**The Application of Machine Learning for Predicting Global Seismicity**

Shkuratskyy Viacheslav1, Aminu Bello Usman1, and Michael ODea1

Department of Computer & Data Science1

York St John University, UK

[viacheslav.shkurat@yorksj.ac.uk](mailto:viacheslav.shkurat@yorksj.ac.uk),[a.usman@yorksj.ac.uk](mailto:a.usman@yorksj.ac.uk), [m.odea@yorksj.ac.uk](mailto:m.odea@yorksj.ac.uk)

# 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 chapter, the authors explored the methods of how machine learning algorithms including k-nearest neighbour, support vector regression, random forest regression, and Long Short-Term Memory network can be applied to predict earthquake and to understand if there is a relationship between solar activity and earthquakes. The authors employed three types of solar activity: sunspots number, solar wind, and solar flares, as well as worldwide earthquake frequencies that ranged in magnitude and depth.

**Keywords:** Supervised Learning, K-Nearest Neighbour, Support Vector regression, random Forest Regression, Long Short-Term Memory Network, Characteristics of Earthquakes, Sunspot Number, Solar Wind, Solar Flares

# Background

Since ancient times cataclysmic disasters such as droughts, floods, earthquakes, volcanic eruptions, storms, and many other types of natural catastrophes, have 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 summer 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 disasters since 1900 excluding epidemics and famines.

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 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).

Diagram

Description automatically generated with medium confidence

*Figure 1 Yearly average global of annual deaths from natural disasters, by decade.*

As seen in Figure 1 the three deadliest natural disasters are droughts, floods, and earthquakes. However, in the last decades, the most dangerous natural disasters for people are considered to be earthquakes, extreme temperature, and floods. Even though the average global death toll from natural disasters in the 21st century is less than in the previous century, the average death rate is still high.

Most of the Earth's meteorological processes are localised, and they make good weather forecasts only in a limited area. Space weather is always global on the planetary scale (Koskinen et al., 2001).

Further, the assumption that solar activities could have an influence on 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, have 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) showed a correlation between solar activity and earthquakes with a magnitude(M) M>5.6.

Modern solar activity data and natural disaster data, as well as worldwide data, which exponentially increase every year with improved or new technologies – they contain a plethora of different parameters for solar but also natural disaster events. Reinse et al. (2018) stated that the International Data Corporation predicts an increase of the global dataset from 33 ZB in 2018 to 175 ZB by 2025. 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 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. Therefore machine learning is one of the most important and vital parts of artificial intelligence (Dunjko & 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 would solve real life problems. A quote from the original article best describes machine learning: *“Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort.”*

With increasing the processing power of computer systems in parallel to the growing amount of solar and climate data, together with implementing powerful data analyses techniques, will allow more accurate prediction levels of risks and threats of disasters, but also time and place of the disasters.

However, there are a lack of studies, that use Machine Learning techniques, which try to find the best and 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.

## Motivation

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 which use machine learning techniques for trying to predict earthquakes based on the solar activity events.

Although, it has not yet been conclusively established that solar activity events influence natural disasters. However, in accordance with findings of Love & 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.

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

## Problem Definition

Solar activity and earthquakes appear to be unrelated at first glance. However, some studies have consistently shown that there is a relationship between solar activity and earthquakes (Gribbin, 1971; Han, 2004; Odintsov et al., 2007; Novikov et al., 2020):

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.

However, it is still not clear of how and to what extent solar activity affects earthquakes. Moreover, all the 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.

## Research Contribution

The connection between solar activity and earthquakes isn't a novel one. However, just a few research studies have used machine learning approaches to investigate this association.

* 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 & Thomas, 2013).
* The study based on seismology finding that were conducted and that would assist them in employing machine learning approaches to support their results.
* Using machine learning techniques, the authors attempted to uncover any probable links between these two events.
* It can serve as a foundation for future earthquake research.

The study given here is only a step toward earthquake prediction using machine learning techniques, but it has important long-term implication.

# Earthquake

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

1. Natural disasters – events that are natural
2. Human-made disasters – events that occurred as a result of human activity.

In the last decades, as can be seen from Figure 1, the most dangerous natural disaster for people are earthquakes. According to Kanamori & 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 & Shan (2013), about 60% of all deaths that occurred due to natural disasters are caused by earthquakes.

An earthquake occurs as a result of global tectonic plate movement. There are several basic parameters that characterise an earthquake, such as depth of an earthquake, hypocentre, and magnitude:

* 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(M): The magnitude is the measure of the size of an earthquake source.

Earthquakes happen at conservative and collisional plate margins (Merle, 2011). In the conservative the tectonic plates are **sliding past** each other, while in the collision the continental tectonic plates are moving **towards** each other.

The significant earthquakes, which took place during 23rd solar cycle, are shown in Figure 2, together with the plate tectonics map in Figure 3 show, that earthquake events mainly take place at the plate tectonic boundary.

A map of the world

Description automatically generated with medium confidence

*Figure 2 Earthquake events map during 23rd solar cycle, distributed by magnitude*

The earthquake data for Figure 2, were collected from the open-source National Geophysical Data Center / World Data Service (NGDC/WDS) (1972). The dataset consists of information about significant earthquakes which meet one of the criteria such as “caused deaths, caused moderate damage (approximately $1 million or more).

