Using data mining to investigate the causes of climate change, construct predictive models and propose possible mitigations

Vo Viet Thuan
Dept. of Computer Science
University of Information Technology
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21521504@gm.uit.edu.vn

Tang Minh Hien
Dept. of Computer Science
University of Information Technology
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520229@gm.uit.edu.vn

Nguyen Thai Thanh Long
Dept. of Computer Science
University of Information Technology
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520334@gm.uit.edu.vn

Chau Thien Long
Dept. of Computer Science
University of Information Technology
Vietnam National University – Ho Chi Minh City
Ho Chi Minh City, Vietnam
21520331@gm.uit.edu.vn

***Abstract*—Starting the Second Industrial Revolution, the Earth begins to get hotter. The development of science and technology continuously harms the Earth in a concerning way. The more greenhouse gases we emit to the atmosphere, the more serious climate change is. Should no action be taken, we will eventually reach the point of no return, climate change will be irreversible. This project uses data mining to find the causes of climate change and predict future trends of related statistics. From that, some mitigations will be proposed to slow down this climate problem. The project is done to raise awareness of the current situation of climate change, so that everyone starts doing something for the Earth.**

Keywords—climate change, mitigation, data mining, linear regression, machine learning, visualization

I. Introduction

As the science and technology of humanity continues growing, climate change has been becoming more and more severe. The Earth becomes hotter, the ice in polars melts quicker, disasters become more destructive. Climate change was first hypothesized in the XIX century, and the first signs were observed in the late 1950s [3]. In more than half a century, scientists have proposed numerous measures to help reduce the speed of climate change. But as technology is still growing, especially in developing countries, these measures are just like a drop in the ocean. Moreover, since we are in the Fourth Industrial Revolution, the era of Artificial Intelligence, the training process of machine learning algorithms requires a lot of power, which in fact emits of a lot greenhouse gases to the atmosphere [4].

In this project, we will apply data mining techniques on a climate change-related dataset to gather useful information about this problem, discover the root causes of it and try to explain with the help of external resources. Then, we use what we have gathered to create predictive models that forecast future values of concerned features. By that, we can see how these values go in the next few decades. Finally, we discuss some mitigations can be made to alleviate climate change.

This project benefits from the dataset provided by IMF. The dataset contains many indicators of climate change that have the scope ranging from worldwide to country-wise or region-wise. We also use dataset about El Niño effect, since El Niño often ties with climate change in many environmental aspects. We also use dataset about El niño effect, since El Niño often ties with climate change in many environmental aspects.

II. Climate Change

A. Definition

Climate change refers to long-term changes in temperatures and weather patterns. These shifts can occur naturally, influenced by factors such as changes in solar activity or significant volcanic eruptions. However, since the 1800s, human activities have become the primary driver of climate change, mainly through the burning of fossil fuels like coal, oil, and gas [9].

B. Current situation

According to World Meteorological Organization report, as of 2019, temperature of the Earth is at least one degree Celsius above preindustrial level. Two-thirds of the world's cities are located in areas at risk of sea level rise. If appropriated actions are not taken, some cities will become Atlantis in the future. Soil degradation is directly caused by climate change, affecting about 500 million people, and threatening 30% of food cultivated in the area. Climate change strengthens natural disasters, causing more damage to infrastructure and inhabitants. [5]

C. Dataset

IMF, International Monetary Fund, provides a dataset about indicators of climate changes [1]. The dataset contains the data in various categories from most countries with the frequency ranging from daily to yearly.

TABLE I. Dataset description

Data name	Scope	Collection frequency
Surface temperature	By countries	Yearly
CO_2 concentration	Worldwide	Monthly
Sea level	By sea areas	Daily
Disaster frequency	By countries	Yearly
Energy transition	By regions	Yearly
Forestation	By countries	Yearly
Land coverage	By countries	Yearly
Greenhouse gases	By regions	Quarterly

Due to limitations, for this project, we only pick eights of them mainly focusing on the environmental factors to analyze. The table above shows some metadata of our chosen dataset.

In addition, we also use a dataset about El niño and La niña effect of NOAA, National Oceanic and Atmospheric Administration. This data will help us see the periodicity of some climate aspects related to that effect.

III. Methodology

In this section, we will explain the methods we used in the project. The first one is the exploratory analysis step. It serves the purpose of visualizing data and making sense of some information that can be easily observed. The next one is the mining step. Here, we use several techniques of statistics and data mining to discover the hidden pattern and correlation between features in the hope that we can obtain more insights about climate change and use that as our advantages to build the forecasting model. Finally, we construct some models to predict the future trends of important features, which may be helpful to further understand the relationship between features and also discover main contributors to climate change.

