Climate Monitoring and Prediction using Supervised Machine Learning

Rudransh Kush[1], Shubham Gautam[2], Sai Suvam Patnaik[3], Tanisha Chaudhary[4]

[1,2,3,4]*Student, Department of Computer Science and Engineering, Bennett University, Greater Noida, Uttar Pradesh, India*

---***---

Abstract – *The objective of the research paper is to firstly analyse about the problems occurring due to the climatic changes happening irregularly, and therefore arising the need of a climate or weather monitoring Machine learning based prediction system. Aim of the paper will be fulfilled by proposing a methodology related to predicting precipitation type and Weather climate by considering several other factors also on which environment depends both directly and indirectly. There are various factors like cloud formations, wind patterns, continentally which impact climate predictions. Current climate systems [1, 2] have been fairly successful in including such factors. But this success comes with some limitations, this is because since recent times there has been another scientifically acclaimed factor added to the list which is human society intervention ranging from deforestation for development purposes to pollution caused by industries. This paper brings forward with itself not only the methodology in order to predict the climate, but also inculcating a need for preservation of our environment by using data visualizations methods which in most cases create a bigger impact in the society.*

Key Words: *Machine Learning, Climate, Monitoring, Prediction, Classification, Regression, Random Forest*

1. INTRODUCTION

As the world's population increases, so do the demands on ecosystems for resources and the consequences of our global impact. Natural resources are neither indestructible nor limitless. The ecological consequences of human actions are becoming more apparent: air and water quality are deteriorating, pests and illnesses are spreading outside their historical ranges, and deforestation is worsening flooding and biodiversity loss downstream. Ecosystem services are not only restricted, but they are also threatened by human activity, as society is increasingly conscious. Long-term ecosystem health and its role in permitting human settlement and economic activities must be prioritized. There are several scientific evidence which are highly suggestive of increase in variability of climatic events. These events have proved to be a major factor affecting human ecosystem and several factors of the general economy. There are several irregularities as seen in the past which has caused disruption to many sectors that are influenced by it, with one such sector being agriculture.

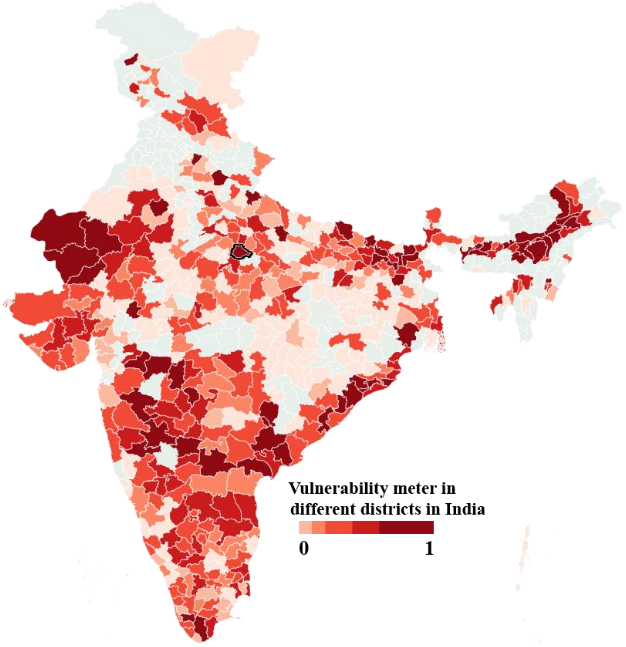


Figure 1: Vulnerability meter in several districts range from 0 – 1.

With increasing cases of sudden Events like soil eruption, sudden rainfall and wind gusts, the most affected niche is the agricultural industry. These irregularities have arisen mostly due to Global warming leading to a dire need for better and improved climate prediction and monitoring systems to be placed. There have been recurring and ambiguous patterns of natural disasters and climate fluctuations which have now been prevalent more than ever. This is a serious problem which absolutely calls for monitoring to help improve the affected economy sectors while also help reducing the human impact on the climate patterns. Climate prediction is necessary to embrace for any natural climatic deformities. Alongside these climate models is also a need for consistent, methodical, and periodical monitoring of climate and various geological events affecting it. Therefore, there is an urgent need to monitor and predict the climate changes happening all over the world due to causes identified till date.