Map

Description automatically generated

*Figure 3 Plate Tectonics Map - Plate Boundary Map. Source: Plate Tectonics Map - Plate Boundary Map (2021)*

There are few ways to predict an earthquake. Animal behaviours: dogs started to bark before earthquakes (Fidani, 2010), behaviour of mice different before and after an earthquake (Li et al., 2009), cows produce less milk than usual (Yamauchi et al., 2017). Other studies focussed on analysing the change in Earth’s water levels (Orihara et al., 2015; Singh et al., 2010) One of the natural triggers of earthquakes can be volcanic eruptions (McNutt & Roman, 2015), however, on the other hand earthquakes also can be a trigger for 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). Furthermore, earthquakes may be caused by man's interference with nature: heavy water pressure in dams (Chander, 1999), mining (Redmayne, 1988), testing nuclear bombs (Tian et al., 2018). Arguably, the methods of using these signs to predict earthquakes are far from being perfect (Schorlemmer et al., 2018).

# Solar activity

## Sunspots and Solar Cycles

A sunspot is a dark area which 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 10 Mm diameter sunspot may last for 2 – 3 days, a 60 Mm – 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 position. During the next solar cycle, the poles revert back. One solar cycle has a period of approximately 11 years (Figure 4). The length of the solar cycle may vary. The solar cycle has a solar minimum and a solar maximum. The solar minimum at the beginning and at the end of each solar cycle is associated with the minimum number of sunspots, while the solar maximum in the middle of the solar cycle is linked to the maximum number of sunspots. The first solar cycle was documented in the 18th century (Priest, 2014). The present, 25th solar cycle, began in December 2019 (Potter, 2020).

The state of the solar cycle is measured by calculating sunspots (Priest, 2014). Figure 4 provides a graphic representation of the solar cycles, based on average quantity of sunspot number per year.

Chart, bar chart

Description automatically generated

*Figure 4 Solar cycle and sunspot number, data source: SILSO World Data Center website (2021)*

## Solar Flares

A solar flare is an explosion of energy also accompanied by coronal mass ejection. This explosion of energy occurs because magnetic fields intersect and reorganise near sunspots. 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 et al. (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-class. However, the method they chose did not work for the prediction of small-sized solar flares.

## Solar Wind

The solar wind, according to the description of the NASA/Marshall solar physics (2014), is a “not uniform” stream of charged participles, that flows from the Sun in all possible directions, with a speed of about 400 km/s.

The source of the solar wind is the Sun’s outer atmosphere, which is called corona. The highest speed of around 800 km/s, occurred over regions where the corona is dark, and the lowest speed of around 300 km/s was observed over large cap-like coronal structures

According to Wood et al., 2009 the measurements of the solar wind are solar wind speed (velocity), proton density, and proton temperature.

The mean distance from the Sun to the Earth is 1.5x1011 m (Meyer-Vernet, 2012). So, the average time when the solar wind reaches the Earth is 4.3 days (1.5x108 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 satellites, which detect the solar wind. For example, the location of ACE (Advanced Composition Explorer) satellite between the Earth and the Sun is about 1.5\*106 km forward of the Earth (*ACE real-time solar wind | NOAA / NWS space weather prediction center*, 2021).

# Earthquakes and Solar Activity

The two events, solar activity and earthquakes, take place at different location within the solar system, on the surface of the Sun and the Earth, separated by, approximately, 1.5x1011 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 & Thomas (2013), using *χ*2 and Student's t tests, 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). On the other hand, they acknowledged that they do not have proof that the notion that solar activity has no affect is correct.

There is an assumption, that earthquakes are influenced by several factors. Bijan et al. (2013) in their study classified earthquakes into two categories (Figure 5):

1. Earthquakes that have happened as a result of the Tectonic Effect or Internal Earth’s Effects, for example rainfall, volcanic eruptions, landslides.
2. Earthquakes that have happened because of the Non-Tectonic Effect or External Earth’s Effects, for example Sun and Moon gravitation, solar activity.

Graphical user interface, text, application

Description automatically generated

*Figure 5 Classification of Triggers of Earthquakes, source: Bijan, Saied and Somayeh (2013)*

The study Novikov et al. (2020) showed the possibility that strong solar flares impact the triggering of earthquakes. In the first part of their study, they discussed the results of the laboratory experiments: electric current, generated by the artificial generator, injected into the Earth, what creates a telluric current, and copied the impact of ionizing radiation from solar flares on earthquake sources. They noticed an increase in the quantity of earthquakes with magnitude M<3, after injection of electric current into the Earth’s crust. They mentioned that the current density in the Earth’s crust depends on its electrical conductivity. It was shown, that if electrical conductivity in the depth of 10 km is higher than the electrical conductivity in the Earth surface, the current density will increase by an order. 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.

