*, Tomsk, Russian Federation: SPIE, p. 34. Available at: https://doi.org/10.1117/12.2502904.

Hainzl, S. *et al.* (2006) 'Evidence for rainfall-triggered earthquake activity', *Geophysical Research Letters*, 33(19), p. L19303. Available at: https://doi.org/10.1029/2006GL027642.

Han, Y. (2004) 'Possible triggering of solar activity to big earthquakes (Ms>8) in faults with near west-east strike in China', *Science in China Series G*, 47(2), p. 173. Available at: https://doi.org/10.1360/03yw0103.

Hassan, D. *et al.* (2016) 'Sunspots and ENSO relationship using Markov method', *Journal of Atmospheric and Solar-Terrestrial Physics*, 137, pp. 53–57. Available at: https://doi.org/10.1016/j.jastp.2015.11.017.

Hathaway, D.H. (2015) 'The solar cycle', *Living Reviews in Solar Physics*, 12(1), p. 4. Available at: https://doi.org/10.1007/lrsp-2015-4.

Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780. Available at: https://doi.org/10.1162/neco.1997.9.8.1735.

Hoogendoorn, M. and Funk, B. (2018) 'Mathematical foundations for supervised learning', in M. Hoogendoorn and B. Funk (eds) *Machine Learning for the Quantified Self: On the Art of Learning from Sensory Data*. Cham: Springer International Publishing (Cognitive Systems Monographs), pp. 101–121. Available at: https://doi.org/10.1007/978-3-319-66308-1_6.

Hothorn, T., Hornik, K. and Zeileis, A. (2006) 'Unbiased recursive partitioning: a conditional inference framework', *Journal of Computational and Graphical Statistics*, 15(3), pp. 651–674. Available at: https://doi.org/10.1198/106186006X133933.

Hoyt, D.V. and Schatten, K.H. (1998) 'Group Sunspot Numbers: A New Solar Activity Reconstruction.', *Solar Physics*, 179(1), pp. 189–219. Available at: https://doi.org/10.1023/A:1005007527816.

Huzaimy, J.M. and Yumoto, K. (2011) 'Possible correlation between solar activity and global seismicity', in *Proceeding of the 2011 IEEE International Conference on Space Science and Communication (IconSpace)*. *Proceeding of the 2011 IEEE International Conference on Space Science and Communication (IconSpace)*, pp. 138–141. Available at: https://doi.org/10.1109/IConSpace.2011.6015869.

Ida, Y. *et al.* (2008) 'Detection of ULF electromagnetic emissions as a precursor to an earthquake in China with an improved polarization analysis', *Natural Hazards and Earth System Sciences*, 8(4), pp. 775–777. Available at: https://doi.org/10.5194/nhess-8-775-2008.

Introduction — statsmodels (2021). Available at: https://www.statsmodels.org/stable/index.html (Accessed: 18 March 2021).

Istiake Sunny, Md.A., Maswood, M.M.S. and Alharbi, A.G. (2020) 'Deep learning-based stock price prediction using lstm and bi-directional lstm model', in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*. *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, Giza, Egypt: IEEE, pp. 87–92. Available at: https://doi.org/10.1109/NILES50944.2020.9257950.

Jacobs, J.P. (2012) 'Bayesian support vector regression with automatic relevance determination kernel for modeling of antenna input characteristics', *IEEE Transactions on Antennas and Propagation*, 60(4), pp. 2114–2118. Available at: https://doi.org/10.1109/TAP.2012.2186252.

Jang, E. *et al.* (2018) 'Grasp2vec: learning object representations from self-supervised grasping', *arXiv:1811.06964 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1811.06964.

Jupyter Notebook Viewer (2021). Available at: https://nbviewer.jupyter.org/gist/talbertc-usgs/18f8901fc98f109f2b71156cf3ac81cd (Accessed: 15 May 2021).

Kanamori, H. and Brodsky, E.E. (2004) 'The physics of earthquakes', *Reports on Progress in Physics*, 67(8), pp. 1429–1496. Available at: https://doi.org/10.1088/0034-4885/67/8/R03.

Kaplan, A. and Haenlein, M. (2019) 'Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence', *Business Horizons*, 62(1), pp. 15–25. Available at: https://doi.org/10.1016/j.bushor.2018.08.004.

Keras: the Python deep learning API (2021). Available at: https://keras.io/ (Accessed: 29 August 2021).

Khan, T. *et al.* (2020) 'A machine learning approach for predicting the sunspot of solar cycle', in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, pp. 1–4. Available at: https://doi.org/10.1109/ICCCNT49239.2020.9225427.

Kim, J. *et al.* (2021) 'A cryptocurrency prediction model using lstm and gru algorithms', in *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*. *2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD)*, Zhuhai, China: IEEE, pp. 37–44. Available at: https://doi.org/10.1109/BCD51206.2021.9581397.

Kononenko, I. and Kukar, M. (2007) *Machine learning and data mining: introduction to principles and algorithms*. Chichester, UK: Horwood Publishing.

Korsós, M.B. *et al.* (2021) 'Testing and validating two morphological flare predictors by logistic regression machine learning', *Frontiers in Astronomy and Space Sciences*, 7, p. 571186. Available at: https://doi.org/10.3389/fspas.2020.571186.

Koskinen, H. *et al.* (2001) 'SPACE WEATHER EFFECTS CATALOGUE', *ESA Space Weather Study (ESWS)*, (2), pp. 11–21.

Kotsiantis, S.B. (2007) 'Supervised machine learning: A review of classification techniques.', in I.G. Maglogiannis (ed.) *Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in eHealth, HCI, information retrieval and pervasive technologies*. Amsterdam ; Washington, DC: IOS Press (Frontiers in artificial intelligence and applications, v. 160), pp. 3–24.

Kuncheva, L.I. (2004) *Combining pattern classifiers: methods and algorithms*. Hoboken, NJ: J. Wiley.

Kutner, M.H. (ed.) (2005) *Applied linear statistical models*. 5th ed. Boston: McGraw-Hill Irwin (The McGraw-Hill/Irwin series operations and decision sciences).

Laurenz, L., Lüdecke, H.-J. and Lüning, S. (2019) 'Influence of solar activity changes on European rainfall', *Journal of Atmospheric and Solar-Terrestrial Physics*, 185, pp. 29–42. Available at: https://doi.org/10.1016/j.jastp.2019.01.012.

Lesk, C., Rowhani, P. and Ramankutty, N. (2016) 'Influence of extreme weather disasters on global crop production', *Nature*, 529(7584), pp. 84–87. Available at: https://doi.org/10.1038/nature16467.

Li, Y. *et al.* (2009) 'Behavioral change related to Wenchuan devastating earthquake in mice', *Bioelectromagnetics*, 30(8), pp. 613–620. Available at: https://doi.org/10.1002/bem.20520.

Lipton, Z. C., Berkowitz, J. and Elkan, C. (2015) 'A critical review of recurrent neural networks for sequence learning'. arXiv. doi: 10.48550/arXiv.1506.00019.

Liu, C.M. *et al.* (2021) 'Comparison of machine learning approaches for tsunami forecasting from sparse observations', *Pure and Applied Geophysics*, 178(12), pp. 5129–5153. Available at: https://doi.org/10.1007/s00024-021-02841-9.

Liu, H. *et al.* (2019) 'Predicting solar flares using a long short-term memory network', *The Astrophysical Journal*, 877(2), p. 121. Available at: https://doi.org/10.3847/1538-4357/ab1b3c.

Logistic regression: from introductory to advanced concepts and applications - sage research methods (2010). Available at: https://methods.sagepub.com/book/logistic-regression-from-introductory-to-advanced-concepts-and-applications (Accessed: 18 April 2021).

Loutas, T.H., Roulias, D. and Georgoulas, G. (2013) 'Remaining useful life estimation in rolling bearings utilizing data-driven probabilistic e-support vectors regression', *IEEE Transactions on Reliability*, 62(4), pp. 821–832. Available at: https://doi.org/10.1109/TR.2013.2285318.

Love, J.J. and Thomas, J.N. (2013) 'Insignificant solar-terrestrial triggering of earthquakes: INSIGNIFICANT TRIGGERING', *Geophysical Research Letters*, 40(6), pp. 1165–1170. Available at: https://doi.org/10.1002/grl.50211.

Ma, H. *et al.* (2019) 'Solar activity modulates the El Niño-Southern Oscillation-induced precipitation anomalies over southern China in early spring', *International Journal of Climatology*, n/a(n/a). Available at: https://doi.org/10.1002/joc.7214.

Mallouhy, R. *et al.* (2019) 'Major earthquake event prediction using various machine learning algorithms', in *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*. *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, Paris, France: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/ICT-DM47966.2019.9032983.