A. Exploratory analysis

We're facing the huge amount of data when it comes to visualization, as our dataset has a wide scope. Therefore, instead of showing everything we've visualized, we will show some important graphs and meaningful insights we've discovered.

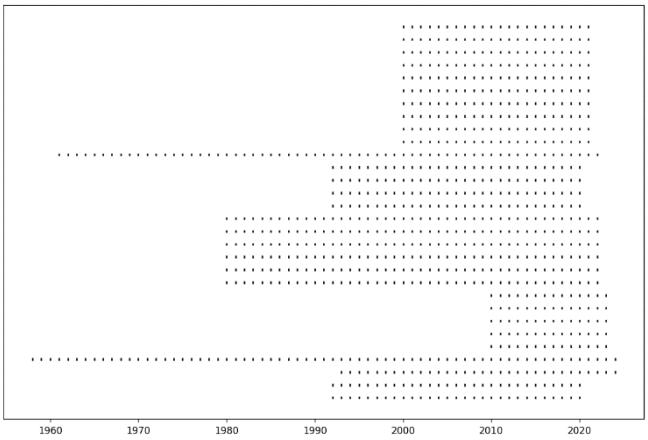

Fig 1. Availability of data of each feature in each year.

The event graph above indicates the availability of data in each features corresponding to each row. The features' names are now shown to not bundle up the graph. The dot shows that there's a valid data in that feature at that time, and the lack of dot indicates that the data is missing. We can observe that some features have data from the 1960s or 1980s, but some others only have the data in the past 10 years. This poses a problem in preprocessing and analyzing the data as the resolution might not be enough for some conclusions.

Starting from the 1980s, the Earth's surface temperature slowly increased, which was overlapped with the happening of the Third Industrial Revolution [2]. The rate has been becoming faster and faster lately and reached the peak in the past ten years.

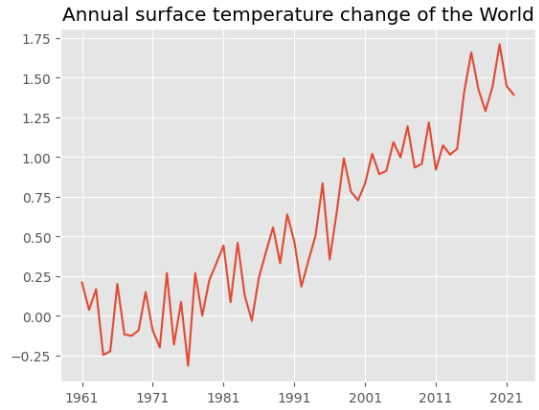

Fig 2. Worldwide surface temperature change compared to a baseline in the 1960s. (unit: degree Celsius)

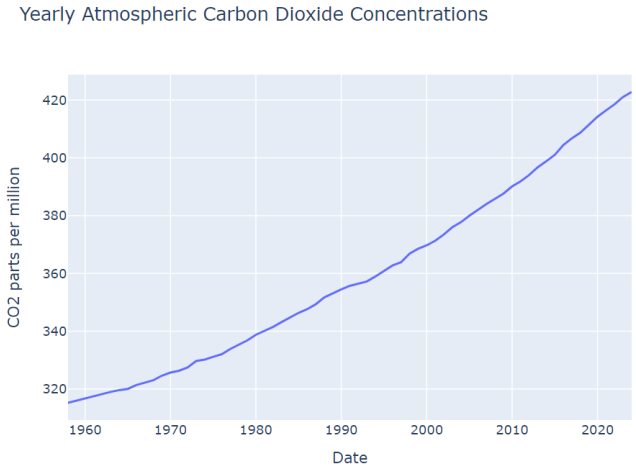

Fig 3. CO_2 concentration in the atmosphere.

CO_2 plays an important role in global warming. CO_2 absorbs energy from infrared radiation and emits it back to the earth, making the Earth hotter [6]. The concentration of CO_2 has been increased faster and faster lately, which is the main cause of the rising in surface temperature.

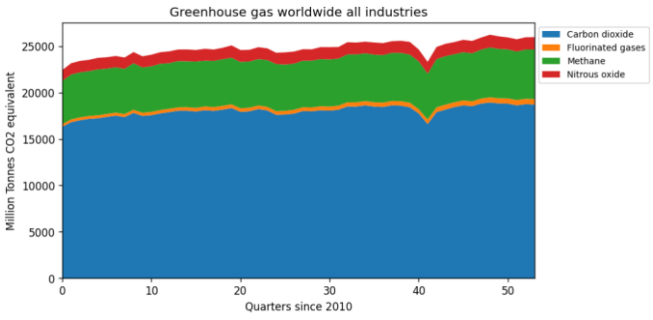

Fig 4. Area chart of each greenhouse gases over time.