Novikov et al. (2020) observed earthquakes between August and September 2017 (on September 6, an X-class solar flare occurred). They used a quantity of global earthquakes, with magnitude M≥4, and a quantity of regional (Greece) earthquakes, with magnitude M≥3. The result is based on the comparison the numbers of earthquakes before and after the solar flare. The number of earthquakes increased in both groups of earthquakes: global (increased by 68%) and regional (increased by 120%), after the solar flare.

Odintsov et al. (2006) and Odintsov et al. (2007) tried to confirm the hypothesis of Sytinskii, (1973) that earthquakes with magnitude M>6.5 matches with high-speed solar winds whose velocity is more than 500 km/s. For their study they used a 27-year period, daily number of earthquakes with magnitude M≥5.5, and solar wind with velocity 500 km/s and above. They identified 307 cases of solar wind with this velocity.

Odintsov et al. (2006) and Odintsov et al. (2007) analysed the quantity of earthquakes on the day of solar wind arrival and a few days before and after solar wind arrivals. They found an increase of the quantity of earthquake on the day of high-speed solar wind arrival and the day after.

Also, Odintsov et al. (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 magnitude M≥7. They compare average yearly sunspot number and average yearly number of earthquakes and found that during the 11-year solar cycle number of earthquakes there are two maxima. The first maximum is the same as the maximum of sunspot number and the second maximum is during the descending phase of the 11-year solar cycle.

What is more, 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. The time of occurrence of individual strong earthquakes with magnitude Μ≥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.

Furthermore, recent data-driven studies have discovered a link between global earthquakes and solar activity. Nishii et al. (2020) set out to determine whether solar activity is a source of earthquakes. For their study they used data from *SPDF - OMNIWeb Service* (2021) for solar activity data and the *Usgs earthquake hazards program* (2021) catalogue for earthquake data. They discovered a link between solar activity and earthquakes using support vector regression, notably for earthquakes with a magnitude M<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 magnitude M≥5.6 over a 20-year period. For the earthquake data they used Storchak et al. (2013) earthquake catalogue; 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.

# Machine Learning

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, neighbour forest. All four algorithms showed accuracy prediction rates 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 predicted earthquakes using four machine learning techniques: pattern recognition neural network, recurrent neural network, random forest, and linear programming boost ensemble classifier. Every algorithm showed a different result 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 the problem and data will help to choose the right machine learning technique. Without this understanding, the choice of the machine learning techniques would be random (Müller & Guido, 2016). The basic stages of machine learning process are presented graphically in the Figure 6.

A picture containing text

Description automatically generated

*Figure 6 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 & Haenlein, 2019), (Figure 7):

* Supervised learning is based on the relationship between inputs and their outputs, based on the result and knowledge gained, which allows to make 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 as finding patterns in data (Ghahramani, 2004).
* Reinforcement learning algorithm is also called the agent. The agent learns from an environment using feedbacks and compares actions based on feedbacks by trying to choose the best one (Sutton and Barto, 2018).

A picture containing text

Description automatically generated

*Figure 7 Taxonomy of machine learning algorithms.*

## Supervised Learning (SL)

Solar activity data and earthquake data are labelled. As previously stated, supervised learning uses labelled data. Hence, supervised learning is the most appropriate option in this case.

SL methods are based on relationship between inputs and their outputs. For example, let the input variables be represented with a label “X” and an 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 of 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 et al. (2020), used SL for finding if solar activity effects on earthquake events. Murwantara et al. (2020) and Mallouhy et al. (2019) used SL predict earthquakes.

There are two major types of SL, Figure 8. The first one is classification, and the other is regression (Mohamed, 2017; Müller & Guido, 2016). A classification problem predicts “a category”, while a regression problem predicts “a number”. (Kotsiantis, 2007).

Graphical user interface, text, application

Description automatically generated

*Figure 8 Supervised learning.*

The main goal of classification supervised machine learning is to predict outputs that are categorised in nature for previously learned input data. Each output is assigned to a specific category or class (Müller & Guido, 2016).

The main goal of regression supervised learning is to estimate the value. The output data attribute of input data attributes is a numeric value. Finding the value of an object is a common application of supervised learning regression (Alpaydin, 2014).

## 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, which 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. Below are the descriptions of the summary of the main supervised ML algorithms.

1. **K-Nearest Neighbours Algorithm (KNN)** is one of the easiest algorithms, it is easy to implement and has been used in natural disasters 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 variations of the datasets and different “K” values and found that with increasing “K” value the accuracy of the identification tsunami is increasing; however, after a certain value of “K” there was no significant change in accuracy.

KNN implementation requires only two parameters: *“K”* value, what 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”* values 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 is normally calculated using Euclidean distance, *equation* (6), as 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.

|  |  |  |
| --- | --- | --- |
|  |  | *(6)* |

Where:

– data represented in n-dimensional vector

– dimension

For the regression task, the closest “K” 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.

1. **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 (Smola & Schölkopf, 2004). SVM is to find the best line, which has 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) dimension. Also, there are data points, which are the closest to hyperplane, they are called support vectors (Figure 9). SVM can be used for classification (mostly) and regression learning types. For regression tasks, SVM is called Support vector regression (SVR).

