Natural Disasters Analysis

|  |  |  |
| --- | --- | --- |
| line 1: 1st Given Name Surname line 2: dept. name of organization (of Affiliation)  line 3: name of organization (of Affiliation)  line 4: City, Country  line 5: email address or ORCID | line 1: 2nd Given Name Surname line 2: dept. name of organization (of Affiliation)  line 3: name of organization (of Affiliation)  line 4: City, Country  line 5: email address or ORCID | line 1: 3rd Given Name Surname line 2: dept. name of organization (of Affiliation)  line 3: name of organization (of Affiliation)  line 4: City, Country  line 5: email address or ORCID |

*Abstract* — **The forecasting data on natural disaster is mainly generated by using different modeling computation as well as analyzing the data using different software and resources. The visualized and processes information are mainly understandable for everyone and considering it, this paper has carried out the analysis of the potential data on natural disaster and its victims for the last century. The main intention of this research is to evaluate the effects of natural disaster and the way it is forecasted about the future natural disaster. This paper particularly has collected the data on natural disaster from online and web dataset and the visualization and processing data has been studied including the data of single point observation, as well as computation data. The information is processed based on the visualization and application server with a two-dimensional style. The achievement of this research can be considered in visualization and processing of forecasting the disasters in future.**

**Keywords — natural disaster, timeseries, forecasting, ARIMA, prediction**

1. INTRODUCTION

There are some developing nations where natural disasters took place frequently and there is lack of restrictive measures to tackle the disaster harm. The damages are mainly causing in the higher developing countries where earthquakes, floods, volcanic eruptions, wild fire, tsunamis are common and these large scale events are sometimes cause massive disruption. The uneven development and the geographically dispersed population have left in large scale where the population has exposed in different disaster levels.

In order to analyze and to see the effectiveness of the environmental calamities as well as the victims of disasters, the effective data has been collected from EM-DAT (The International Disaster Database), was taken from its website [www.emdat.be.](http://www.emdat.be/) The data collected was from years 1900 to 2021.

Then, the information obtained was measured through different data mining processes to track useful information and insights from the large level data scale and the data mining processes including regression, classification, outlier detection, clustering, sequence analysis, association rules, social network analysis, time series analysis, sentiment analysis etc.

Time series analysis is mainly a particular process of sequencing crucial data points which have been collected over the time by maintaining a particular interval of time. In this analysis all the data points have been recorded with a particular time period at consistent intervals, rather than randomly recording the data points. Analyzing time series mainly needs a large scale amount of information points for ensuring reliability and consistency. Considering an extensive set of data generally ensures a proper representative sample size which can be reduced by noisy data. It has also confirm that any discovered patterns or trends are not accounted or outlier as seasonal variance. For forecasting purpose, time series data can be analyzed and used for predicting the future information based on historical information.

The forecasting of time series mainly happened during any kind of scientific predictions which are related to historical time stamped information. It includes developmental models using historical analysis and use of making observation and driving effective future strategic decision making.

Time series analysis highlights the way information changes with the time and effective forecasting can track the path of changing data.

1. RELATED WORK

A lot of research in data and records of natural disasters are available online. There exists many open sources for accessibility of these records. Research is conducted so that various techniques can be used for doing the analysis of the disasters and detection of this disaster is to be done for restricting it to become fatal. It is quite an effective and popular aspect that AI and ML (Artificial Intelligence and Machine Learning) play a crucial role in geospatial industry. A compact data analysis can be done efficiently with different machine learning models and these models are highly effective to measure pandemic predictions and it can be recorded with scaled and transformed manner. They can be analyzed in depth for better accuracy and predictions of data.

There are several researches have done on the data and different models associated with Machine Learning are used for prediction of the natural disasters in the various locations.

Multiple approaches have been taken into consideration based on the natural disasters dataset and they can be summarized below in a table.