Mangalathu, S. *et al.* (2020) 'Classifying earthquake damage to buildings using machine learning', *Earthquake Spectra*, 36(1), pp. 183–208. Available at: https://doi.org/10.1177/8755293019878137.

Marchitelli, V. *et al.* (2020) 'On the correlation between solar activity and large earthquakes worldwide', *Scientific Reports*, 10(1), p. 11495. Available at: https://doi.org/10.1038/s41598-020-67860-3.

Masci, F. and Thomas, J.N. (2015) 'Are there new findings in the search for ULF magnetic precursors to earthquakes?: ON ULF MAGNETIC EARTHQUAKES PRECURSORS', *Journal of Geophysical Research: Space Physics*, 120(12), p. 10,289-10,304. Available at: https://doi.org/10.1002/2015JA021336.

Math — mathematical functions (2021) *Python documentation*. Available at: https://docs.python.org/3/library/math.html (Accessed: 1 February 2021).

Matplotlib: Python plotting — Matplotlib 3.4.2 documentation (2021). Available at: https://matplotlib.org/ (Accessed: 18 March 2021).

McNutt, S.R. and Roman, D.C. (2015) 'Volcanic seismicity', in *The Encyclopedia of Volcanoes*. Elsevier, pp. 1011–1034. Available at: https://doi.org/10.1016/B978-0-12-385938-9.00059-6.

McPhaden, M.J., Zebiak, S.E. and Glantz, M.H. (2006) 'ENSO as an integrating concept in earth science', *Science*, 314(5806), pp. 1740–1745. Available at: https://doi.org/10.1126/science.1132588.

Merle, O. (2011) 'A simple continental rift classification', *Tectonophysics*, 513(1–4), pp. 88–95. Available at: https://doi.org/10.1016/j.tecto.2011.10.004.

Métivier, L. *et al.* (2009) 'Evidence of earthquake triggering by the solid earth tides', *Earth and Planetary Science Letters*, 278(3), pp. 370–375. Available at: https://doi.org/10.1016/j.epsl.2008.12.024.

Meyer-Vernet, N. (2012) *Basics of the solar wind*. Cambridge: Cambridge University Press. Available at: https://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9780511535765 (Accessed: 7 June 2021).

Mohamed, A.E. (2017) 'Comparative Study of Four Supervised Machine Learning Techniques for Classification', 7(2), p. 14.

Mohamed, M.A. and El-Mahdy, M.E.-S. (2021) 'Impact of sunspot activity on the rainfall patterns over Eastern Africa: a case study of Sudan and South Sudan', *Journal of Water and Climate Change*, p. jwc2021312. Available at: https://doi.org/10.2166/wcc.2021.312.

Mohd Razali, N. and Yap, B. (2011) 'Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests', *J. Stat. Model. Analytics*, 2.

Moldwin, M. (2008) *An introduction to space weather*. Cambridge: Cambridge University Press. Available at: https://doi.org/10.1017/CBO9780511801365.

Muhamedyev, R. (2015) 'Machine learning methods: An overview', *CMNT*, 19, pp. 14–29.

Muis, S. *et al.* (2018) 'Influence of el niño-southern oscillation on global coastal flooding', *Earth's Future*, 6(9), pp. 1311–1322. Available at: https://doi.org/10.1029/2018EF000909.

Müller, A.C. and Guido, S. (2016) *Introduction to machine learning with Python: a guide for data scientists*. 1st edn. Sebastopol, CA: O'Reilly Media, Inc.

Murwantara, I.M., Yugopuspito, P. and Hermawan, R. (2020) 'Comparison of machine learning performance for earthquake prediction in Indonesia using 30 years historical data', *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 18(3), p. 1331. Available at: https://doi.org/10.12928/telkomnika.v18i3.14756.

NASA/Marshall solar physics (2014). Available at: https://solarscience.msfc.nasa.gov/ (Accessed: 2 June 2021).

National Centers for Environmental Information(NCEI) (2021) *National Centers for Environmental Information (NCEI)*. Available at: http://www.ncei.noaa.gov/node (Accessed: 6 March 2021).

National Flood Relief Commission (1933) *Report Of The National Flood Relief Commission 1931 1932*. The Comacrib Press. Available at: http://archive.org/details/reportofthenatio032042mbp (Accessed: 2 February 2021).

National Geophysical Data Center / World Data Service (NGDC/WDS) (1972) 'NCEI/WDS Global Significant Earthquake Database'. NOAA National Centers for Environmental Information. Available at: https://doi.org/10.7289/V5TD9V7K.

Nguyen, Q.H. *et al.* (2021) 'Influence of data splitting on performance of machine learning models in prediction of shear strength of soil', *Mathematical Problems in Engineering*. Edited by Y.-S. Shen, 2021, pp. 1–15. Available at: https://doi.org/10.1155/2021/4832864.

Nishii, R., Qin, P. and Kikuyama, R. (2020) 'Solar activity is one of triggers of earthquakes with magnitudes less than 6', in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Waikoloa, HI, USA: IEEE, pp. 377–380. Available at: https://doi.org/10.1109/IGARSS39084.2020.9323381.

Nishimura, T. (2017) 'Triggering of volcanic eruptions by large earthquakes: Triggering of Volcanic Eruptions', *Geophysical Research Letters*, 44(15), pp. 7750–7756. Available at: https://doi.org/10.1002/2017GL074579.

North Atlantic Oscillation (NAO) | Teleconnections | National Centers for Environmental Information (NCEI) (no date). Available at: https://www.ncdc.noaa.gov/teleconnections/nao/ (Accessed: 1 February 2021).

Novianty, A. *et al.* (2019) 'Tsunami potential identification based on seismic features using knn algorithm', in *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC)*. *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC)*, Melaka, Malaysia: IEEE, pp. 155–160. Available at: https://doi.org/10.1109/ICSPC47137.2019.9068095.

Novikov, V. *et al.* (2020) 'Space weather and earthquakes: possible triggering of seismic activity by strong solar flares', *Annals of Geophysics*, 63(5), p. 13. Available at: https://doi.org/10.4401/ag-7975.

Novikov, V.A. *et al.* (2017) 'Electrical triggering of earthquakes: results of laboratory experiments at spring-block models', *Earthquake Science*, 30(4), pp. 167–172. Available at: https://doi.org/10.1007/s11589-017-0181-8.

Numpy and Scipy Documentation — Numpy and Scipy documentation (2021). Available at: https://docs.scipy.org/doc/ (Accessed: 18 March 2021).

Odintsov, S. *et al.* (2006) 'Long-period trends in global seismic and geomagnetic activity and their relation to solar activity', *Physics and Chemistry of the Earth, Parts A/B/C*, 31(1–3), pp. 88–93. Available at: https://doi.org/10.1016/j.pce.2005.03.004.

Odintsov, S.D., Ivanov-Kholodnyi, G.S. and Georgieva, K. (2007) 'Solar activity and global seismicity of the earth', *Bulletin of the Russian Academy of Sciences: Physics*, 71(4), pp. 593–595. Available at: https://doi.org/10.3103/S1062873807040466.

OFDA/CRED International Disaster Data (2021) *Our World in Data*. Available at: https://ourworldindata.org/ofdacred-international-disaster-data (Accessed: 4 February 2021).

Orihara, Y., Kamogawa, M. and Nagao, T. (2015) 'Preseismic Changes of the Level and Temperature of Confined Groundwater related to the 2011 Tohoku Earthquake', *Scientific Reports*, 4(1), p. 6907. Available at: https://doi.org/10.1038/srep06907.

Oxford English Dictionary (2021). Available at: https://oed.com/ (Accessed: 19 February 2021).

Pal, M. *et al.* (2020) 'Long-lead prediction of enso modoki index using machine learning algorithms', *Scientific Reports*, 10(1), p. 365. Available at: https://doi.org/10.1038/s41598-019-57183-3.

Pandas - python data analysis library (2021). Available at: https://pandas.pydata.org/ (Accessed: 2 February 2021).

Pararas-Carayannis, G. and Zoll, P. (2017) 'Incipient evaluation of temporal El Nino and other climatic anomalies in triggering earthquakes and tsunamis – case study: The earthquake and tsunami of 16th April 2016 in Ecuador', *Science of Tsunami Hazards*, 36, pp. 262–291.

Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in Python', *MACHINE LEARNING IN PYTHON*, p. 6.

Pham, B.T. *et al.* (2020) 'A novel hybrid soft computing model using Random Forest and particle swarm optimization for estimation of undrained shear strength of soil', *Sustainability*, 12(6), p. 2218. Available at: https://doi.org/10.3390/su12062218.