Not only CO_2 contributes to the greenhouse effect, other gases like CH_4, NO_2 and F-gases also come into play. Different gases have different capabilities in trapping heat [7]. However, we can notice a common trend that all those gases tend to increasing more and more. Noticeably, at around the quarter 40 to 42, which is in the year of 2020, an outlier of the trend is easily observed. This indicates that the human activity was stalled out by Sars-Cov-2 pandemic, leading to the sudden drop of every greenhouse gases. Therefore, it is not

unreasonable to hypothesize that human activity is also a cause of climate change.

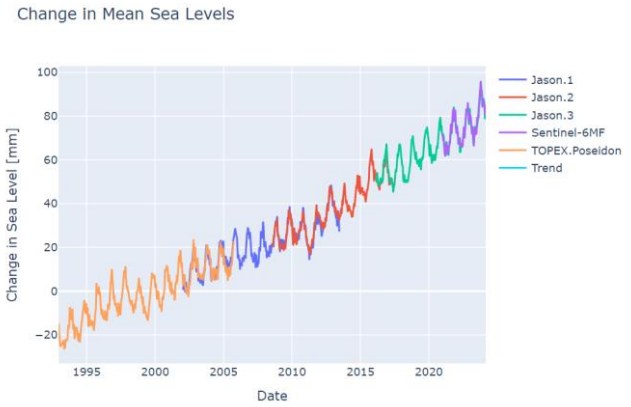

Fig 5. The average change in mean sea levels

With the constantly increasing surface temperature, the melt of polar ice is inevitable. This leads to the rising of sea level which will threaten the habitat of many coastal animals and also affect the daily life of people living near the beach.

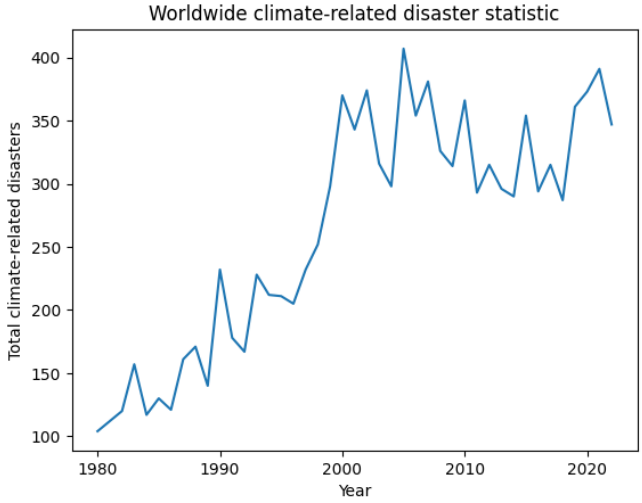

Fig 6. The number of natural disasters on Earth from 1980 to 2020

Climate change is also responsible for the higher frequency of disasters, including drought, heat wave, flood, landslide, storm and wildfire. As of 2000, the total number of such disasters each year is around 300 and above, which is doubled compared to the 1980s.

It can be observed that most countries near the North Pole have significantly increased surface temperatures compared to the standard, and the temperature rise is greater than in other regions. The root of this problem is that the Arctic is more vulnerable to climate change, due to the ice-albedo feedback [11]. Ice and snow are white and therefore reflect a lot of sunlight. After the initial warming and melting of the snow and ice, the white surface is replaced by a darker surface of the open ocean, which absorbs more sunlight, thus leading to further local warming [8]. Europe is heating up so quickly because of its proximity to the Arctic, where the effects of climate change are more pronounced. Another reason is that the ocean and atmospheric currents around Europe are generally warmer than those at similar latitudes in other parts of the world [10].

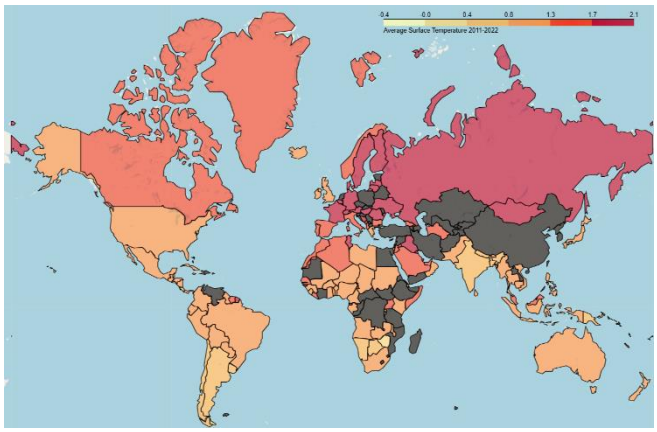

Fig 7. Average surface temperature of each country from 2011 to 2022

The Antarctica also has the ice-albedo feedback loop, but it is far more complex. As the ice melts from the bottom, and the ice sheet in Antarctica is thicker, this retains the albedo. Also, Antarctica is surrounded by the Antarctic circumpolar and Antarctic subpolar current. Thus maintains the ocean temperature better. [12]