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Author** | **Year** | **Findings** |
| 1. | Prihandoko, Bertalya | 2016 | Implementation of K-means has been done for analysis where clustering technique has been done. The outcome of the research has highlighted the crucial condition associated to the occurrence of natural disaster; however, the geographic condition is the major cause of the issue. |
| 2. | Muhammad [Aamir,](https://sciprofiles.com/profile/1098101) Zeeshan, Samar Alqhtani | 2021 | Deep CNN is used for predictions. A great accuracy has been achieved is comparable as well as competitive with the algorithms associated with state-of-the-art. |

Table 2.1 Literature Review

|  |  |  |  |
| --- | --- | --- | --- |
| **S. No** | **Methodology name** | **Outcomes** | **Weakness** |
| 1. | Support vector machine, Naive Bayes | Classify the natural disasters on various parameters | Limited for only early stages of natural disasters |
| 2. | Neural network and back propagation | Prediction occur on past dataset | Dynamic prediction is very much crucial for this system |
| 3. | Clustering for multivariable time series | Proposed a dynamic clustering approach for time series analysis and self- optimize organizing mapping technique | Dynamic time series data required for clustering process |
| 4. | Machine learning technique | The gradient boosting tree and CLIPER  model used for cyclone prediction | Model is still weak to produce velocity sensitivities |

Table 2.2 Comparison of state-of-the-art techniques.

1. METHODOLOGY

The first and foremost step in any project is to define a clear goal. Hence, at this point, it is important to learn every minute detail about the project. Now that the goal has been set, the next step is to find, explore and clean the data necessary for analysis. This stage takes up a lot of time but helps in finding answers to many important questions. Once the data is ready, the next step is to find meaningful insights from the data. Depending on the nature of the business problem we are dealing with we can make use of any of the following data modelling techniques to gather such insights. Now that we have built our model, we need to determine whether it meets our goals. At this stage we have achieved a desirable model. A model that meets all the requirements and goals we set for ourselves at the beginning of the project. [2]



Figure 3.1. The methodology of research

For this natural disaster analysis, we took the information from ‘The International Disaster Database’ or EM-DAT which is considered as a research center on the Epidemiology of Disasters. Moreover, the dataset contains the data from the years 1900 to 2021.

We chose two datasets, i.e., one dataset contains the data from 1900 to 2021 with 16126 records and the other dataset contains the data from 1970 to 2021 with 14644 records. We then combine both the datasets and removed the duplicate records.

The research starts with gathering the data from the EM- DAT database and loading the dataset into the Jupyter notebook. The two data sources are combined after that in order to integrate the analysis quite effective manner. The information is processed in the stage of pre- processing, where the information is sorted. The information is then used for visualization to get some insights and patterns from the data. The data features are transformed and scaled for better analysis. The data is then used for time series forecasting using the ARIMA model.

The tools and technologies we used in this research project are described below.

**Python** is considered as an effective general purpose and high level programming language which mainly has used for Machine Learning, Web Development application. It is a cutting-edge technology for the software industry. Its use in the business environment is multi-purpose because it allows mainly Procedural paradigms and Object-Oriented programming. It has significant collection which can be utilised in GUI Applications, Machine Learning, Web Frameworks, Web Scraping, Scientific Computing and many more.

**EDA** or **Exploratory Data Analysis** is considered as an effective approach for analysing the information with different visualisation processes. It is mainly used for discovering patterns, trends, and measuring assumptions with the support of graphical representations and statistical summary.

**ML** or **Machine Learning** is an effective field which allows the computers to deliver the capacity of learning without any explicitly programming. It is an innovative and unique technology which allows the computers to develop it as similar to humans- ‘The ability of learning’.

The general purpose of **NumPy** is to promote array-processing package which delivers a high-performance tools and object with multidimensional array for working with other arrays. It is also known as a fundamental package which can support scientific computing with Python.

**Pandas** is considered as an librarial open-source which is referred as Python package in order to offer different data structures as well as operations to manipulate the time series and numeric data. Pandas is highly used for analysing and importing the information within less time and higher productivity and performance.

**Scikit – Learn** is an open-source Python library that mainly applicable through different machine learning ranges with the usage of cross-validation, pre-processing, as well as visualisation algorithm within an unified interface. It is developed based on the top of SciPy, NumPy as well as Matplotilb.