Diagram, engineering drawing

Description automatically generated

*Figure 9 SVM, two dimensions, adapted from Smola & Schölkopf, 2004.*

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 form, *equation* *(8)*.

|  |  |  |
| --- | --- | --- |
|  |  | *(8)* |

Where:

– kernel

– vectors

– feature space

The Kernel approach, which basically aids in solving the linearity and non-linearity of the equation in a very straightforward manner, is the most often utilised and useful feature of SVM. 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 et al., 2013; Jacobs, 2012).

|  |  |  |
| --- | --- | --- |
|  |  | *(9)* |

|  |  |  |
| --- | --- | --- |
|  |  | *(10)* |

Where:

– kernel parameters, c≥0, aN

|  |  |  |
| --- | --- | --- |
|  |  | *(11)* |

Where:

– kernel parameters, σ>0

|  |  |  |
| --- | --- | --- |
|  |  | *(12)* |

Where:

– kernel parameters, λ>0, ψ<0

Nishii et al., (2020), using SVR, found that solar activity effects 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 earthquake 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 of 3≤M<5 most strongly.

Murwantara et al. (2020) in their study compared three algorithms to predict earthquakes in Indonesia. They made predictions based on available earthquake data. As for the programme environment they used R with its machining learning library and methods. They found that the algorithm, which showed the best result in earthquake prediction, was SVM.

The following algorithms are tree-based algorithms. Tree-based algorithms can be used for both classification and regression learning types. Tree-based algorithms are the collection of nested “If-Else” conditions Figure 10. Tree-based algorithms start with a full population and split the data based on some condition. The splitting will continue until the stopping criteria is met (Verdhan & Kling, 2020).

1. **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).

Graphical user interface

Description automatically generated with medium confidence

*Figure 10 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 regression problem – the average of all subtrees results is the final outcome. The greater number of trees the higher the accuracy of the prediction. RF does not have an overfitting problem, it uses a random subspace method (Breiman, 2001; Hothorn et al., 2006).

Mallouhy et al. (2019) compared in their study eight different algorithms for predicting earthquake events, based on earthquake data. As for the programme environment they used Matlab. They found that the best prediction percentage was RF. Very close to RF was KNN. NB and LR yielded the worst prediction percentage.

## Dimension Reduction

When analysing data of moderate or high dimensions, it is frequently beneficial to look for ways to restructure the data and lower its dimensionality 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).

|  |  |  |
| --- | --- | --- |
|  |  | *(15)* |

Where:

– the observations

– is the mixing matrix

– 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. Calculation the eigenvector and eigenvalue of the covariance matrix is the third step. The final stage is to project raw data into a new dimensional subspace.

Dimensional reduction is an unsupervised learning technique. For efficient implementation of this unsupervised learning algorithm, there is a Scikit-learn library for Python (*Unsupervised learning — scikit-learn 0.24.2 documentation*, 2021).

## Neural Networks

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

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

One of the most 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 & Schmidhuber, 1997).

According to Hochreiter & 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)*.

|  |  |  |
| --- | --- | --- |
|  |  | *(16)* |

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

|  |  |  |
| --- | --- | --- |
|  |  | *(17)* |
|  |  | *(18)* |

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

|  |  |  |
| --- | --- | --- |
|  |  | *(19)* |
|  |  | *(20)* |
|  |  | *(21)* |

Where:

– input vectors

– hidden vectors

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

– sigmoid function

– weights

– bias vector

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

– cell input activation vector, between -1 and 1

– cell state activation vector

– 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 & Kling, 2020).

One of the most important advantages of NN compared to traditional ML algorithms is that NN can work well with an 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 an LSTM network to forecast sea surface temperature, based on the sea surface temperature dataset. Yuan et al. (2019) based on historical North Atlantic Oscillation index data created an LSTM network to predict North Atlantic Oscillation index.

## Evaluation metrics in supervised learning, regression

Evaluation metrics determine how accurate a prediction is. 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 & Smith, 1998). According to Draper & Smith (1998) the challenges that require estimating a numeric value are known as regression predictive modelling, like in this situation. As a result, the authors looked at the regression metrics 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 & 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 (R2), *equation* *(1)* – is the fraction of the variance in the dependent variable that can be predicted by the independent variable. The closer R2 to “1” the better the model fits data.

|  |  |  |
| --- | --- | --- |
|  |  | *(1)* |

Where:

– the number of data points

– predicted values

– actual values

– mean of actual values

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

|  |  |  |
| --- | --- | --- |
|  |  | *(2)* |

Where:

– the number of data points

– predicted values

– 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 to the actual values.

|  |  |  |
| --- | --- | --- |
|  |  | *(3)* |

Where:

– the number of data points

– predicted values

– 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.

|  |  |  |
| --- | --- | --- |
|  |  | *(4)* |

Where:

– the number of data points

– predicted values

– actual values

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

|  |  |  |
| --- | --- | --- |
|  |  | *(5)* |

Where:

– the number of data points

– predicted values

– actual values

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

## 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 & Kukar, 2007).

Regarding 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). Nguyen et al. (2021) in their study claimed, 70/30 is the best ratio, however they used rather small dataset of 538 samples. However, Rácz et al. (2021) suggested that 80/20 ratio is likely to be superior, especially for large datasets.