Energy generation emits a huge amount of greenhouse gases into the atmosphere. As the climate change gets more severe, the energy transition is necessary. From 2001 to 2020, the proportion of energy generated by renewable sources increased by almost 10%. This shows that we are actually aware of climate change lately.

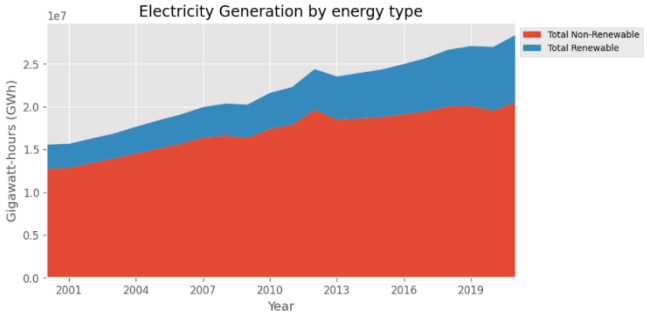

Fig 8. Amount of energy generated from renewable and non-renewable sources

B. Mining useful information

After doing exploratory analysis, we start mining useful information in the dataset. The dataset is segmented by countries or by regions depending on the category. However, we will mainly focus on the worldwide scope, unless necessary.

1. Anomaly analysis

Anomaly analysis allows us to detect unusual pattern appear on the data. With that, we will then try explaining what was going on and obtain some interesting insights that otherwise we would not be able to observe. We apply data mining techniques to detect anomalies and analyze it.

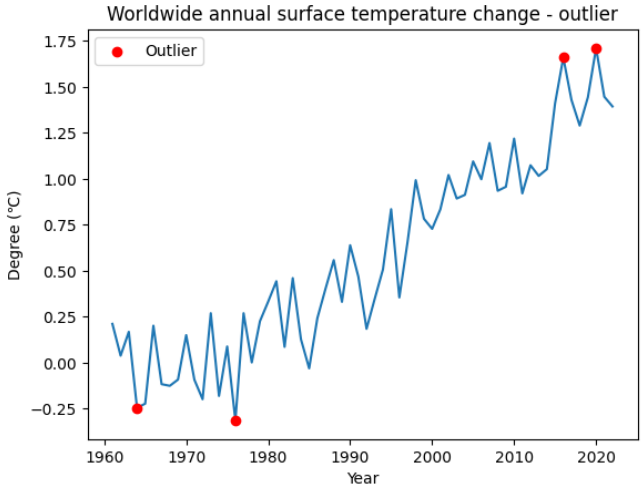

Fig 9. Anomalies in surface temperature trends

Comparing with the statistic of El Niño effect, we see a high index in 2016 and 2019. This explains why the temperatures in those years are exceptionally high. A similar observation can also be seen in the year 1963 and 1976 but with La Niña effect however.

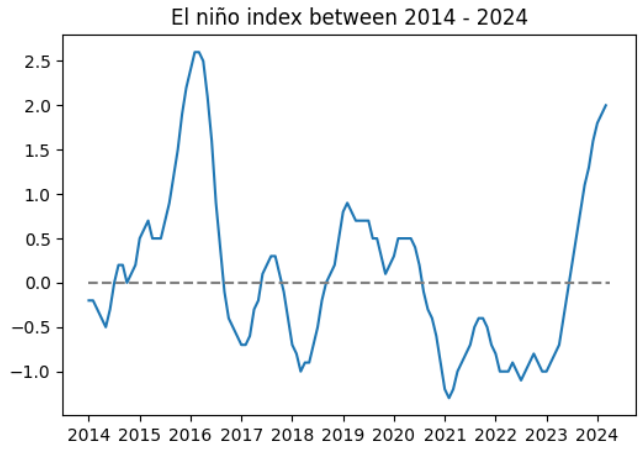

Fig 10. El Niño index from 2014 to 2024. The positive value indicates El Niño, the negative value indicates La Niña

2. Correlation coefficient

Next, we will calculate the correlation between time series. However, correlation of time series is not as straightforward as regular data, as the relationship may be delayed several timestep into the past or the future. Therefore, we will use cross-correlation since it gives us a better understanding of relations between features across different lags. Formally speaking, if x_t and y_t are two time series, the cross-correlation with lag s can be computed as:

$$\frac{\sum(x_t - \bar{x})(y_{t-s} - \bar{y})}{\sigma_x \sigma_y}$$

where $\bar{x}, \bar{y}$ are the sample mean and σ_x, σ_y are the sample variance.