Plate Tectonics Map - Plate Boundary Map (2021). Available at: https://geology.com/plate-tectonics.shtml (Accessed: 15 May 2021).

Potter, S. (2020) *Solar Cycle 25 Is Here. NASA, NOAA Scientists Explain What That Means*, *NASA*. Available at: http://www.nasa.gov/press-release/solar-cycle-25-is-here-nasa-noaa-scientists-explain-what-that-means (Accessed: 27 February 2021).

Pribadi, S. *et al.* (2013) 'Characteristics of Earthquake-Generated Tsunamis in Indonesia Based on Source Parameter Analysis', *Journal of Mathematical and Fundamental Sciences*, 45(2), pp. 189–207. Available at: https://doi.org/10.5614/j.math.fund.sci.2013.45.2.8.

Priest, E. (2014) *Magnetohydrodynamics of the sun*. Cambridge: Cambridge University Press. Available at: https://doi.org/10.1017/CBO9781139020732.

Probst, P. and Boulesteix, A.-L. (2018) 'To Tune or Not to Tune the Number of Trees in Random Forest', *Journal of Machine Learning Research*, (18 (2018) 1–1), pp. 6673–6690.

Rácz, A., Bajusz, D. and Héberger, K. (2021) 'Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification', *Molecules*, 26(4), p. 1111. Available at: https://doi.org/10.3390/molecules26041111.

Raju, V.N.G. *et al.* (2020) 'Study the influence of normalization/transformation process on the accuracy of supervised classification', in *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India: IEEE, pp. 729–735. Available at: https://doi.org/10.1109/ICSSIT48917.2020.9214160.

Rasouli, K., Hsieh, W.W. and Cannon, A.J. (2012) 'Daily streamflow forecasting by machine learning methods with weather and climate inputs', *Journal of Hydrology*, 414–415, pp. 284–293. Available at: https://doi.org/10.1016/j.jhydrol.2011.10.039.

Razali, N., Ismail, S. and Mustapha, A. (2020) 'Machine learning approach for flood risks prediction', *IAES International Journal of Artificial Intelligence (IJ-AI)*, 9(1), p. 73. Available at: https://doi.org/10.11591/ijai.v9.i1.pp73-80.

Reddy, G.T. *et al.* (2020) 'Analysis of dimensionality reduction techniques on big data', *IEEE Access*, 8, pp. 54776–54788. Available at: https://doi.org/10.1109/ACCESS.2020.2980942.

Redmayne, D.W. (1988) 'Mining induced seismicity in UK coalfields identified on the BGS National Seismograph Network', *Geological Society, London, Engineering Geology Special Publications*, 5(1), pp. 405–413. Available at: https://doi.org/10.1144/GSL.ENG.1988.005.01.45.

Reinse, D., Gantz, J. and Rydning, J. (2018) *The Digitization of the World From Edge to Core*. #US44413318. Seagate IDC. Available at: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf (Accessed: 19 March 2021).

Rodriguez-Galiano, V.F. *et al.* (2012) 'An assessment of the effectiveness of a random forest classifier for land-cover classification', *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, pp. 93–104. Available at: https://doi.org/10.1016/j.isprsjprs.2011.11.002.

Samuel, A.L. (1959) 'Some studies in machine learning using the game of checkers', *IBM Journal of Research and Development*, 3(3), pp. 210–229. Available at: https://doi.org/10.1147/rd.33.0210.

Schorlemmer, D. *et al.* (2018) 'The collaboratory for the study of earthquake predictability: achievements and priorities', *Seismological Research Letters*, 89(4), pp. 1305–1313. Available at: https://doi.org/10.1785/0220180053.

seaborn: statistical data visualization — seaborn 0.11.1 documentation (2021). Available at: https://seaborn.pydata.org/ (Accessed: 1 February 2021).

Shabnam, N. (2014) 'Natural disasters and economic growth: a review', *International Journal of Disaster Risk Science*, 5(2), pp. 157–163. Available at: https://doi.org/10.1007/s13753-014-0022-5.

Shcherbakov, M.V. *et al.* (2013) 'A Survey of Forecast Error Measures', p. 7.

SILSO | World Data Center for the production, preservation and dissemination of the international sunspot number (2021). Available at: https://wwwbis.sidc.be/silso/home (Accessed: 30 April 2021).

Singh, S. *et al.* (2010) 'Radon monitoring in soil gas and ground water for earthquake prediction studies in north west himalayas, india', *Terrestrial, Atmospheric and Oceanic Sciences*, 21(4), p. 685. Available at: https://doi.org/10.3319/TAO.2009.07.17.01(TT).

de Siqueira Santos, S. *et al.* (2014) 'A comparative study of statistical methods used to identify dependencies between gene expression signals', *Briefings in Bioinformatics*, 15(6), pp. 906–918. Available at: https://doi.org/10.1093/bib/bbt051.

Smola, A.J. and Schölkopf, B. (2004) 'A tutorial on support vector regression', *Statistics and Computing*, 14(3), pp. 199–222. Available at: https://doi.org/10.1023/B:STCO.0000035301.49549.88.

Solar Flare Data | NCEI, N.G.D. (2021) *Solar Terrestrial Physics*. U.S. Department of Commerce. Available at: https://www.ngdc.noaa.gov/stp/solar/solarflares.html (Accessed: 2 February 2021).

SPDF - OMNIWeb Service (2021). Available at: https://omniweb.gsfc.nasa.gov/ (Accessed: 30 April 2021).

Spiegelhalter, D. J. (2019) *The art of statistics: learning from data*. [London] UK: Pelican, an imprint of Penguin Books.

Spiess, A.-N. and Neumeyer, N. (2010) 'An evaluation of R2 as an inadequate measure for nonlinear models in pharmacological and biochemical research: a Monte Carlo approach', *BMC Pharmacology*, 10(1), p. 6. Available at: https://doi.org/10.1186/1471-2210-10-6.

Statistical functions (Scipy. Stats) — SciPy v1.7.1 Manual (2021). Available at: https://docs.scipy.org/doc/scipy/reference/stats.html (Accessed: 16 June 2021).

Storchak, D.A. *et al.* (2013) 'Public release of the isc-gem global instrumental earthquake catalogue(1900-2009)', *Seismological Research Letters*, 84(5), pp. 810–815. Available at: https://doi.org/10.1785/0220130034.

Sunspot number Data | NCEI, N.G.D. (2021) *Solar-Terrestrial Physics and World Data Center for STP*. U.S. Department of Commerce. Available at: https://www.ngdc.noaa.gov/stp/solar/ssndata.html (Accessed: 1 February 2021).

Sunspot-numbers - monthly (2021). Available at: https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-indices/sunspot-numbers/american/lists/list_aavso-arssn_monthly.txt (Accessed: 15 February 2021).

Supervised learning — scikit-learn 0.24.2 documentation (2021). Available at: https://scikit-learn.org/stable/supervised_learning.html (Accessed: 2 February 2021).

Sutton, R.S. and Barto, A.G. (2018) *Reinforcement learning: an introduction*. Second edition. Cambridge, Massachusetts: The MIT Press (Adaptive computation and machine learning series).

Sytinskii, A.D. (1973) 'Relation between seismic activity of the earth and solar activity.', *Uspekhi Fizicheskih Nauk*, 111(10), p. 367. Available at: https://doi.org/10.3367/UFNr.0111.197310i.0367.

Thompson, D.W.J. and Wallace, J.M. (1998) 'The Arctic oscillation signature in the wintertime geopotential height and temperature fields', *Geophysical Research Letters*, 25(9), pp. 1297–1300. Available at: https://doi.org/10.1029/98GL00950.

Thorndike, E.L. (2000) *Animal intelligence: experimental studies*. New Brunswick, N.J: Transaction Publishers.

Tian, D., Yao, J. and Wen, L. (2018) 'Collapse and earthquake swarm after North Korea's 3 September 2017 nuclear test', *Geophysical Research Letters*, 45(9), pp. 3976–3983. Available at: https://doi.org/10.1029/2018GL077649.

Trappenberg, T.P. (2020) *Fundamentals of machine learning*. First edition. Oxford, United Kingdom: Oxford University Press.

Unsupervised learning — scikit-learn 0.24.2 documentation (2021). Available at: https://scikit-learn.org/stable/unsupervised_learning.html (Accessed: 2 February 2021).

Usgs earthquake hazards program (2021). Available at: https://earthquake.usgs.gov/ (Accessed: 7 April 2021).

Verdhan, V. and Kling, E.Y. (2020) *Supervised learning with Python: concepts and practical implementation using Python*. New York, NY: Apress (For professionals by professionals).