There are over 8000 records in the data used in this study. This indicate that, an 80/20 ratio would be the best option for the data.

## Types of normalising

Data normalising is the processes of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values (Muhamedyev, 2015). Normalising helps to reduce data redundancy. Normalising helps to remove anomalies and minimisation null values, what the dataset contains null values. 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 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. The authors will use boxplots to compare the results of these methods in order to find the best method.

***MinMaxScaler*** – For each component, the base guess 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)*.

|  |  |  |
| --- | --- | --- |
|  |  | *(22)* |

Where:

– data point

– minimum value in a dataset

– maximum value in a dataset

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

|  |  |  |
| --- | --- | --- |
|  |  | *(23)* |

Where:

– 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)*.

|  |  |  |
| --- | --- | --- |
|  |  | *(24)* |

Where:

– data point

– mean of a dataset

– 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)*.

|  |  |  |
| --- | --- | --- |
|  |  | *(25)* |

Where:

– data point

– median of a dataset

– 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.

Based on the comparison of all the methods, the *QuantileTransformer* showed the best result, which is why the authors chose this method.

# Experiment Method

The experiment is based on the findings of previous seismological studies. Novikov et al. (2020) and Odintsov et al. (2007) showed the relationship between strong earthquakes (magnitude M>5.5) and solar activity. However, in their studies, they did not use earthquakes with magnitude M<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, Odintsov et al. (2007) used solar wind velocity.

On the other hand, Nishii et al. (2020) using Support Vector Regression demonstrated that solar activity affects earthquakes with magnitude M<5, and that earthquakes with magnitude 3 and 4 the most strongly influenced by solar activity. For their study they used quantity of earthquakes with magnitude M≥3 and nine measurements of solar activity, including sunspot number, solar wind velocity, proton temperature. They also showed that not all nine measurements of solar activity affected earthquakes.

Therefore, the authors decided to use:

* Solar activity:
  + Sunspot Number
  + Solar Wind (speed, proton density, and proton temperature)
  + Solar Flares (A, B, C, M, X classes)
* Earthquake:
  + Earthquakes with magnitude M<5.5
  + Earthquakes with magnitude M≥5.5

Also, Novikov et al. (2020) noticed the increasing of earthquakes with magnitude M<3 after the influence of electric current on the Earth crust, what similar to solar exposure. The Earth crust has different electrical conductivity in different regions. According to Novikov et al. (2020) and Novikov et al. (2017) the higher electrical conductivity the higher current density in the lower Earth crust levels, what also leads to an earthquake. It can be assumed that, in different regions of the Earth, solar activity may have an influence on earthquakes depending on their depth. Since the authors are studying global earthquakes, the authors decided to use two earthquake options:

* Global earthquakes
* Global earthquakes divided by their depth:
  + Shallow zone earthquakes
  + Intermediate zone earthquakes
  + Deep zone earthquakes

For the experiment, the authors chose algorithms with a variety of backgrounds from both traditional machine learning and deep learning. Also, a non-linear relationship between the earthquakes and solar activity was taken into account:

* KNN – uses Euclidean distance
* SVR – kernel-based algorithm
* RFR – tree-based algorithm
* LSTM – neural networks.

For the evaluation method, it is suggested to use a variety of metrics (Chai & Draxler, 2014). Spiess & Neumeyer (2010) study used various simulation models and found that R-squared leads to false conclusions which nonlinear models are better. Also, MAPE, is not suitable here, as the earthquake dataset has a lot of zero value, would give a division by zero, *equation (2)*. Furthermore, Willmott & Matsuura (2005) indicated that RMSE is not a good measure of model performance, and it may be a deceptive indicator of average error. They suggested MAE is a preferable metric. On the other hand, Chai & Draxler (2014) stated that avoiding RMSE is not the best practice. Also, RMSE avoid using absolute values what is the benefit over MAE. Hence, for the evaluation, the authors decided to use:

* MAE
* RMSE

Based on the experiment results, the LSTM model had the best accuracy. However, the accuracy of the prediction of all algorithms is very close to each other. Moreover, the authors made a prediction that the LSTM model, compared to the other models, has more potential for increasing accuracy in further work as it has more changeable parameters than other algorithms.

# Conclusion

In this chapter, the authors explored whether machine learning is effective in predicting earthquakes based on solar activity. To attempt this, the authors first tried to find characteristics and types of earthquakes and solar activity that should be used. Based on the seismological studies, the authors chose global earthquakes ranging in magnitude and depth. As for the solar activity, the authors chose sunspot number, solar wind, and solar flares. To evaluate the efficacy and effectiveness of the machine learning algorithms used in the study, the authors opted for normalised values of RMSE and MAE. During the experiment, it was found that the relationship between earthquakes and solar activity is nonlinear, which was one of the conditions for choosing algorithms. The experiment showed that the lowest accuracy had the KNN algorithm, while the highest accuracy had the LSTM algorithm. However, all algorithms' prediction accuracy was quite close to one another. Moreover, it can be assumed that there is a possible connection between solar activity and earthquakes. However, to be sure about this statement, there are some steps that need to be taken in future work.