In the graph below, we use cross-correlation to calculate the relationship between Forest area and Surface temperature change:

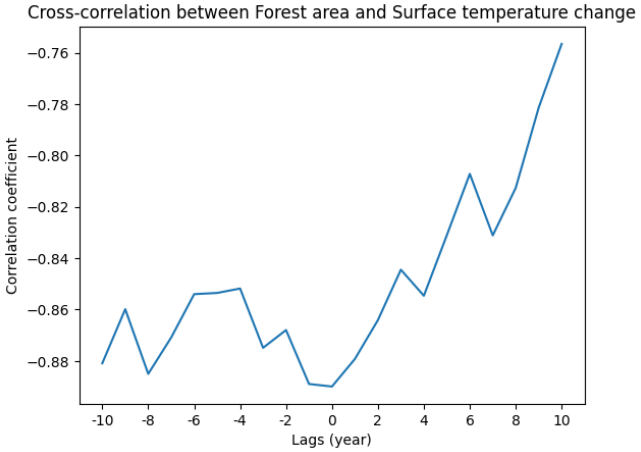

Fig 11. Cross-correlation between forest area and surface temperature across lags from -10 years to 10 years

The negative lags mean that the past values of Forest area are paired with the future values of Surface temperature, whereas the positive lags are vice versa. According to the cross-correlation, we can see that forest area negatively correlates with Surface temperature in that year or one year later. This makes sense since forests regulate the concentration of CO_2 in the atmosphere, which helps control climate change.

Upon investigation, we also noticed an interesting fact related to disaster statistics. Usually, one thought that climate change strongly influences the rate of disasters happening on Earth. But as we calculated the cross-correlation, we discovered that there are little to no disasters that have a strong correlation with the change of surface temperature. In the graph below, we can see that all features but "Flood" have a weak or no correlation to surface temperature.

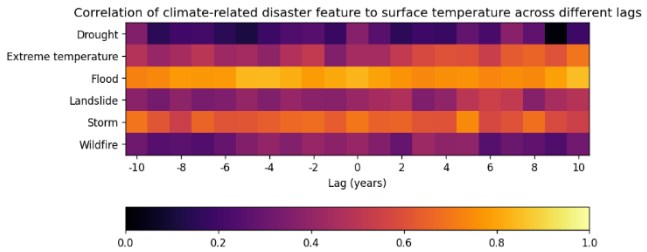

Fig 12. Cross-correlation between disasters and surface temperatures

Intuitively, the ones that are heat-related like Drought, Extreme temperatures should be strongly influenced by climate change. Counter-intuitively, almost all these disasters depend on the precipitation, which controlled mainly by El Niño effect [13]. In fact, if we plot one disaster (Drought, for example) and El Niño index together, we can see the relationship more obviously. This observation allows us to skip some features that might not be useful in follow-up prediction models.

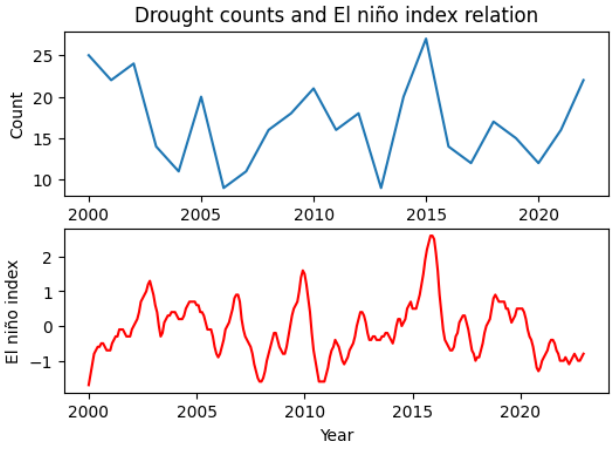

Fig 13. Relationship between the number of Drought and El Niño index

Flood stands out of these features as it has a strong correlation with surface temperature. It is because it negatively correlates with the Forest area as forests neutralize the strong flow of water in the heavy rain, which causes floods. Therefore, it indirectly correlates with surface temperature. For that reason, this feature is not useful either.