**Time series analysis** is considered as an ideal approach for evaluating the collected data points sequences throughout a limited interval of time. In time series analysis, the analysts are used to record the crucial data points through a consistent interval of time period without recording the information points randomly. [4]

**Time series forecasting** is mainly used while making any kind of scientific predictions by considering the historical time stamped information. Moreover, it contains developing models with historical analysis by using observation and driving strategic decision making. [4]

The first step of the analysis is to import the required libraries into the python environment and load the dataset into a Pandas data frame. For this natural disaster analysis, we took the data from ‘The International Disaster Database’ or EM-DAT which is a Centre for research on the Epidemiology of Disasters. The dataset contains the data from the years 1900 to 2021. We chose two datasets, i.e., one dataset contains the data from 1900 to 2021 with 16126 records and the other dataset contains the data from 1970 to 2021 with 14644 records. We then combine both the datasets and removed the duplicate records. The statistical information of the data is then analysed.

The second step of the analysis is to process the data where the data is cleaned. The pre-processing stage includes type conversion, missing value treatment, feature transformation and scaling and outlier detection.

The type conversion function converts the original column datatypes into best possible datatypes for the respective columns.

The dataset may contain some missing values. This treatment can be done in four ways – remove the rows containing missing values, remove the columns containing missing values, imputing the missing values in a numerical column with its mean or median and imputing the missing values in a categorical column with its mode.

Outliers are treated in two ways – remove the rows containing outliers, replace the outliers in a column with its median value.

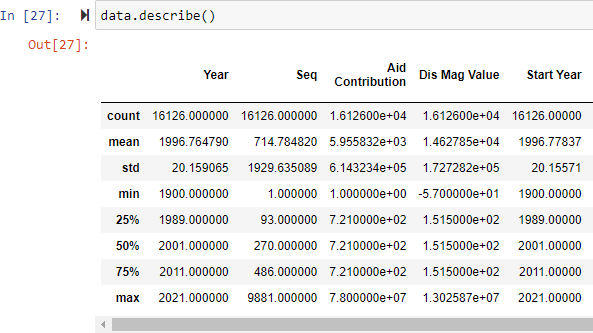
The next step is to transform and scale the features in the dataset. Feature pre-processing is a crucial step in development of a Machine learning or ML model. It is the only way to gain a better score and also a crucial point how you represent your data and feed it to a target model. It is associated with mathematical transformation which can be applied with a mathematical formula for featuring or transforming the useful values for further analysis. On the other hand, Feature Scaling is considered as a standardise process which can independently feature the fixed range of the data. It is utilised in the pre-processing phase of the data for managing the highly varying magnitudes. There are various techniques like Min Max Scaling, Robust Scaling, Log Transformation, Max Abs Scaling, Label Encoding and many more.

Finally, the data is used for forecasting and time series analysis. Generally, time series analysis contains with different methods for analysis of the time series data for extracting the relevant statistics as well as the characteristics of the information. On the other hand, time series forecasting is mainly used for predicting future values by using specific model considering the previously observed values. The components associated with the time series are effective to address the general direction of time series data for a long time period. Exhibiting the trends through Seasonality may repeats in terms of direction, timing as well as the magnitude. Further, the irregular variation mainly fluctuates with the time series data and has become quite evident with the removal of cyclical and trend variation. Decomposition of ETS is also used for separating the different components associated with the time series.

In this natural disasters analysis, we use the data from 1900 – 2021 and try to forecast the natural disasters in the future i.e., from the year 2022, using the ARIMA model.

ARIMA, stands for Autoregressive Integrated Moving Average Model with three order parameters:

*(p, d, q).*

1. AR(*p*) Autoregression – It is mainly a regression model which is used to build a dependent relationship among the present observation and the observation of past. Moreover, an auto-regressive component mainly refers the usage of past values within a regression equation for time series..
2. I(*d*) Integration – It mainly uses for differentiation of observation to make stationary time series. Differencing mainly refers subtraction of the present values of a time series with past values d number of times.
3. MA(*q*) Moving Average – It is a model which is been used for addressing the dependency between a residual error and an observation from a moving average component applied for lagging observations. A moving average component mainly addresses the error of the model as a combo of post error terms. Here the order q refers the number of terms has included in the model. [5]

There are different types of ARIMA model which are stated below.

