





# Introduction

The main objective of our project was to work with a dataset containing historical earthquake data and use the coordinates of the earthquakes in this dataset to predict the coordinates of future earthquakes.

An additional objective was to put the coordinates of the predicted earthquakes through a geolocation API to find the country of origin and get an estimated death toll of the predicted earthquake..

This document contains a brief description of our project, the main outcomes of doing the project, an analysis of the project's success and finishes with conclusions for each of the document's authors. This document's purpose is to document the successes and failures each one of the authors faced while completing this project and share each of their experiences.

# Project description

Thousands of people die from earthquakes every year without counting disasters such as the 2010 Haiti earthquake which caused 220 thousand casualties[1].

In order to contribute to society and experience the machine learning process, the team tried to predict the time and location of the next earthquake occurrence while using machine learning algorithms and data science analysis methods.

The project involved earthquake[2] and population[3] data sets from Kaggle, which contained past earthquake occurrences since 1965 and the world population data set of each country since 1960.

The project followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) while using python and Jupyter Notebook for processing. Additionally, datasets have been stored/accessed via Dropbox URLs.

# Project outcomes

Several things have been learned during the course of the project.

1. One of the most important parts of Machine Learning (M.L) is understanding data that was put off for too long. Without that knowledge the rest of the process is pointless.

2. Even though the data has been clean previously, It has been discovered how to proceed with additional aggregation of different types of tables to create a more suitable dataset.

3. Which models work well on different types of data. Some examples may be :

a. Linear Regression does not work with data that appear to be curved on the graph or categorical data needed to be encoded to use it. Additionally, it only works with the numerical type of data.

b. Decision trees work well with many different data types but have an overfitting problem.

4. It has been learned how to use Linear regression and the way it tries to predict the next value in a straight line. Additionally to see if the data can work well together Correlation matrix (heatmap) can be created which displays the connection between variables.

5. Decision tree regression is been also been used to discover that the data is too diverse since there were too many earthquakes happening in one place and at some places had only a few.

6. Random forest regression has been used where it similarly was unstable but with higher scores.

7. Use of Pandas, Seaborn, Matplotlib, Sklearn and other python libraries that are needed in trying to achieve the goals.

8. The author has learned how to use the DBSCAN Clustering ML model where the most important hyper values used in DBSCAN sklearn library is ‘eps’ and ‘min\_samples’ since it controls the distance between the main point and how many other data points are needed to become the main point.

In order to discover the best possible hyper values for the clustering the KNNs ML model has been used with the recommended number of “n\_neighbors” hyper values, which created a k-distance elbow and plotted using matplotlib.[1]

The Author achieved the following:

Using DBSCAN, KNNs ML algorithms as discussed above, the outliners from data has been removed which leaves the main clusters.

Additionally, the average time between the earthquakes for each of the clusters has been calculated and attached to the data set, therefore the way the user can predict the time of the next earthquake is by providing the Latitude and Longitude values of the location they wish to enquire.

Taking that value, the Decision tree classification algorithm has been used to discover the most possible earthquake cluster that the location values are part of and add an average time to the previous cluster occurrence time. Additionally, the user can view the cluster that the location is part of, in regards to the map of clusters.

The path that would be taken if started again would be:

1. Pick a data set with more features.
2. Evaluate the data in more detail in order to understand it.

For this Author the main outcome of the project is understanding the use of clustering and its usefulness in a project such as this. Although the author had previous experience with Python libraries such as sk-learn and pandas, K-Means and hierarchical clustering were novel aspects of the project for the Author.

In this project the Author used K-means clustering to cluster the earthquake dataset based on latitude and longitude. In doing this models could be built based on a single cluster. A prediction made on that cluster could then be said to roughly affect any populus inside the cluster's circumference. The idea was to make an accurate prediction of the magnitude of an earthquake if it occurred inside a cluster in a given number of days. Therefore the devastation of the earthquake may be predicted given certain variables. To do this a “gap” was added to each cluster's dataset. This was the number of days since the last earthquake occurred inside that specific cluster.

Once the data clustered a support vector regression (SVR), multi-linear regression, polylinear regression and decision tree regression models were modeled on all of the clusters individually. Then the mean squared error was calculated for each model on each cluster. The results were averaged with SVR returning the most accurate model.