2. METHODOLOGY

Dataset Preparation. The methodology will be demonstrated in this section in order to layout proper flow of the model as proposed to predict and monitor the irregular climate changes happening all over the world. The methodology will be consisting of various step by step mechanisms since machine learning techniques [12] will be used in order to predict changing climate. The proposed model will be end product available to the user which in our case will be Government organization responsible for the handling of the climate control over their territory. Since it will be a machine learning based model the starting point for the development is the exploration of the dataset consisting of various climate parameters along with their geographical areas. The [7] dataset by Kaggle consist of the climate changes happening all over the world in the period 1961-2019. Forecasting and predicting climate changes relies on several different factors such as rainfall, earth surface temperature.

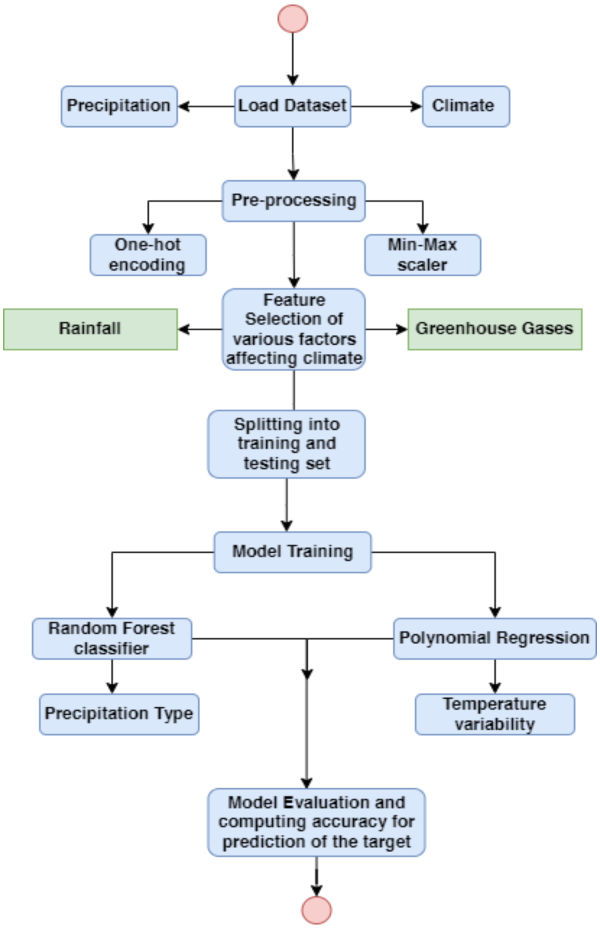


Figure 2: Flowchart of the Methodology

And rise of greenhouse effect over last few decades. So, to handle this issue and in order to analyze the climate change effectively individual datasets were utilized for the development of the model. For the factor considering the above conditions datasets [8, 9] were also taken as an input for the starting phase of the model.

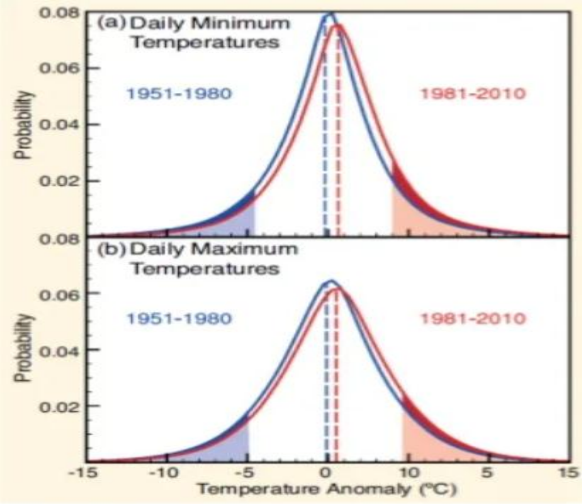


Figure 3: Extreme climate chances or events in the world.

For developing the model, we utilized google colab in which we imported several machine learning libraries like sklearn, pandas, matplotlib, seaborn and numpy based on python language.

Data Pre-processing. The data pre-processing is a very essential task since we have to handle and take care of the unrefined and coarse data and convert them into well-structured and proper information collection. These are the problems that are taken care of by preprocessing techniques such as one hot encoding, simple imputation, minmax scaling and multivariate analysis. In our concatenated dataset there are several categorical columns which can't be handled by machine learning [12] operations therefore for this we use onehot encoding which converts categorical data into numeric data. For our model, observation was made regarding the high correlation among the amount of rainfall happening in a year to the amount of floods occurring within a particular geographical region and on the basis of the multivariate analysis, we have utilized and trained the mode using the amount of rainfall within a country insertion (Figures related to rainfall, temperature rise, time series analysis.