C. *Predictive models*

Predictive models are often used to forecast future trends of data based on previous observations. It will illustrate how the trends might look like in the future, which helps us know what to prepare for. In this project, we will mainly use two types of models: statistical model and machine learning model.

1. Statistical model

ARIMA, Autoregressive Integrated Moving Average, is a statistical model that often be applied to time series data to forecast future values [14]. Its prediction is based on the previous data in a specific window size. ARIMA is the combined term of AR – autoregression, I – integrated, MA – moving average. The AR part refers to how the prediction is calculated: It applies linear regression to its own past data. The I part is when the non-stationary of time series is removed, which integrated the original ARIMA model on more general data. The MA part uses moving average to predict the future value. It is used in conjunction with the AR to form ARIMA model, which is only applied on stationary time series. The ARMA model can be written as:

$$X_t = \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Here, the second term on the right side is the AR part, and the third term is the MA part. For the I part, or often called as "differencing", the non-stationary property is removed simply by create a new time series from the reference one as:

$$X'_t = X_t - X_{t-1}$$

It is obvious that the mean in time of this series is zero, since mean function is linear. In some cases, we might need to differencing the differenced time-series, which is called second-order differencing. In addition, if seasonality is presented, the number 1 in the above formula will be substituted with the value m, which is the length of the season in timesteps.

ARIMA model therefore has 3 hyperparameters: p – window of the AR, q – window of the MA and d – order of differencing. To evaluate the model on our dataset, we will fit the model on the time series but the last 5 datapoints. Then the model will forecast the next 20 datapoints. The first 5 of them will be used to evaluate the model using MAPE, Mean Absolute Percentage Error, which is the metrics often be used for forecasting problems. MAPE is calculated as (A is the target values, P is the predicted values):

$$MAPE(A,P) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_t - P_t}{A_t}\right|$$

Here is the prediction of ARIMA trained on the surface temperature time series with some hyperparameters. MAPE score of each model is 9%, 7.3%, 7.6% respectively.

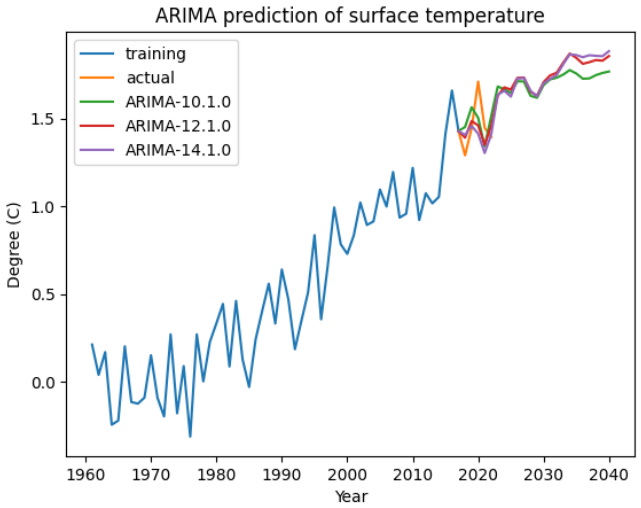

Fig 14. Prediction of ARIMA on surface temperature.

According to the graph and models' prediction, it is suggested that the surface temperature of the Earth will gradually increase in the next 2 decades, which is not a good sign.

Another aspect of climate change we can try predicting is the mean sea level. Using ARIMA, the graph below shows what the sea level will be in the next 20 years.

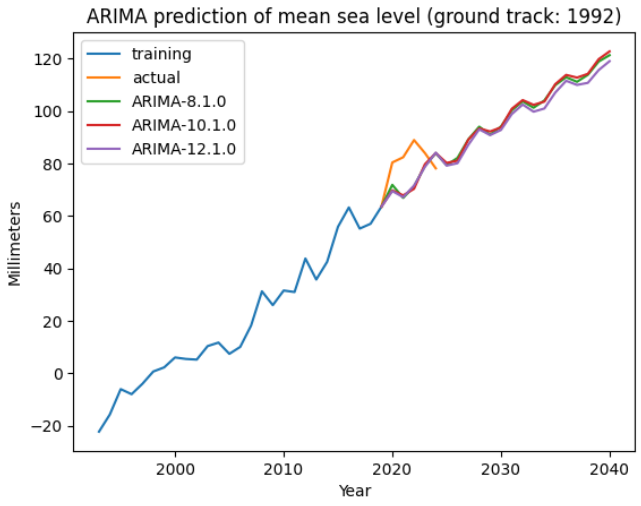

Fig 15. Prediction of ARIMA on mean sea level

Although the prediction does not match exactly with the actual data, the MAPE score is about 13%, the trend is clearly

depicted. The model suggests that the mean sea level will continue to increase, and by the year 2040, the sea level might rise 12cm or above compared to the based line in 1992.