1. ARIMA – Non seasonal Autoregressive Integrated Moving Averages
2. SARIMA – Seasonal ARIMA
3. SARIMAX – Seasonal ARIMA with exogenous variables

The ‘auto\_arima’ function from the ‘pmdarima’ library to support the relevancy of different optimal parameters associated with ARIMA modeland delivers a suitable ARIMA model.

1. RESULTS AND EVALUATION

The data is used in visualization for getting some useful insights and patterns.

From the core, Statistics can be relate with a sophisticated Machine Learning or ML algorithms to capture and translate the data patterns in actgionable evidence.

Figure 4.1 Statistical Distribution of Data

Correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable value increases, the other variables values decrease. Correlation can also be neutral or zero, means that the variables are unrelated. If the highly correlated features are not picked out, the model may end up with predicting the biased results.

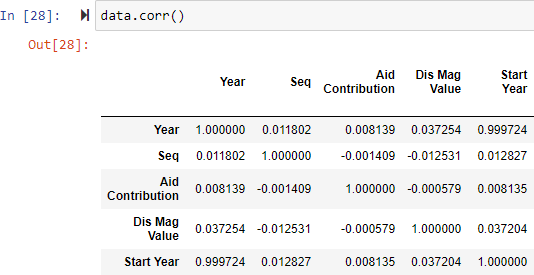


Figure 4.2 Correlation of Data

The below graphs gives us some useful insights about the natural disasters.

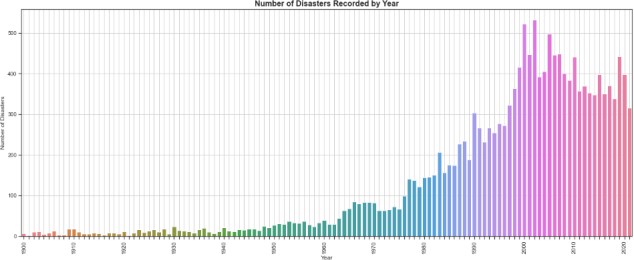
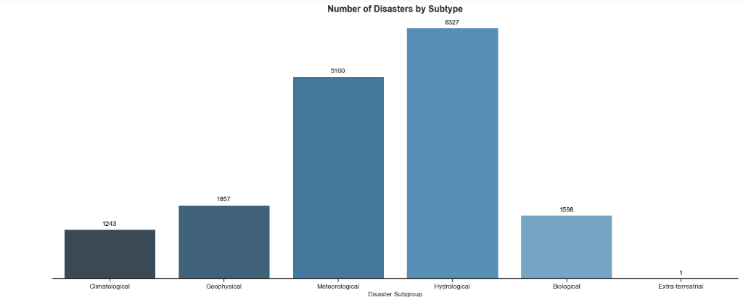
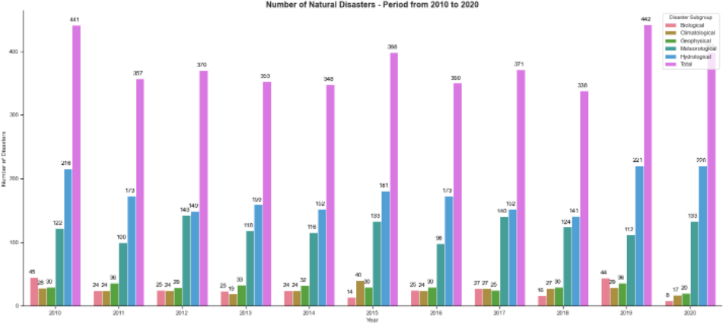
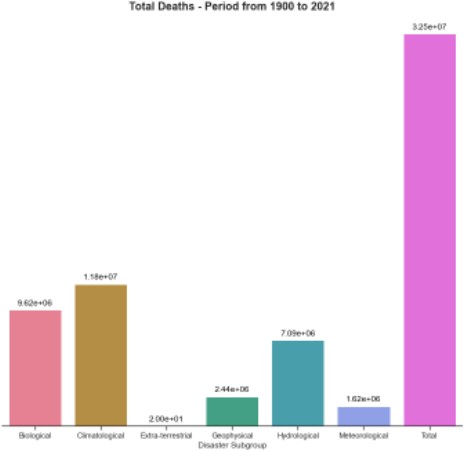