Wang, H., Xu, P. and Zhao, J. (2021) 'Improved knn algorithm based on preprocessing of center in smart cities', *Complexity*. Edited by Z. Lv, 2021, pp. 1–10. Available at: https://doi.org/10.1155/2021/5524388.

Wang, Q. *et al.* (2019) 'Improved multi-agent reinforcement learning for path planning-based crowd simulation', *IEEE Access*, 7, pp. 73841–73855. Available at: https://doi.org/10.1109/ACCESS.2019.2920913.

Wang, S. *et al.* (2015) 'Combined effects of the pacific decadal oscillation and el niño-southern oscillation on global land dry–wet changes', *Scientific Reports*, 4(1), p. 6651. Available at: https://doi.org/10.1038/srep06651.

Willmott, C. and Matsuura, K. (2005) 'Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance', *Climate Research*, 30, pp. 79–82. Available at: https://doi.org/10.3354/cr030079.

Wirasinghe, S.C. *et al.* (2013) 'Preliminary Analysis and Classification of Natural Disasters'. Available at: https://doi.org/10.13140/RG.2.1.4283.5041.

Witten, I.H. and Frank, E. (2017) *Data mining practical machine learning tools and techniques*. San Diego, CA, USA: Elsevier Science & Technology Books.

Wolf, Rod. (1853) 'On the periodic return of the minimum of sun-sport; the agreement between those periods and the variations of magnetic declination', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 5(29), pp. 67–67. Available at: https://doi.org/10.1080/14786445308646906.

Wood, B. *et al.* (2009) 'The three-dimensional morphology of a corotating interaction region in the inner heliosphere', *The Astrophysical Journal Letters*, 708, p. L89. Available at: https://doi.org/10.1088/2041-8205/708/2/L89.

Yamauchi, H. *et al.* (2017) 'Statistical Evaluations of Variations in Dairy Cows' Milk Yields as a Precursor of Earthquakes', *Animals*, 7(12), p. 19. Available at: https://doi.org/10.3390/ani7030019.

Yuan, S. *et al.* (2019) 'Prediction of north atlantic oscillation index with convolutional LSTM based on ensemble empirical mode decomposition', *Atmosphere*, 10(5), p. 252. Available at: https://doi.org/10.3390/atmos10050252.

Zhang, Q. *et al.* (2017) 'Prediction of sea surface temperature using long short-term memory', *IEEE Geoscience and Remote Sensing Letters*, 14(10), pp. 1745–1749. Available at: https://doi.org/10.1109/LGRS.2017.2733548.

Zou, K.H., Tuncali, K. and Silverman, S.G. (2003) 'Correlation and simple linear regression', *Radiology*, 227(3), pp. 617–628. Available at: https://doi.org/10.1148/radiol.2273011499.

Appendix A Codes

```
# Source: https://www.python-graph-gallery.com/313-bubble-map-with-folium
map_world  = folium.Map(location=[20,0], tiles="cartodbpositron", zoom_start=2)

colours=['#DCDF34','#57DF2A', '#3495DF', '#7D26DF', '#DF1DA5', '#DF1021']
value_mag_EQ = [2.5, 5.4, 6.0, 6.9, 7.9, 8.0]

def setMagnitudeColour(mag, value_mag_EQ, colours):
    for i in range(len(value_mag_EQ)):
        if mag < value_mag_EQ[i]:
            return colours[i]

for i in range(len(df_earthquake)):
    folium.Circle(location=[df_earthquake.iloc[i]["Latitude"], df_earthquake.iloc[i]["Longitude"]],radius=1000,
                  color=setMagnitudeColour(df_earthquake.iloc[i]["Magnitude"], value_mag_EQ, colours)).add_to(map_world)
#Add Legend
addLegend(map_world)
map_world
```

Figure A - 1 Python code for the Earthquake events map, code for the legend was taken from the Jupiter Notebook website (*Jupyter Notebook Viewer*, 2021).

```
# read data from the source
data_list = pd.read_html("https://ourworldindata.org/ofdacred-international-disaster-data")
```

```
data_disaster = data_list[0].copy() # The source has two tables, we need the first table
data_disaster.head()
```

	Yearly average global annual deaths from natural disasters, by decade	Drought	Earthquake	Extreme temperature	Flood	Impact	Landslide	Mass movement (dry)	Storm	Volcanic activity	Wildfire
0	1900s	130000	17302	0	63	0	5	13	1801	4494	0
1	1910s	8500	6280	0	10138	0	0	12	5995	648	107
2	1920s	472400	54935	0	428	0	43	0	11999	514	10
3	1930s	0	23770	169	436147	0	103	4	9384	318	7
4	1940s	345000	16187	0	10103	0	1753	0	12712	213	25

```
# change index:
data_disaster.index = data_disaster['Yearly average global annual deaths from natural disasters, by decade']
data_disaster = data_disaster.drop(columns = ['Yearly average global annual deaths from natural disasters, by decade'])
data_disaster.head()
```

Yearly average global annual deaths from natural disasters, by decade	Drought	Earthquake	Extreme temperature	Flood	Impact	Landslide	Mass movement (dry)	Storm	Volcanic activity	Wildfire
1900s	130000	17302	0	63	0	5	13	1801	4494	0
1910s	8500	6280	0	10138	0	0	12	5995	648	107
1920s	472400	54935	0	428	0	43	0	11999	514	10
1930s	0	23770	169	436147	0	103	4	9384	318	7
1940s	345000	16187	0	10103	0	1753	0	12712	213	25

```
plot_df = data_disaster.plot.barh(figsize=(15,12))
plot_df.set_xlabel("Yearly average global annual deaths")
plot_df.set_ylabel("Years")
plot_df.set_title("Yearly average global annual deaths from natural disasters, by decade")
plot_df.set_xlim(0,472400)
```

Figure A - 2 Python code for the graph of the yearly average global annual deaths from natural disasters, by decade.

```
# data source: https://wwwbis.sidc.be/silso/home
path_ssn = './SN_d_tot_V2.0.csv'
df_ssn = pd.read_csv(path_ssn, sep=';', usecols=[0,1,2,4])
df_ssn.head()
```

	Year	Month	Day	SSN
0	1818	1	1	-1
1	1818	1	2	-1
2	1818	1	3	-1
3	1818	1	4	-1
4	1818	1	5	-1

```
df_ssn = df_ssn.groupby('Year')['SSN'].mean()
df_ssn.head()
```

```
Year
1818    30.476712
1819    25.969863
1820    14.442623
1821     7.479452
1822     6.016438
Name: SSN, dtype: float64
```

Figure A - 3 Generation SSN data

```
n = 10
ax = df_ssn.plot(kind='bar', x='Year', y='SNN', title='Solar cycles: Average sunspot number', figsize=(20,7),
                 color='Purple')
ticks = ax.xaxis.get_ticklocs()
ticklabels = [l.get_text() for l in ax.xaxis.get_ticklabels()]
ax.xaxis.set_ticks(ticks[::n])
ax.xaxis.set_ticklabels(ticklabels[::n])
ax.set(ylabel='Sunspot Number')
ax.figure.show()
```

Figure A - 4 Python code of solar cycles graph

Appendix B Testing Null Hypothesis, codes and graphs

Date	M<5.5	M≥5.5	S_M<5.5	S_M≥5.5	I_M<5.5	I_M≥5.5	D_M<5.5	D_M≥5.5	SSN	S_Wind_Speed	Proton_Density	Proton_Temperature	A_class	B_class	C_
1996-01-03	193	1	173	1	19	0	1	0	22	485.0	5.2	132150.0	0.0	15.0	
1996-01-04	225	1	205	1	15	0	5	0	35	443.0	5.5	93895.0	0.0	22.0	
1996-01-05	191	0	181	0	8	0	2	0	56	416.0	10.2	50191.0	0.0	10.0	
1996-01-06	213	1	200	1	13	0	0	0	55	389.0	13.5	31361.0	0.0	3.0	
1996-01-07	324	2	310	1	11	0	3	1	48	350.0	11.7	19810.0	0.0	7.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2020-01-06	901	2	871	2	29	0	1	0	7	496.0	3.6	108093.0	0.0	0.0	
2020-01-07	1173	5	1132	4	39	1	2	0	4	433.0	5.0	62078.0	0.0	0.0	
2020-01-08	758	1	712	1	41	0	5	0	4	386.0	4.1	47508.0	0.0	0.0	
2020-01-09	510	1	475	1	31	0	4	0	15	446.0	5.4	111653.0	0.0	0.0	
2020-01-10	444	1	414	1	28	0	2	0	4	508.0	3.6	119082.0	0.0	2.0	