Also, while it is noted that the neural network showed the most potential, it is far from perfect in predicting earthquakes based on solar activity. Besides improving neural network performance, one of the first steps, based on the finding of a non-linear relationship, is to research the non-linearity of the relationship. However, neural networks require more energy and more expensive resources; therefore, additional experiments with different neural networks and traditional machine learning algorithms are required.

Furthermore, another thing to pay attention to is data. The data that were used is a good starting point, but it is better to try to use data from different sources to supplement them to improve the accuracy. Also, increasing the quantity of characteristics of earthquakes, such as earthquake location, solar activity events, such as solar energetic particles, and additional events, such as distance from the Earth to the Sun, will lead to improved prediction using machine learning algorithms. Additionally, to improve the accuracy, besides the quantity of solar flares, it is worthwhile to try the images of solar flares.

The combination of all these factors, together with experiments using different algorithms and their settings, should help improve the quality of the prediction.

# Key Terms and Definitions

Earthquake – event shaking the Earth’s surface.

Earthquake magnitude – the measure of the size of an earthquake source.

Earthquake depth – indicates where an earthquake can occur between the Earth’s surface and 700 kilometres below the surface.

Solar Activity – sunspot, solar flares, and solar wind the example of solar activity.

Sunspots – a dark area which appears on the Sun’s surface

Solar cycles – the magnetic field, generated by the sun, reverses and the north and south poles of the sun switch position.

Sunspot number – calculation of sunspots to measure the state of the solar cycle.

Solar flare – an explosion of energy also accompanied by coronal mass ejection.

Solar wind – “not uniform” stream of charged participles, that flows from the Sun in all possible directions.

Machine learning – a system that can learn from the data.

Algorithm – a step-by-step procedure for solving a problem.

Supervised learning – the relationship between inputs and their outputs that allows to make a future prediction, uses labelled data.

Dimension reduction – the process of reducing the number of traits, variables, and characteristics.

Neural networks – network of neurons to solve problems.

Evaluation metrics – determine how accurate a prediction is.

# REFERENCES

*ACE real-time solar wind | NOAA / NWS space weather prediction center*. (2021). <https://www.swpc.noaa.gov/products/ace-real-time-solar-wind>

Alpaydin, E. (2014). *Introduction to machine learning* (Third edition). The MIT Press.

Asaly, S., Gottlieb, L.-A., & 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*, 1469–1481. <https://doi.org/10.1109/JSTARS.2020.3044470>

Asim, K. M., Martínez-Álvarez, F., Basit, A., & Iqbal, T. (2017). Earthquake magnitude prediction in Hindukush region using machine learning techniques. *Natural Hazards*, *85*(1), 471–486. <https://doi.org/10.1007/s11069-016-2579-3>

Benkedjouh, T., Medjaher, K., Zerhouni, N., & Rechak, S. (2015). Health assessment and life prediction of cutting tools based on support vector regression. *Journal of Intelligent Manufacturing*, *26*(2), 213–223. <https://doi.org/10.1007/s10845-013-0774-6>

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

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

Calvo, P., Gagliano, M., Souza, G. M., & Trewavas, A. (2020). Plants are intelligent, here’s how. *Annals of Botany*, *125*(1), 11–28. <https://doi.org/10.1093/aob/mcz155>

Chai, T., & 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), 1247–1250. <https://doi.org/10.5194/gmd-7-1247-2014>

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

Dao, D. V., Adeli, H., Ly, H.-B., Le, L. M., Le, V. M., Le, T.-T., & Pham, B. T. (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), 830. <https://doi.org/10.3390/su12030830>

Das, S., Samui, P., Khan, S., & Sivakugan, N. (2011). Machine learning techniques applied to prediction of residual strength of clay. *Open Geosciences*, *3*(4). <https://doi.org/10.2478/s13533-011-0043-1>

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

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed). Wiley.

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

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), 967–978. <https://doi.org/10.5194/nhess-10-967-2010>

Ghaedi, M., Rahimi, M. reza, Ghaedi, A. M., Tyagi, I., Agarwal, S., & Gupta, V. K. (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*, 425–434. <https://doi.org/10.1016/j.jcis.2015.09.024>

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, U. von Luxburg, & 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* (pp. 72–112). Springer. <https://doi.org/10.1007/978-3-540-28650-9_5>

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

Hainzl, S., Kraft, T., Wassermann, J., Igel, H., & Schmedes, E. (2006). Evidence for rainfall-triggered earthquake activity. *Geophysical Research Letters*, *33*(19), L19303. <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), 173. <https://doi.org/10.1360/03yw0103>

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. <https://doi.org/10.1198/106186006X133933>

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), 2114–2118. <https://doi.org/10.1109/TAP.2012.2186252>

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

Kaplan, A., & 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), 15–25. <https://doi.org/10.1016/j.bushor.2018.08.004>

*Keras: The Python deep learning API*. (2021). <https://keras.io/>

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

Koskinen, H., Tanskanen, E., Pirjola, R., Pulkkinen, A., Dyer, C., Rodgers, D., Cannon, P., Mandeville, J.-C., & Boscher, D. (2001). SPACE WEATHER EFFECTS CATALOGUE. *ESA Space Weather Study (ESWS)*, *2*, 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* (pp. 3–24). IOS Press.