Figure 4.3 Number of Disasters recorded per Year

Figure 4.4 Number of Disasters by Subtype

Figure 4.5 Number of Disasters from 1900 – 2020

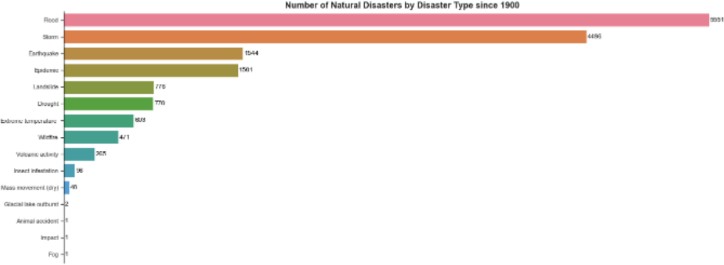
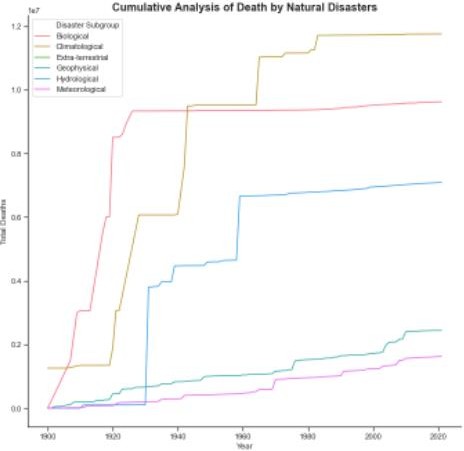
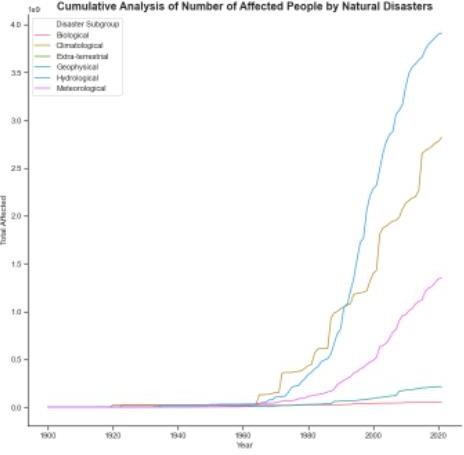