8718 rows × 17 columns

Figure B - 1 Final data, earthquake events, range by magnitude and depth

```
def understand_pattern(data, period):
    plt.figure(figsize=(12,9))
    #Source: https://seaborn.pydata.org/generated/seaborn.lineplot.html
    SSN_time = sn.lineplot(data=data, x=data.index.year, y=data.SSN, color="#DF444F", legend=False, label='SSN')
    Quantity_time = SSN_time.twinx()
    Quantity_time = sn.lineplot(data=data, x=data.index.year, y=data.EQ,
                                color="#E9B458", legend=False, label='Quantity')
    SSN_time.set_title("SSN and Quantity Of Earthquakes Over the " + period + " Solar Cycles")
    plt.xlabel('xlabel')
    SSN_time.figure.legend()
    plt.show()
```

Figure B - 2 Method for building Sunspot Number and Earthquakes graph

```
def testing_linera_nonlinear(data):
    # Prepare data for testing linear/nonlinear relationchip
    solar_data = data.copy()
    solar_data = solar_data.drop(columns = ['EQ'])
    eq_data = data.copy()
    eq_data = eq_data.drop(columns = ['SSN'])

    # Applying Linear Regression
    regressor = LinearRegression().fit(solar_data, eq_data)
    # Checking the accuracy
    return r2_score(regressor.predict(solar_data), eq_data)
```

Figure B - 3 Testing for liner/nonlinear relationship

```python
def data_exploration(data, colour, name):
    # Detect outliers:
    f1 = plt.figure()
    sn.set(style="whitegrid")
    sn.boxplot(x=data, color=colour)
    plt.title(name +' Data detect outliers')

    # Build distribution
    # Source: https://seaborn.pydata.org/generated/seaborn.distplot.html
    f2 = plt.figure()
    data_values = pd.Series(data, name = name)

    #ax = sn.distplot(data_values, color ="#E9B458")
    ax = sn.histplot(data_values, color ="#E9B458", kde=True, stat="density", linewidth=0)
    ax.set_title("Distribution of " + name)
    ax.set_ylabel("Density")

    data_mean = data.mean()
    data_median = data.median()

    Mean_legend = name +" mean = " + str(round(data_mean,2))
    Median_legend = name + " median = " + str(round(data_median,2))

    # Source: https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.axvline.html
    plt.axvline(data_mean,0,1, color="#0037FF",linestyle='--', linewidth=2, label=Mean_legend)
    plt.axvline(data_median,0,1, color="#FF1E1A",linewidth=2, label=Median_legend)

    ax.legend()

    #Probability-plot
    f3 = plt.figure()
    stats.probplot(data, plot=plt)
    plt.title('Probability plot of ' + name)
    plt.show()
```

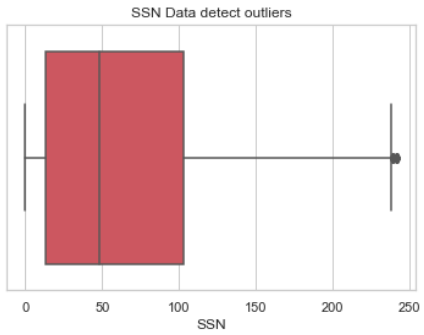
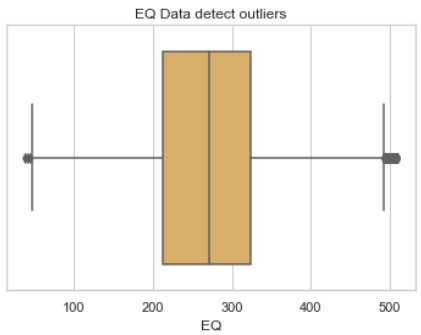
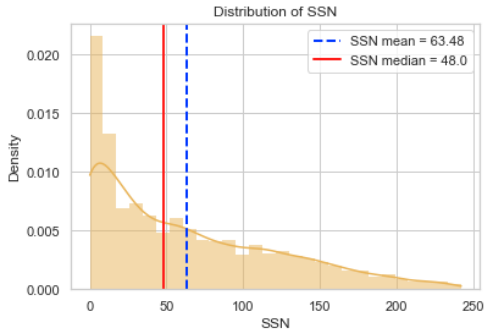
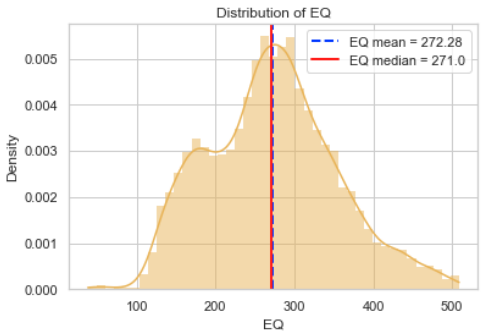
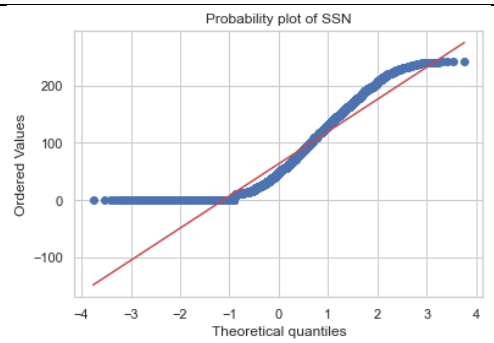
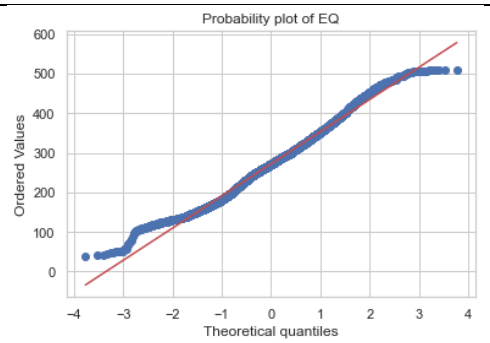
Figure B - 4 Method for building Boxplot, Distribution plot, and Probability plot

Date	EQ	SSN
1996-01-03	194	22
1996-01-04	226	35
1996-01-05	191	56
1996-01-06	214	55
1996-01-07	326	48
...	...	...
2019-12-30	460	0
2019-12-31	509	0
2020-01-04	460	12
2020-01-05	482	14
2020-01-10	445	4

8240 rows × 2 columns

Figure B - 5 Compact dataset

Table B - 1 Testing for normality of the compact data

Sunspot Number	Earthquakes
	
	
	

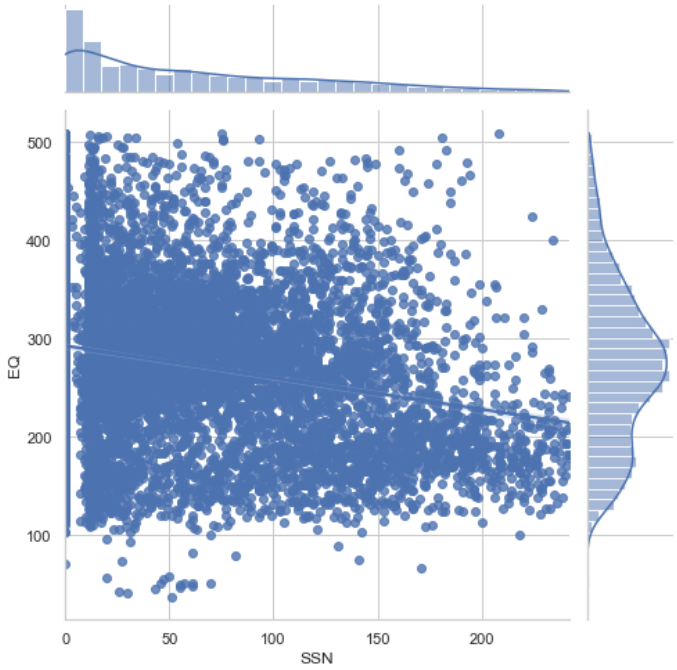


Figure B - 6 Sunspot Number and Earthquakes, compact data, relationship

*Compact data: Testing **Speraman's rho** correlation coefficient*

```
# https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html
stats.spearmanr(data_without_outliers.SSN, data_without_outliers.EQ_All, axis=None)
```

```
SpearmanrResult(correlation=-0.20099561992531892, pvalue=7.273096102097901e-76)
```

Figure B - 7 Testing Speraman's rho correlation coefficient, compact data

Date	EQ	SSN
1996-01-03	194	22
1996-01-04	226	35
1996-01-05	191	56
1996-01-06	214	55
1996-01-07	326	48
...	...	...
2008-12-27	374	0
2008-12-28	453	0
2008-12-29	420	0
2008-12-30	361	0
2008-12-31	365	0