2. Machine learning model

Machine learning models are also utilized in data mining as it can learn hidden features better than statistical models. In this section, we introduce two models used to predict future trends of surface temperature and sea levels.

a. Feature engineering

As described in the previous section, some features are not related to surface temperature. Therefore, dropping those features before training are necessary.

Moreover, since our data contains a lot of missing values, imputation is also a must. Lack of data will make our models underfit and the prediction will not be accurate. For simplicity, we will use forward fill and backward fill.

Since we are doing forecasting task, models also need future values of the features they use to predict. For this problem, we will utilize ARIMA to generate future values of those features as described above. This creates values that are more realistic than a simple extrapolation.

b. Training

Linear regression is a simplest machine learning model yet has a lot of applications. Due to its simplicity, we will utilize it to predict future trends of surface temperature and sea levels. The implementation is straightforward as we already have data and target values. After training, the MAPE score for this model is 34% and 146% for sea levels and surface temperatures respectively.

Linear regression is a simple model. But it is also so simple that it ignores the temporal relationship in the dataset. Therefore, we can use a slightly more complicated model like LSTM to capture such features. The major problem of doing this is the lack of high resolution data. In fact, by training the model on the dataset we prepared, with the window size of 5, the MAPE score of LSTM is 72% and 307% for sea levels and surface temperatures respectively. This means that LSTM is not suitable for this data at all. The graphs below show the prediction of 3 models we have discussed.

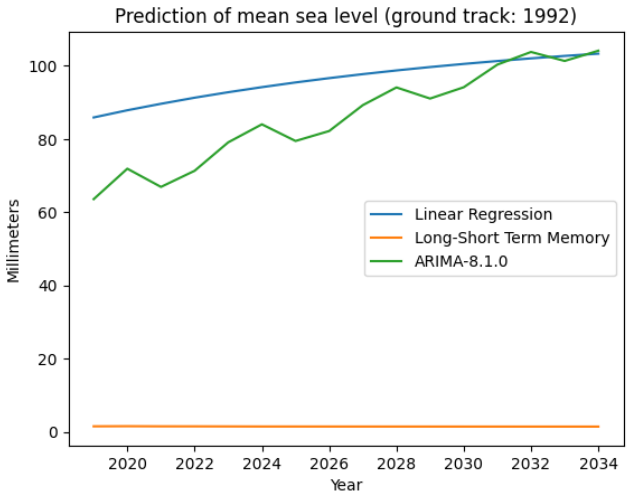

Fig 16. Prediction of all three models on mean sea levels

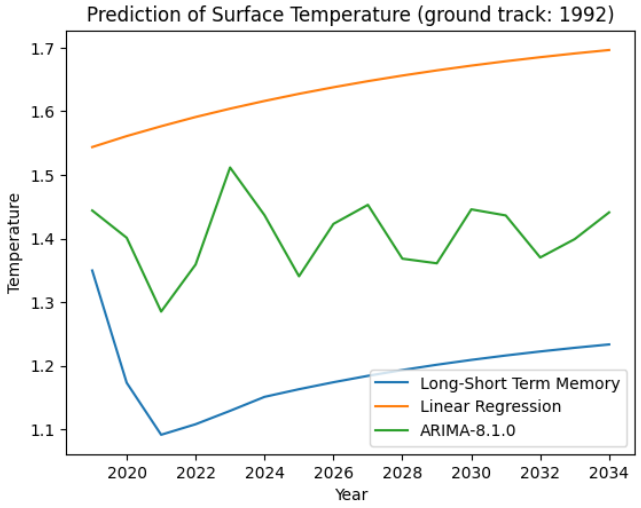

Fig 17. Prediction of all three models on mean sea levels

IV. Result Analysis

A. *Interpretation*

All of the above analyzation allow us to make some deduction about climate change.

1. Climate change is happening, and it is happening at a faster and faster rate. The more greenhouse gas we emit to the atmosphere, the worse climate change becomes. Eventually we hit a point of no return, that is when these change are irreversible [15].

2. El niño is a common climate phenomenon, but it has a good synergy with climate change, which will amplify the effects of climate change in a much worse way. In fact, the two previous year that have extremely high temperatures, 2016 and 2019, are the result of amplification of El niño in conjunction with climate change. In 2023-2024, the El niño index is also high (1.8 to 2.0), this explains why we have a hot weather in the month of April, 2024.