Figure 4.6 Number of Disasters by Type

Figure 4.7 Cumulative Analysis of Death by Natural Disasters



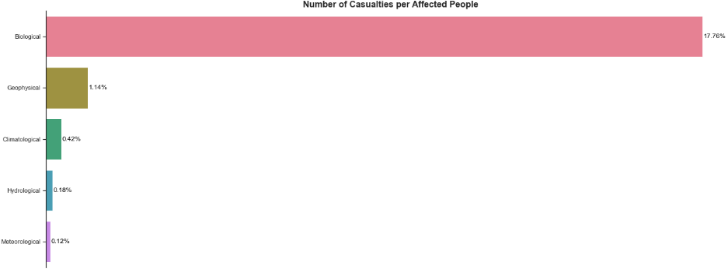


Figure 4.8 Cumulative Analysis of Number of Affected People by Natural Disasters

Figure 4.9 Total Deaths

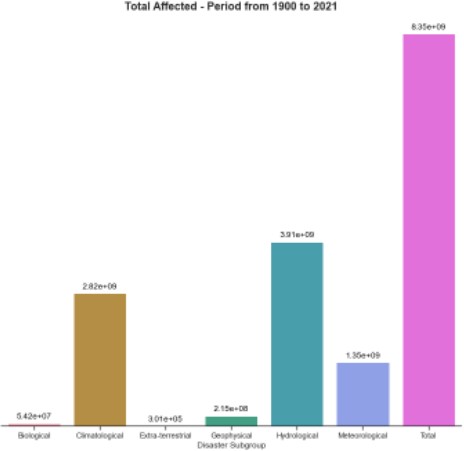
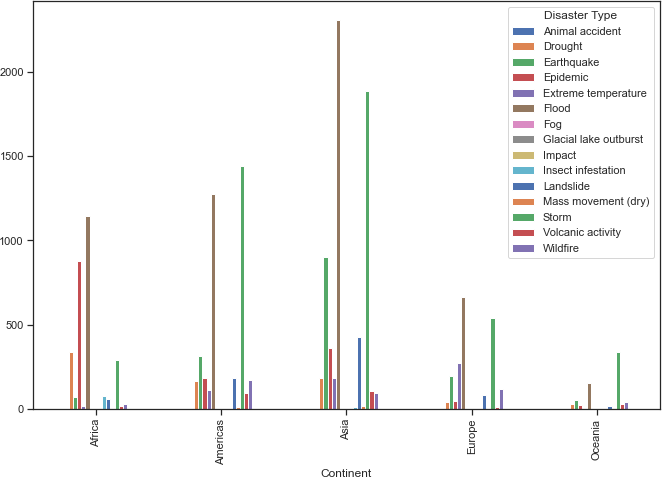


Figure 4.10 Total Affected

Figure 4.11 Number of Causalities per Affected People

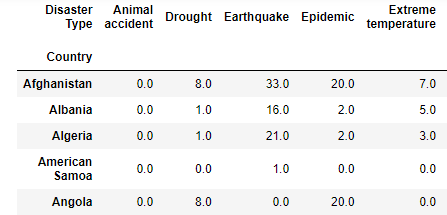
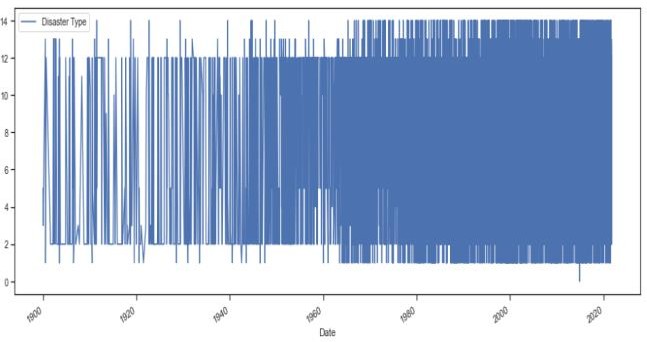


Figure 4.12 Number of Disasters for each continent

Figure 4.13 Number of Disasters for each country

**Time Series Forecasting with ARIMA**

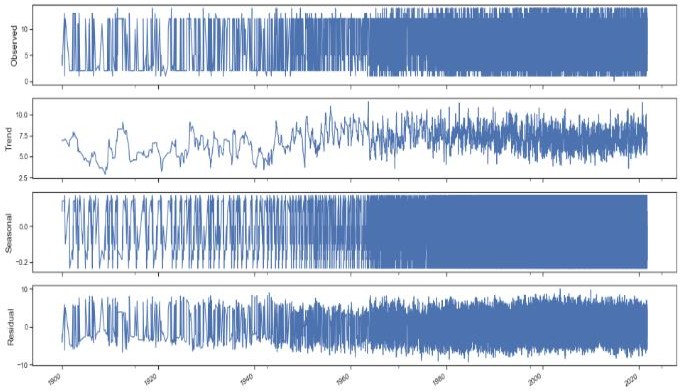
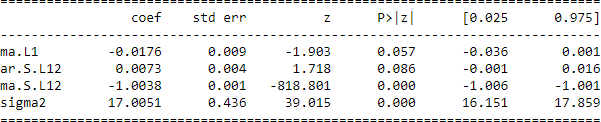
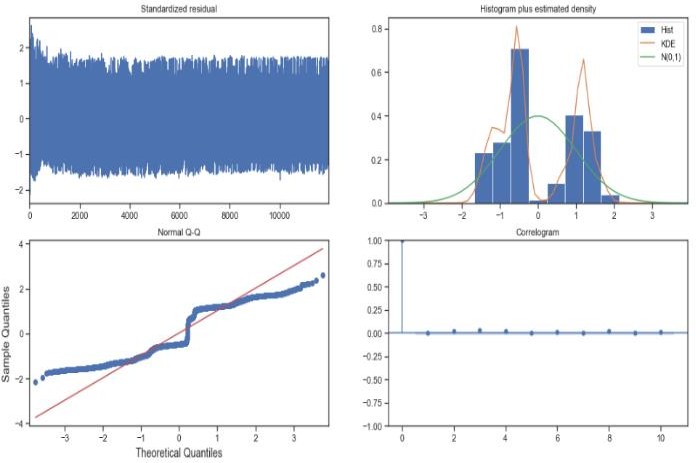