4694 rows × 2 columns

Figure B - 8 23rd Solar Cycle original dataset

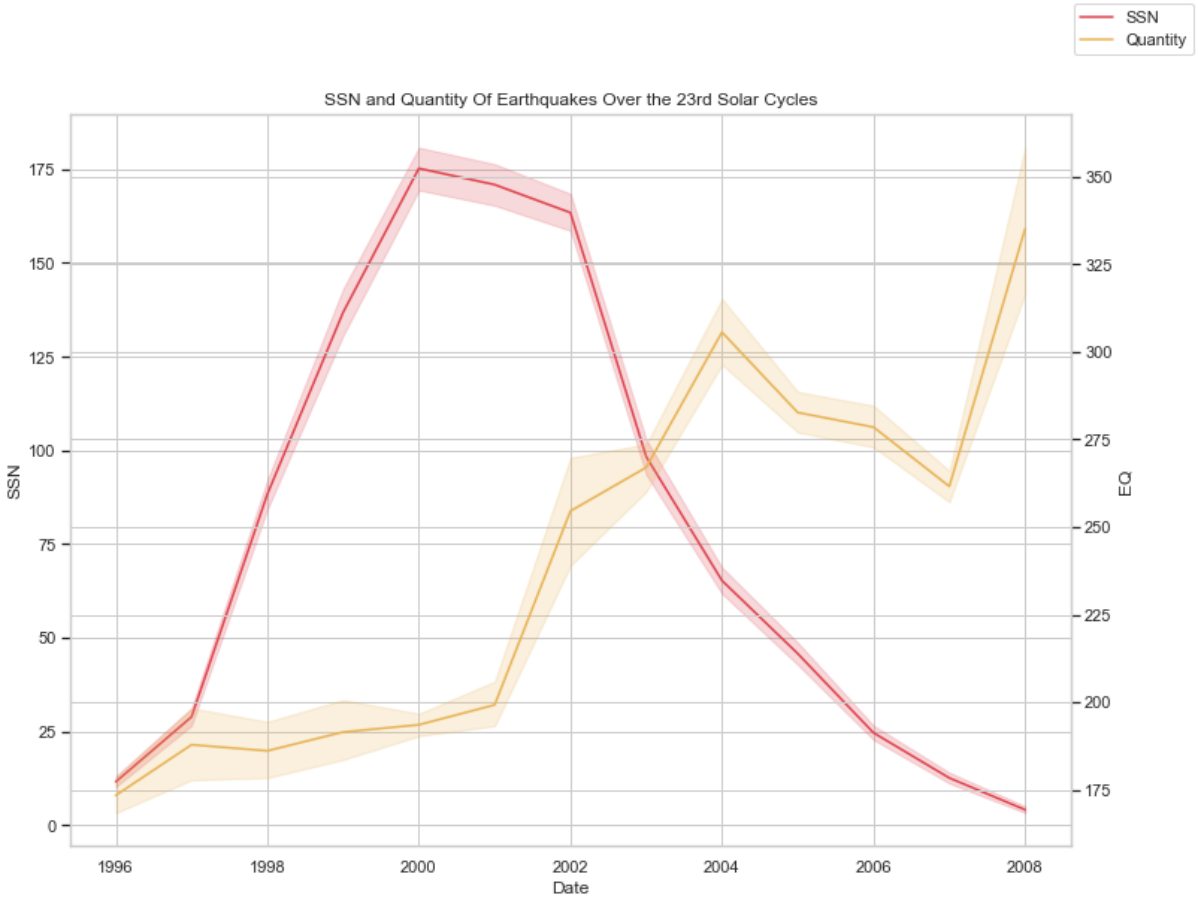


Figure B - 9 SSN and Quantity of Earthquakes Over the " 23rd Solar Cycle

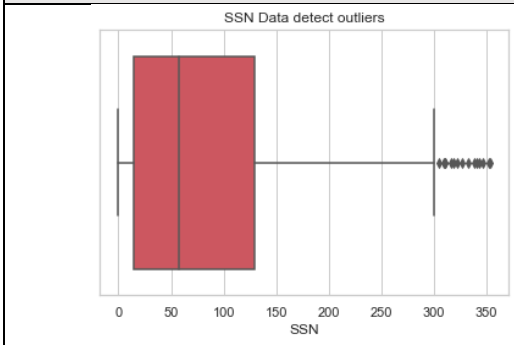
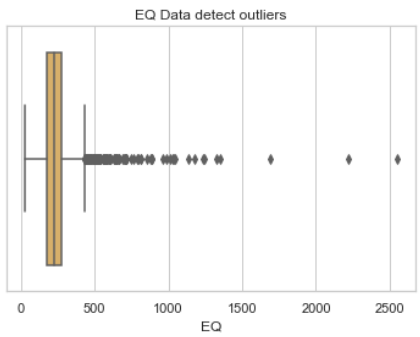
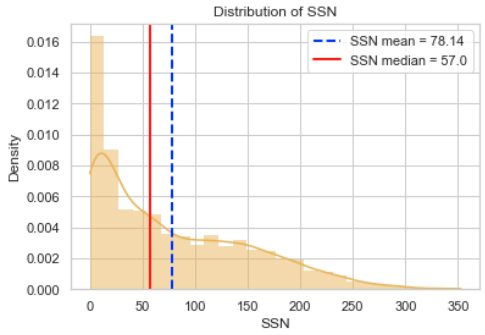
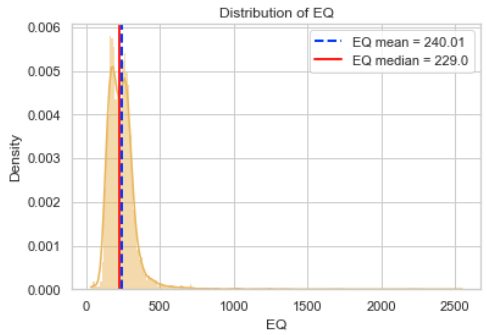
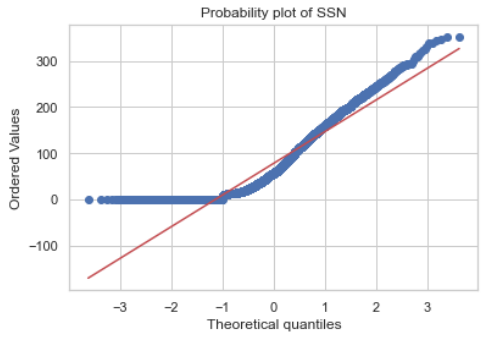
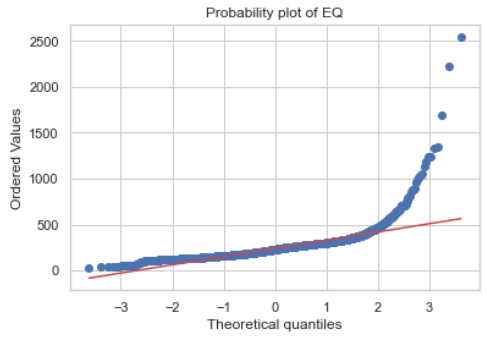
Testing linear/nonlinear

```
1 testing_linera_nonlinear(data_23)
```

```
-34.203007338333144
```

Figure B - 10 23rd Solar Cycle – testing linear/nonlinear

Table B - 2 23rd Solar Cycle, Testing for normality of the original data

Sunspot Number	Earthquakes
	
	
	

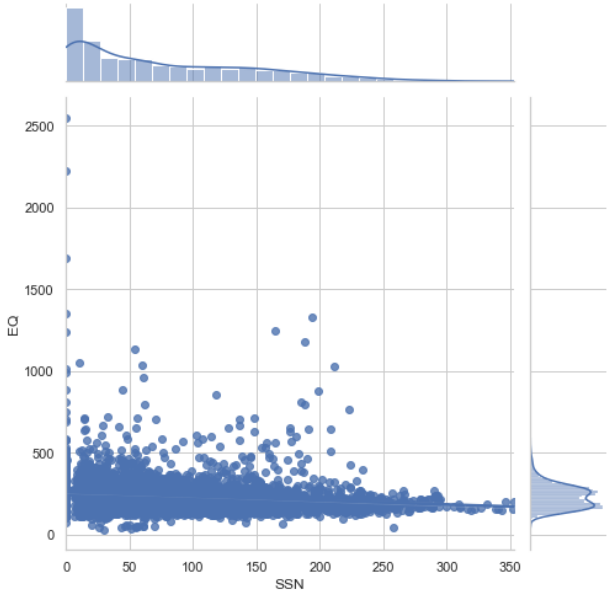


Figure B - 11 23rd Solar Cycle, Sunspot Number and Earthquakes, original data, relationship