Li, Y., Liu, Y., Jiang, Z., Guan, J., Yi, G., Cheng, S., Yang, B., Fu, T., & Wang, Z. (2009). Behavioral change related to Wenchuan devastating earthquake in mice. *Bioelectromagnetics*, *30*(8), 613–620. <https://doi.org/10.1002/bem.20520>

Loutas, T. H., Roulias, D., & Georgoulas, G. (2013). Remaining useful life estimation in rolling bearings utilizing data-driven probabilistic e-support vectors regression. *IEEE Transactions on Reliability*, *62*(4), 821–832. <https://doi.org/10.1109/TR.2013.2285318>

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

Mallouhy, R., Jaoude, C. A., Guyeux, C., & Makhoul, A. (2019). Major earthquake event prediction using various machine learning algorithms. *2019 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, 1–7. <https://doi.org/10.1109/ICT-DM47966.2019.9032983>

Mangalathu, S., Sun, H., Nweke, C. C., Yi, Z., & Burton, H. V. (2020). Classifying earthquake damage to buildings using machine learning. *Earthquake Spectra*, *36*(1), 183–208. <https://doi.org/10.1177/8755293019878137>

Marchitelli, V., Harabaglia, P., Troise, C., & De Natale, G. (2020). On the correlation between solar activity and large earthquakes worldwide. *Scientific Reports*, *10*(1), 11495. <https://doi.org/10.1038/s41598-020-67860-3>

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

Métivier, L., de Viron, O., Conrad, C. P., Renault, S., Diament, M., & Patau, G. (2009). Evidence of earthquake triggering by the solid earth tides. *Earth and Planetary Science Letters*, *278*(3), 370–375. <https://doi.org/10.1016/j.epsl.2008.12.024>

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

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

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

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

Murwantara, I. M., Yugopuspito, P., & 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), 1331. <https://doi.org/10.12928/telkomnika.v18i3.14756>

*NASA/Marshall solar physics*. (2014). <https://solarscience.msfc.nasa.gov/>

National Flood Relief Commission. (1933). *Report Of The National Flood Relief Commission 1931 1932*. The Comacrib Press. <http://archive.org/details/reportofthenatio032042mbp>

National Geophysical Data Center / World Data Service (NGDC/WDS). (1972). *NCEI/WDS Global Significant Earthquake Database* [Data set]. NOAA National Centers for Environmental Information. <https://doi.org/10.7289/V5TD9V7K>

Nguyen, Q. H., Ly, H.-B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., Prakash, I., & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, *2021*, 1–15. <https://doi.org/10.1155/2021/4832864>

Nishii, R., Qin, P., & Kikuyama, R. (2020). Solar activity is one of triggers of earthquakes with magnitudes less than 6. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 377–380. <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), 7750–7756. <https://doi.org/10.1002/2017GL074579>

Novianty, A., Machbub, C., Widiyantoro, S., Meilano, I., & Irawan, H. (2019). Tsunami potential identification based on seismic features using knn algorithm. *2019 IEEE 7th Conference on Systems, Process and Control (ICSPC)*, 155–160. <https://doi.org/10.1109/ICSPC47137.2019.9068095>

Novikov, V. A., Okunev, V. I., Klyuchkin, V. N., Liu, J., Ruzhin, Y. Ya., & Shen, X. (2017). Electrical triggering of earthquakes: Results of laboratory experiments at spring-block models. *Earthquake Science*, *30*(4), 167–172. <https://doi.org/10.1007/s11589-017-0181-8>

Novikov, V., Ruzhin, Y., Sorokin, V., & Yaschenko, A. (2020). Space weather and earthquakes: Possible triggering of seismic activity by strong solar flares. *Annals of Geophysics*, *63*(5), 13. <https://doi.org/10.4401/ag-7975>

Odintsov, S., Boyarchuk, K., Georgieva, K., Kirov, B., & Atanasov, D. (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), 88–93. <https://doi.org/10.1016/j.pce.2005.03.004>

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

*OFDA/CRED International Disaster Data*. (2021). Our World in Data. <https://ourworldindata.org/ofdacred-international-disaster-data>

Orihara, Y., Kamogawa, M., & Nagao, T. (2015). Preseismic Changes of the Level and Temperature of Confined Groundwater related to the 2011 Tohoku Earthquake. *Scientific Reports*, *4*(1), 6907. <https://doi.org/10.1038/srep06907>

*Oxford English Dictionary*. (2021). <https://oed.com/>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*, 6.