3. Natural disasters are strongly associated with El niño rather than climate change, but some evidence suggests that climate change is also a main factor that strengthen El niño event. Stronger El niño effect creates more catastrophic disasters. [16]

4. Climate change increases the ocean temperature, which cause the ice in the polar melts. The sea level continues to increase, which will threaten the habitat of many animals and also flood many coastal area in the world.

B. *Possible mitigations*

As the main cause of climate change are greenhouse gases, it is required to reduce their emission as much as possible. Advanced in environmental-friendly science introduces many technology that use renewable resource which does not emit or emit at low rate greenhouse gases. Forests play an important role in regulating the CO_2 concentration. Therefore, it is necessary to recover the forest area. Lack of forests also leads to devastating floods that damages the life of local habitat. As some industries require emitting large amount of CO_2 as a by-product, developing artificial carbon sinks can help reducing the emission amount significantly. Other methods have also been proposed to capture CO_2 in the atmosphere or industrial waste to process and use again or safely store.

V. LIMITATIONS

The project is concluded, but there are some problems that we have not addressed. As mentioned, our dataset has a broad scope that is not only worldwide but also country-wise or region-wise. We did not fully address the cases to better study how climate change impacts each region individually, as well as predict the trends in such regions. In addition, El niño index dataset is evidently an interesting information from which we did not fully benefit. More techniques can be used on it to gain a deeper insight into climate change. Moreover, the dataset IMF provides contains low resolution time series. We can actively collect data from many sources to have dataset that was collected at higher frequency, thus improves our model as well as provides a more accurate insight.

VI. CONCLUSION

In this project, we used data mining techniques on climate change-related dataset to gather useful information about a critical problem of humanity. By applying methods to the dataset, we made sense of some facts about climate change and explained how it happens. The models provide us how climate change might be in the next decades, which is in fact close to what scientists predict. This project also offers us opportunities to research about climate change, raising awareness to the ongoing environmental issue.

REFERENCES

[1] International Monetary Fund. 2022.Climate Change Indicators Dashboard. Accessed on [2024-04-25]

[2] Ghosh, I. (Feb 9, 2021), *Since 1850, these historical events have accelerated climate change*, in collaboration with Visual Capitalist. World Economic Forum.

[3] History.com Editors, (October 6, 2017), *Climate Change History*, (last updated, June 9, 2023). HISTORY. A&E Television Networks. Accessed on [2024-05-18]

[4] Alokya, K., (July 18, 2023). *The Green Dilemma: Can AI Fulfil Its Potential Without Harming the Environment?*, Earth.org, Access on [2024-05-18]

[5] United Nations, *The Climate Crisis – A race we can win*, un.org, Access on [2024-05-19]

[6] Sarah, F. (February 25, 2021), *How Exactly Does Carbon Dioxide Cause Global Warming?*, State of the Planet, Columbia Climate School, Access on [2024-05-20]

[7] *Understanding Global Warming Potentials*, epa.gov, United States Environment Protection Agency, Access on [2024-05-21]

[8] Matthew, H., (February 11, 2022). *"Why does the Arctic warm faster than the rest of the planet?"*, carbonbrief.org, Access on [2024-05-20]

[9] United Nations. "*What is Climate Change?*", un.org, Access on [2024-05-20]

[10] Rebecca, H., "*Why Europe is the fastest-warming continent on Earth*", npr.org, Access on [2024-05-20]

[11] Curry, J. A., Schramm, J. L., & Ebert, E. E. (1995). Sea ice-albedo climate feedback mechanism. Journal of Climate, 8(2), 240-247.

[12] Laine, V. (2007). Antarctic ice albedo, temperature and sea ice concentration trends, 1981-2000. In Annals of 2007 EUMETSAT Meteorological Satellite Conference and 15th AMS Conference, Amsterdam, The Netherlands (p. 50).

[13] Kimutai, J, Zachariah, M, Nhantumbo, B, Nkemelang, T, et al., (2024), *El Niño key driver of drought in highly vulnerable Southern African countries*, Royal Netherlands Meteorological Institute, Grantham Institute, Imperial College London, UK

[14] Prabhat, P., (Apr, 2024), *Building an ARIMA Model for Time Series Forecasting in Python*, Analytic Vidhya, Access on [2024-05-21]

[15] Bill, MG., (Dec, 2023), *The point of no return: how close is the world to irreversible climate change?*, Scientists for Global Responsibility, Responsible Science Journal, no. 6.

[16] Michael, M., (July, 2023), *Has climate change already affected ENSO?*, Science & Information for a climate-smart nation.