*Original data, **23rd** Solar Cycle: Testing **Speraman's rho** correlation coefficient*

```
# https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html
stats.spearmanr(data_23.SSN, data_23.EQ, axis=None)
```

```
SpearmanrResult(correlation=-0.2351157178972734, pvalue=5.690239394386403e-60)
```

Figure B - 12 23rd Solar Cycle: Testing Speraman's rho correlation coefficient, original data

Date	EQ	SSN
1996-01-03	194	22
1996-01-04	226	35
1996-01-05	191	56
1996-01-06	214	55
1996-01-07	326	48
...	...	...
2008-12-26	267	0
2008-12-27	374	0
2008-12-29	420	0
2008-12-30	361	0
2008-12-31	365	0

4538 rows × 2 columns

Figure B - 13 23rd Solar Cycle compact dataset

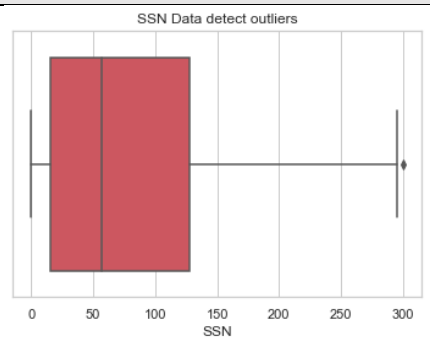
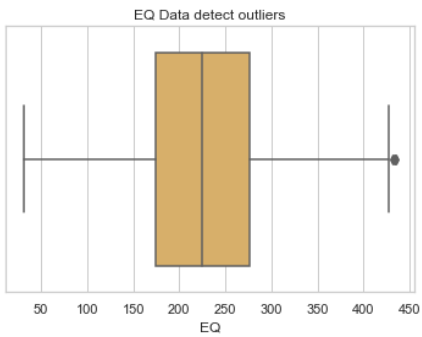
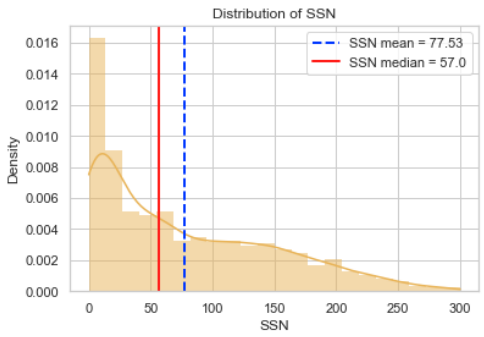
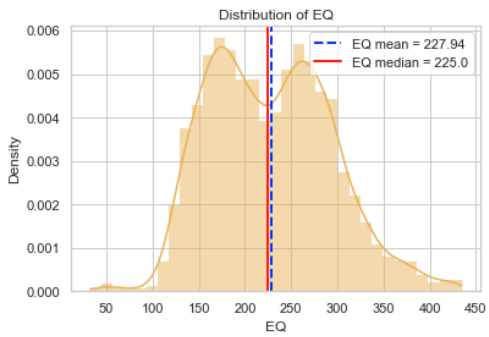
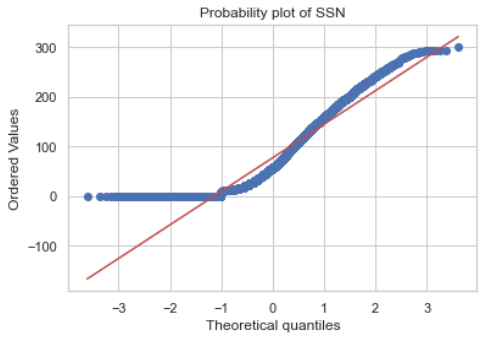
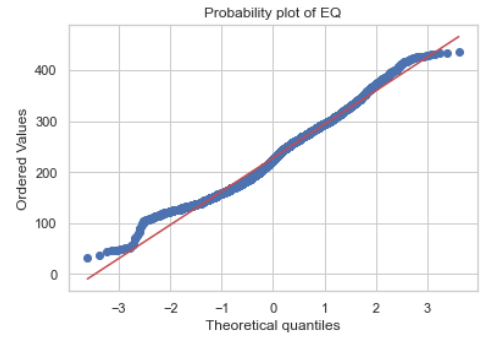
Testing linear/nonlinear

```
1 testing_linera_nonlinear(data_without_outliers_23)
```

```
-13.596039550988815
```

Figure B - 14 23rd Solar Cycle compact dataset – testing linear/nonlinera

Table B - 3 23rd Solar Cycle, Testing for normality of the compact data

Sunspot Number	Earthquakes
	
	
	

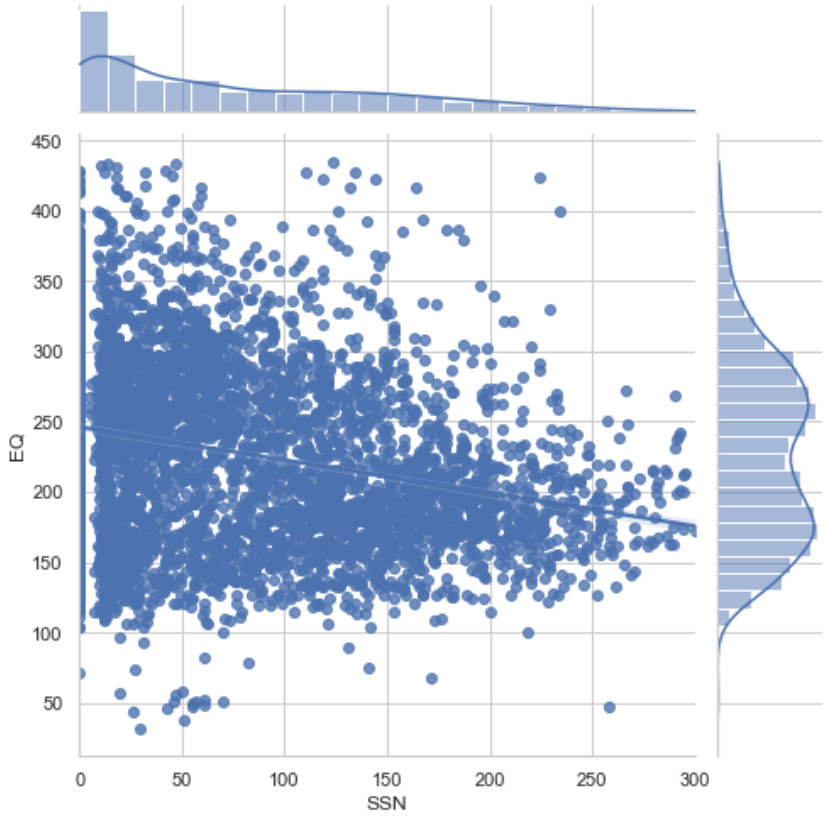


Figure B - 15 23rd Solar Cycle, Sunspot Number and Earthquakes, compact data, relationship

Compact data, **23rd** Solar Cycle: Testing Speraman's rho correlation coefficient

```
# https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html
stats.spearmanr(data_without_outliers_23.SSN, data_without_outliers_23.EQ, axis=None)
```

```
SpearmanrResult(correlation=-0.24227346713602532, pvalue=1.2785628058313536e-61)
```

Figure B - 16 23rd Solar Cycle: Testing Speraman's rho correlation coefficient, compact data

Appendix C The results of the ANOVA and Shapiro-Wilk tests

Solar activity and global earthquakes with Richter magnitude less than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
p-value =  0.17788830399513245
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.7576268911361694
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.9346261620521545
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.5397152900695801
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 17.141354107886347
p-value: 9.46731261444152e-06
There is a difference between the results.