Figure 4.14 Visualizing Time Series data

We can also visualize our data using a method called time- series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and noise.

Figure 4.15 Seasonal Decomposition of data

We used ARIMA model to forecast disaster type in future years. The below shows the statistical summary of our ARIMA model.

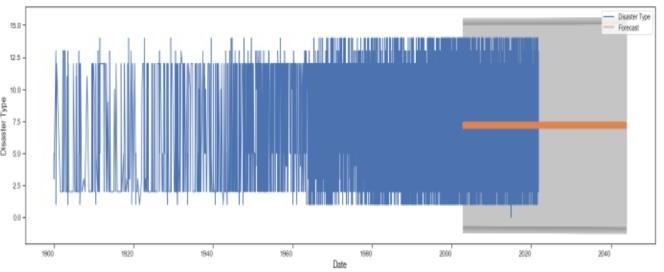


Figure 4.16 Statistical Analysis of ARIMA model

The model has forecasted the disaster type for the years 2019 – 2021 and are compared with the original data.

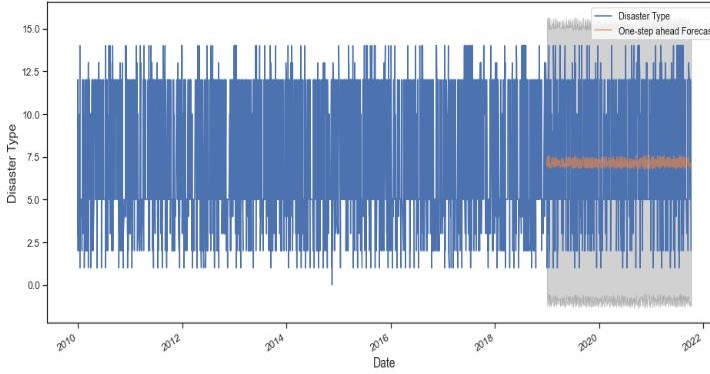


Figure 4.17 Validating forecasts

Our model clearly captured disaster type seasonality for the years 2019 - 2050.

Figure 4.18 Visualizing forecasts for future years

As we forecast further out into the future, it is natural for us to become less confident in our values. This is reflected by the confidence intervals generated by our model, which will grow larger as we move further out into the future.

1. CONCLUSION AND FUTURE WORK

The natural disasters are frequently happened due to the geographical position of the country. Thus, natural disasters mostly occurred as an impact of natural condition. This research has completed an analysis to the data published by EM-DAT to find out the correlation between the natural disasters happened, the number of victims. The analysis also helped us to identify that floods and storm are most common natural disasters that occurred frequently. Also, the analysis helped in forecasting the future occurrence of natural disasters in various countries.

In our research analysis, we have forecasted the disaster type that may occur in the future year. In the future, the research will be continued to forecast the disaster type in the future year specific to the country and the continent.

1. REFERENCES

[1] Prihandoko, Bertalya, “A Data Analysis Of The Impact Of Natural Disaster Using K-Means Clustering Algorithm”, vol. 8, no. 4, December 2016

[2] W. I. D. Mining, “Data Mining: Concepts and Techniques,” Morgan Kaufinann, 2006.K. Elissa, “Title of paper if known,” unpublished.

[3] [www.kaggle.com](http://www.kaggle.com/)

[4] [www.towardsdatascience.com](http://www.towardsdatascience.com/)

[5] [www.geeksforgeeks.org](http://www.geeksforgeeks.org/)