Model Training and Testing. After pre-processing is done, model is trained using various classifiers and regression algorithms, depending on what the target outcome is in the scenario. We have considered solving and create methodologies for two type of problems, first one is in which we will be classifying about the precipitation type like rain, snow and drought in a particular geographical area. Second one is prediction of the temperature variability due to side effects of greenhouse gases [3, 4] in the past decades, which finally will lead to take preventive measures and monitor the climate in advance, by getting ready for the future has in for us. Weather data has been collated in past for historical analysis and training our time- series [10] statistical model.

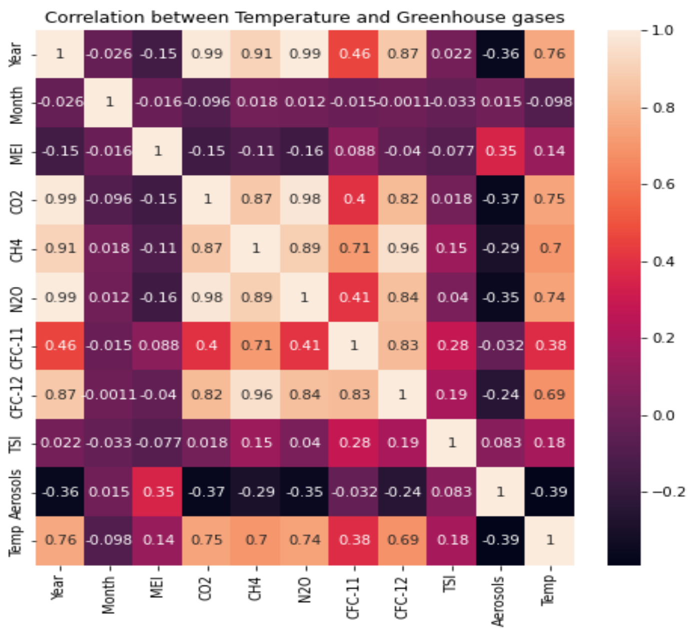


Figure 4: Correlation Matrix between the data points.

(1) Logistic Regression – In order to predict the probability of the independent variables, in our case is precipitation type, and classifying it using sigmoid activation function, logistic regression was used initially, but due to non-linearity of the attributes present in the dataset, logistic regression could not differentiate the classes, leading to need for search for another model classifier.

(2) Decision Tree Classifier – Considering each attribute of the dataset and mapping it into a parent node, following a tree like structure, decision tree was formed with each parent node having the least gini score, but in this algorithm greedy approach is followed which sometimes leads to overfitting of the training dataset ultimately effecting the testing phase and computational power.

(3) Random Forest Classifier – Extension and application of Bootstrap aggregation, random forest helps in removing the overfitting issues and considers all possible permutation and combinations that can be utilized in the training phase by integrating several decision trees, and finally averaging the probabilities, leading to the best possible target outcome that was to be predicted.

(4) Polynomial Regression for weather prediction - After knowing about the precipitation type, methodology involves using regression-based models in order to predict the temperature changes happening due to the side effects of the greenhouse gases over the past decades. Regression model, which also considers some non-linear attributes and labels is polynomial regression since it covers most of the data points which otherwise are ignored in the normal linear regression. The degree of the polynomial is what makes it a complex model, and because of that degree factor also needs to be moderated and cannot be too high or low, since it will then lead to under fitting and overfitting of the training and testing data. As indicated by certain statistics and using past research, a degree of 2-3 works best for predicting the values using polynomial regression.