Pham, B. T., Qi, C., Ho, L. S., Nguyen-Thoi, T., Al-Ansari, N., Nguyen, M. D., Nguyen, H. D., Ly, H.-B., Le, H. V., & Prakash, I. (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), 2218. <https://doi.org/10.3390/su12062218>

*Plate Tectonics Map—Plate Boundary Map*. (2021). <https://geology.com/plate-tectonics.shtml>

Potter, S. (2020, September 15). *Solar Cycle 25 Is Here. NASA, NOAA Scientists Explain What That Means* [Text]. NASA. <http://www.nasa.gov/press-release/solar-cycle-25-is-here-nasa-noaa-scientists-explain-what-that-means>

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

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

Raju, V. N. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020). Study the influence of normalization/transformation process on the accuracy of supervised classification. *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 729–735. <https://doi.org/10.1109/ICSSIT48917.2020.9214160>

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, 54776–54788. <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), 405–413. <https://doi.org/10.1144/GSL.ENG.1988.005.01.45>

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

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

Schorlemmer, D., Werner, M. J., Marzocchi, W., Jordan, T. H., Ogata, Y., Jackson, D. D., Mak, S., Rhoades, D. A., Gerstenberger, M. C., Hirata, N., Liukis, M., Maechling, P. J., Strader, A., Taroni, M., Wiemer, S., Zechar, J. D., & Zhuang, J. (2018). The collaboratory for the study of earthquake predictability: Achievements and priorities. *Seismological Research Letters*, *89*(4), 1305–1313. <https://doi.org/10.1785/0220180053>

*SILSO | World Data Center for the production, preservation and dissemination of the international sunspot number*. (2021). <https://wwwbis.sidc.be/silso/home>

Singh, S., Kumar, A., Bajwa, B. S., Mahajan, S., Kumar, V., & Dhar, S. (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), 685. <https://doi.org/10.3319/TAO.2009.07.17.01(TT)>

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

*SPDF - OMNIWeb Service*. (2021). <https://omniweb.gsfc.nasa.gov/>

Spiess, A.-N., & 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), 6. <https://doi.org/10.1186/1471-2210-10-6>

Storchak, D. A., Di Giacomo, D., Bondar, I., Engdahl, E. R., Harris, J., Lee, W. H. K., Villasenor, A., & Bormann, P. (2013). Public release of the isc-gem global instrumental earthquake catalogue(1900-2009). *Seismological Research Letters*, *84*(5), 810–815. <https://doi.org/10.1785/0220130034>

*Sunspot-numbers—Monthly*. (2021). <https://www.ngdc.noaa.gov/stp/space-weather/solar-data/solar-indices/sunspot-numbers/american/lists/list_aavso-arssn_monthly.txt>

*Supervised learning—Scikit-learn 0.24.2 documentation*. (2021). <https://scikit-learn.org/stable/supervised_learning.html>

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

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

Thorndike, E. L. (2000). *Animal intelligence: Experimental studies*. Transaction Publishers.

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

*Unsupervised learning—Scikit-learn 0.24.2 documentation*. (2021). <https://scikit-learn.org/stable/unsupervised_learning.html>

*Usgs earthquake hazards program*. (2021). <https://earthquake.usgs.gov/>

Verdhan, V., & Kling, E. Y. (2020). *Supervised learning with Python: Concepts and practical implementation using Python*. Apress.

Willmott, C., & 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*, 79–82. <https://doi.org/10.3354/cr030079>

Wirasinghe, S. C., Caldera, H. J., S.W. Durage, & J. Ruwanpura. (2013). *Preliminary Analysis and Classification of Natural Disasters*. <https://doi.org/10.13140/RG.2.1.4283.5041>

Witten, I. H., & Frank, E. (2017). *Data mining practical machine learning tools and techniques*. 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), 67–67. <https://doi.org/10.1080/14786445308646906>

Wood, B., Howard, R., Thernisien, A., & Socker, D. (2009). The three-dimensional morphology of a corotating interaction region in the inner heliosphere. *The Astrophysical Journal Letters*, *708*, L89. <https://doi.org/10.1088/2041-8205/708/2/L89>

Yamauchi, H., Hayakawa, M., Asano, T., Ohtani, N., & Ohta, M. (2017). Statistical Evaluations of Variations in Dairy Cows’ Milk Yields as a Precursor of Earthquakes. *Animals*, *7*(12), 19. <https://doi.org/10.3390/ani7030019>

Yuan, S., Luo, X., Mu, B., Li, J., & Dai, G. (2019). Prediction of north atlantic oscillation index with convolutional LSTM based on ensemble empirical mode decomposition. *Atmosphere*, *10*(5), 252. <https://doi.org/10.3390/atmos10050252>

Zhang, Q., Wang, H., Dong, J., Zhong, G., & Sun, X. (2017). Prediction of sea surface temperature using long short-term memory. *IEEE Geoscience and Remote Sensing Letters*, *14*(10), 1745–1749. <https://doi.org/10.1109/LGRS.2017.2733548>

APPENDIX

Python Code for the Earthquake events map, code for the legend was taken from the Jupiter Notebook website (Jupyter Notebook Viewer, 2021).

# 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

Python Code for the Solar cycle and sunspot number image.

# 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 = df\_ssn.groupby('Year')['SSN'].mean()

n = 10

ax = df\_ssn.plot(kind='bar', x='Year', y='SNN', figsize=(16,6), 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.figure.show()

ax.set\_ylabel("Sunspot Number", size=20)

ax.set\_xlabel("Year", size=20)