This author's main outcome was certainly the learning outcomes. This project was a great way to introduce our group to the different machine learning algorithms and modeling techniques.This author worked with two different machine learning algorithms:

1. Logistic Regression
2. K Nearest Neighbors.

Both of these are classification methods, meaning that they are used to determine which group or “class” a data point belongs to. This meant that ultimately they weren't the right fit for our dataset. While working with the Logistic Regression model, this author discovered that because the y variable in the Logistic Regression algorithm, which in our case is a coordinate, or a continuous , must be a classification type e.g., 0 or 1. Therefore this author was unable to use Logistic Regression to complete the project's objective.

The second machine learning algorithm this author worked with was K Nearest Neighbors. The. K Nearest Neighbors is used to determine the likelihood that a data point will become a member of a group of data based on which group the data points nearest to it belong to. This proved to be more successful, but still not quite what our group was looking for. While using the K Nearest Neighbors algorithm, this author was able to use it to return the nearest “k” neighbors, with k being whatever value you assign to it. Despite receiving a prediction score of circa 92%, the author again discovers that this classification method is not suitable for the group's project and its objective.

# Analysis of project success

As outlined in the groups project specification, the main goal of the project was to create a project which could somewhat accurately predict the date and devastation of future earthquakes while following CRISP methodology. In doing so the group also strived to be exposed to machine learning topics and methods.

The group's most successful achievement, at least in this author's eyes, is knowledge and know-how the group gained from the struggles of this project. The group learned through the hardship that the most difficult part of any machine learning project is the decision on which machine learning method or algorithm best suits the data. In trying to find the answer to this question the group learned the basics of, implemented and tested many different methods of predictive analysis such as data clustering, regression as well as others. This has given the group a backbone of first hand experience with the issues and the possibilities these methods can return.

Another objective the group achieved was a semi-accurate prediction of an earthquake's magnitude. To achieve this the author clustered the data based on longitude and latitude. Then each cluster was passed through support vector regression, decision tree regression, polynomial regression and multi linear regression methods. The average mean squared (MSE) error was calculated for each of the methods by averaging the individual results from each cluster. “The mean squared error (MSE) tells you how close a regression line is to a set of points” [4]. So the closer to zero the MSE is the more accurate the model is. It was discovered that on average ,across the clusters, support vector regression gave back the most accurate results having returned a MSE of 0.35. This proved that there was much less error using this method rather than linear and polynomial regression which had a MSE of around 0.45 and decision tree regression which surprisingly returned an average MSE of 0.54.

Linear and multilinear regression accuracy that has been measured by sklearn score function were below zero, as for decision trees, a higher score was produced (84%) but with an extremely high MSE(above 4), which meant that decision trees are overfitting the data.

As for the random forest regression, the score received was 87% but with a high MSE measure of 2.4, which also meant that is overfitting the data.

In conclusion the group also had some success in applying the CRISP methodology to their work on the project. The project was a very successful learning experience for the group who successfully cleaned and prepared the data but struggled to make an accurate model or prediction which was novel or useful.

# Conclusion(s)

Personally, Author thought that project was extremely difficult and it needed more structure to work the timeline and its deliverables. Even though the team displayed ambition to succeed, the time was against them. The end product does speak for itself, therefore the largest reward has been taken as a glimpse into the ML world and the team thinks that the next project will use this experience wisely.

For this Author the use of clustering paired with regression was a fantastic learning experience for the Author who was reasonably happy with the results. Their next machine learning project will definitely be greatly aided by this and its results will be much more useful and accurate.

This author will conclude by adding that he too agrees that the project was a fantastic learning opportunity for our group. It gave a good glimpse into the vast topic of machine learning and all of its enormous potential. It was also a good introduction to working with big data sets. The one aspect this author would change would be the project topic. I feel as if the group bit off more than we could chew by selecting the topic of predicting earthquakes.

# References

[1] Humanitarian A,12 Jan 2022, “UN marks anniversary of devastating 2010 Haiti earthquake”, <https://news.un.org/en/story/2022/01/1109632>

[2] US Geological Survey, 26 Jan 2017, “Significant Earthquakes, 1965-2016”,

<https://www.kaggle.com/usgs/earthquake-database>

[3] Vaishnavi V, Aishwarya & Raja CSP, 20 Sep 2020, “World Population”,

<https://www.kaggle.com/vaishnavivenkatesan/world-population>

[4] statisticshowto

https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-squared-error

# Contributors

Dylan Scott -

Ignas Rocas -

Andrew Bashorum -