3. EXPECTED RESULTS OR OUTCOME

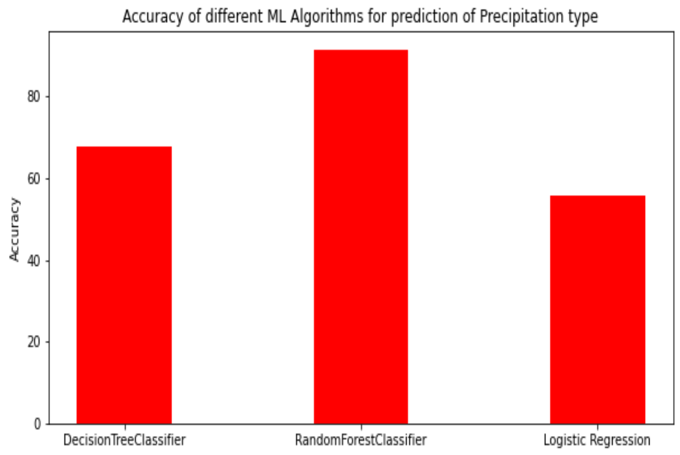
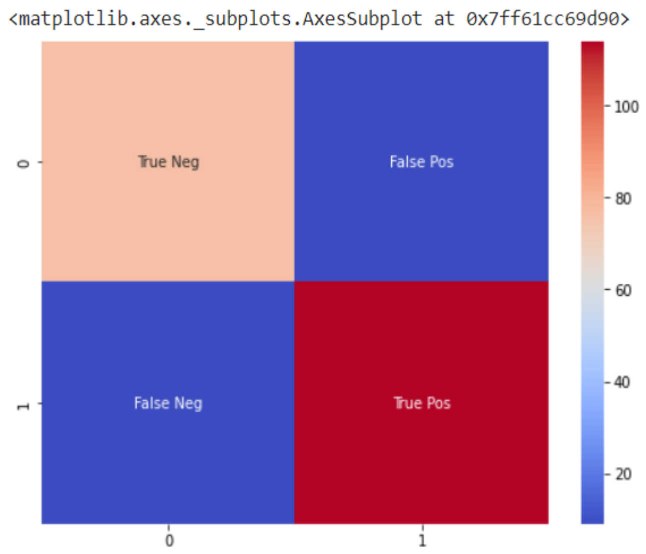


Figure 5: Accuracy Comparison graph of various classifier

On comparing the classification models, we found out Random Forest to be the best classifier with an accuracy score of about 87% as it uses ensemble learning by combining all the possible outcomes that we get from each tree and perform the max voting algorithm over it. So, in layman terms Random Forest is focusing on each and every aspect that climate is dependent on, since it is denoting every feature ranging from rainfall, soil, greenhouse gases a particular node with efficient depth. Even though Decision tree is performing better but it's overfitting the training dataset which is reducing the generalization of the prediction by focusing its attention on a particular climate attribute and thereafter we are getting a high variance which is not preferred generally in the prediction of the target feature in our case the precipitation type. Although it has a low bias but due to the higher complexity and high memory capturing power, various regression techniques were applied on it like decision tree.



```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff61cc69d90>
```

Figure 6: Confusion Matrix of the classification

Pruning and cross validation but it didn't affect the precision score as our dataset was non uniform and the time series data variation was too sparse. Logistic regression failed with a large extent with an accuracy of 51% as it classified the dependent variable by considering a threshold value of the predicted probability. This model becomes very much dependent on the threshold values set up by the researcher due to which there's a large amount of fluctuation of the True positive and true negative values ultimately affecting the confusion matrix.

4. SUMMARY AND CONCLUSIONS

Compressing the above facts and analyzing the points carefully, Natural climatic change is not much harmful for the earth but when some external force is imposed then this force affects the balance of the weather cycle and misbalances the natural flow which results in a negative impact over the biotic components surviving on earth and the abiotic factors which helps in reshaping the earth's surface and needs to be taken well care off else will destroy them completely. Predicting the climate beforehand by statistically research and analysis will help in overcoming all these problems and can force the respective organizations to make proper rules according by studying the present and past data. Due to such high studies about environment, Awareness can be spread among people regarding reducing environment pollution and how can they contribute towards making a clean and green surrounding. Concluding the points, this research will surely make a positive impact and will surely bring the natural climatic change flow back to track by teaching and motivating people what to do and what not to do which can affect the Mother Nature.

REFERENCES

[1] https://www.sciencedirect.com/science/article/abs/pii/S0308521X01000580

[2] https://books.google.co.in/books?id=dsD5UdpEOPsC&lpg=PA64&dq=climate%20prediction&lr

[3] http://nopr.niscair.res.in/bitstream/123456789/11079/1/IJTK%2010%281%29%20183-189.pdf

[4] https://www.frontiersin.org/articles/10.3389/fclim.2020.571245/full

[5] https://www.ncei.noaa.gov/access/monitoring/us-trends/

[6] https://www.ncdc.noaa.gov/climate-monitoring/

[7] https://www.kaggle.com/code/sevgisarac/climate-change/data

[8] https://www.kaggle.com/datasets/econdata/climate-change

[9] https://www.kaggle.com/datasets/muthuj7/weather-dataset

[10] https://www.researchgate.net/publication/293062099_Forecast_of_weather_parameters_using_time_series_data

[11] https://link.springer.com/article/10.1007/s42979-021-00592-x

[12] https://www.researchgate.net/publication/318338750_Supervised_Machine_Learning_Algorithms_Classification_and_Comparison