Results for NMAE by mean
Shapiro-Wilk test for  KNN :
p-value =  0.11578981578350067
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.6425475478172302
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.716153621673584
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.7536829710006714
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 10.973481270630913
p-value: 0.00017813849485294112
There is a difference between the results.
```

Solar activity and global earthquakes with Richter magnitude equal and greater than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
p-value =  0.2568373382091522
```

```
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.5153138041496277
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.6657636165618896
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.6015990972518921
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 81.78487331971033
p-value: 2.112690765350821e-11
There is a difference between the results.
```

Results for NMAE by mean

```
Shapiro-Wilk test for  KNN :
p-value =  0.7982276082038879
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.98912113904953
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.6951642632484436
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.06597564369440079
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 21.75406563907471
p-value: 1.6551583449969694e-06
There is a difference between the results.
```

Solar activity and Shallow zone earthquakes with Richter magnitude less than 5.5

Results for NRMSE by SD

```
Shapiro-Wilk test for  KNN :
p-value =  0.18843567371368408
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.6705337166786194
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.6331677436828613
NRMSE by SD data have a normal distribution.
```

```
Shapiro-Wilk test for  LSTM :
p-value =  0.7835633754730225
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 16.645127261251286
p-value: 1.1644355779257726e-05
There is a difference between the results.
```

Results for NMAE by mean

```
Shapiro-Wilk test for  KNN :
p-value =  0.27412766218185425
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.7041400671005249
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.9487019181251526
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.2370591163635254
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 20.01063593802911
p-value: 3.0894873790568735e-06
There is a difference between the results.
```

Solar activity and Shallow zone earthquakes with Richter magnitude equal and greater than 5.5

Results for NRMSE by SD

```
Shapiro-Wilk test for  KNN :
p-value =  0.9246004819869995
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.5712231397628784
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.2598402202129364
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.29754874110221863
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 105.32973983243016
```

```
p-value: 2.0099364096415475e-12
There is a difference between the results.

Results for NMAE by mean
Shapiro-Wilk test for  KNN :
p-value =  0.006755472160875797
NMAE by mean data do not have a normal distribution
Shapiro-Wilk test for  SVR :
p-value =  0.0013201335677877069
NMAE by mean data do not have a normal distribution
Shapiro-Wilk test for  RFR :
p-value =  0.0021910914219915867
NMAE by mean data do not have a normal distribution
Shapiro-Wilk test for  LSTM :
p-value =  0.02673717774450779
NMAE by mean data do not have a normal distribution

 ANOVA test for NMAE by mean :
NMAE by mean Not all data have a normal distribution
```

Solar activity and Intermediate zone earthquakes with Richter magnitude less than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
p-value =  0.44137734174728394
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.1362820565700531
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.2664940655231476
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.9428536295890808
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 7.045178085800416
p-value: 0.00203888974565398
There is a difference between the results.

Results for NMAE by mean
Shapiro-Wilk test for  KNN :
p-value =  0.3183441460132599
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.07979367673397064
```

```
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.945838451385498
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.3366086184978485
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 3.811407188280396
p-value: 0.026070334708980063
There is a difference between the results.
```

Solar activity and Intermediate zone earthquakes with Richter magnitude equal and greater than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
p-value =  0.7315348386764526
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.01050021592527628
NRMSE by SD data do not have a normal distribution
Shapiro-Wilk test for  RFR :
p-value =  0.45113614201545715
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.9072741270065308
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
NRMSE by SD Not all data have a normal distribution

Results for NMAE by mean
Shapiro-Wilk test for  KNN :
p-value =  0.69315505027771
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.02082839235663414
NMAE by mean data do not have a normal distribution
Shapiro-Wilk test for  RFR :
p-value =  0.029222922399640083
NMAE by mean data do not have a normal distribution
Shapiro-Wilk test for  LSTM :
p-value =  0.5133634805679321
NMAE by mean data have a normal distribution.
```

```
 ANOVA test for NMAE by mean :
NMAE by mean Not all data have a normal distribution
```

Solar activity and Deep zone earthquakes with Richter magnitude less than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
p-value =  0.38038596510887146
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.5624624490737915
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.09868592023849487
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.5761890411376953
NRMSE by SD data have a normal distribution.

 ANOVA test for NRMSE by SD :
F-statistic: 79.56460379368163
p-value: 2.7216374497043484e-11
There is a difference between the results.

Results for NMAE by mean
Shapiro-Wilk test for  KNN :
p-value =  0.5104153752326965
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.2166089564561844
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.7081069350242615
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.13067221641540527
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 4.51210989205787
p-value: 0.01422752499125534
There is a difference between the results.
```

Solar activity and Deep zone earthquakes with Richter magnitude equal and greater than 5.5

```
Results for NRMSE by SD
Shapiro-Wilk test for  KNN :
```

```
p-value =  0.3110499382019043
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.41629543900489807
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.3134954273700714
NRMSE by SD data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.0007865511579439044
NRMSE by SD data do not have a normal distribution

 ANOVA test for NRMSE by SD :
NRMSE by SD Not all data have a normal distribution
```

Results for NMAE by mean

```
Shapiro-Wilk test for  KNN :
p-value =  0.5279027819633484
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  SVR :
p-value =  0.39597877860069275
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  RFR :
p-value =  0.33919471502304077
NMAE by mean data have a normal distribution.
Shapiro-Wilk test for  LSTM :
p-value =  0.8993145227432251
NMAE by mean data have a normal distribution.

 ANOVA test for NMAE by mean :
F-statistic: 11.378279252702814
p-value: 0.00014281265688313277
There is a difference between the results.
```

Appendix D The explanation of the machine learning process that has been implemented in the code in Chapter 3

The implementation of machine learning in Chapter 3 is divided into several segments. The segments are import libraries, set the line style, load data, independent and dependant variables, normalisation, testing linear/nonlinear, machine learning, and save result.

Import libraries.

To implement the machine learning algorithms the necessary libraries were uploaded. The libraries provide an access to the functions and classes, which can be used to implement the machine learning algorithms and evaluate their performance:

- Pandas (*Pandas - python data analysis library*, 2021) and Numpy (*Numpy and Scipy documentation* 2021) are commonly used for data manipulation and analysis.
- Matplotlib (*Matplotlib: Python plotting, 2021*) and Seaborn (*seaborn: statistical data visualization, 2021*) for creating static, interactive, and visualizations. The libraries are useful for exploring data and understanding their patterns and relationships.
- Scikit-learn (*Supervised learning — scikit-learn 0.24.2 documentation*, 2021) provides a wide range of tools and algorithms for tasks such as classification, regression, clustering, and dimensionality reduction. Also, the library provides tools for error calculation and data normalisation.
- Math *(Math — mathematical functions* 2021*)* provides mathematical functions and constants, used for RMSE calculation.
- Cycler (*Composable cycles — cycler 0.11.0 documentation* 2021) is creating custom colour palettes for data visualisations or plots.

Set the line style.

The implementation of the Cycle library for setting the line style and colours for better visualisation.

Load Data.

Load cleaned solar activity and earthquake data using the Pandas library.

Independent and dependent variables.

Two methods were created. The first method to get the independent/dependent variables. The second method to define if there are outliers in the data.

Normalisation.

The method was created by using the Scikit-learn library to normalise the data using different normalisation scalers and return a dictionary with normalised data to compare the normalisation results. Also, the earthquake data were divided by their characteristics (magnitude and depth). For the solar activity data, using the Scikit-learn library, PCA had been applied.

Testing linear/nonlinear.

Two methods had been created to define if there was a linear or nonlinear relationship. The first method used the Scikit-learn library for applying linear regression and counting R^2. The other method took different datasets and sent them to the first method for the calculation of R^2, then printed the results.

Machine learning.

Finding the optimal K-value

For the purpose of finding K-values, two methods were created. The method *"calculation_knn_rmse"* uses a set of training and testing datasets. The method creates an array of RMSEs that were calculated using the Scikit-learn library for a number of "K". For each "K"-value, an object (a model) was created. Then the model has been trained using independent and dependent training datasets. The next step is to use the model to make a prediction. Afterwards, the RMSE was calculated, and the RMSE value was added to the array.

The method *"rmse_graph"* creates a set of training and testing datasets and receives the array of the RMSE values from the *"calculation_knn_rmse"* method. And creates graphs for each dataset.

KNN & SVR & RFR

Several methods were created to implement the machine learning algorithms. Also, for containing actual and predicting data, there were created dictionaries (global_eq_less_5, global_eq_more_5, etc) for each part of the experiment.

For fitting the models of the traditional machine learning algorithms, the method *"fit_traditional_model_for_save_frame"* was created. The method fits the relevant model and

makes a prediction. Then it converts the results to the dataframe and saves them to the appropriate dictionary. The method uses the Scikit-learn library.

Deep learning

The method *"fit_lstm_model_for_save_frame"* was used to implement the LSTM model. The method uses the Keras library (*Keras: the Python deep learning API* 2021). Keras is a high-level deep learning Python library that is used for building and training deep neural networks. The method prepares parameters for the LSTM model. Then it builds, compiles, fits the model, and makes a prediction. Also, it converts the results to the dataframe and saves them to the appropriate dictionary.

For running the methods *"fit_traditional_model_for_save_frame"* and *"fit_lstm_model_for_save_frame"* the method "run_ml" was used. After running the "run_ml" method, actual and predicted data were saved in the dictionaries (global_eq_less_5, global_eq_more_5, etc).

Error calculation & Build plots

For the calculation of the models errors, the methods *"calculate_errors"* and *"run_error_calcuation"* were used. The methods use dictionaries, which contain actual and predicting data. The errors were calculated using the Math and Scikit-learn libraries.

Save result

To save the dictionaries that contain actual and predicting data (global_eq_less_5, global_eq_more_5, etc) appropriate paths and the method *“save_eq_frame”* were used.