Predicting Climate Induced Floods using Machine Learning

by

Chowdhury Nafis Saleh
18101450
Farhan Alam
18101197
Md. Jawad Khan
18101268

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ch. Nafis Saleh

Chowdhury Nafis Saleh
18101450

Farhan Alam

Farhan Alam
18101197

Jawad Khan

Md. Jawad Khan
18101268

Approval

The thesis/project titled "Predicting Climate Induced Floods using Machine Learning" submitted by

1. Chowdhury Nafis Saleh (18101450)
2. Farhan Alam (18101197)
3. Md. Jawad Khan (18101268)

Of Fall, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on January 21, 2022.

Examining Committee:

Supervisor:
(Member)

Arif Shakil
Lecturer
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Climate change has been causing devastation on the economy of the country and the world as a whole. The study aims to determine how climate change would impact the frequency and severity of one of the major natural disasters which is flood. Data sets containing information about the rise of global temperatures, annual rainfall, sea level rise and flood occurrences were surveyed and assembled. Different attributes from the assembled datasets were then taken out and spliced into datasets that suit the scope of our research. A plethora of machine learning algorithms have been used to develop different prediction models based on the constructed datasets. Algorithms employed in the development of flood prediction models include: "Logistic Regression", "Decision Tree", "K Nearest Neighbors", "Support Vector Machine", "Random Forest" and "Ensemble Learning". Projection models were then trained by employing an "Autoregressive" approach for generating projection data, which were a prerequisite for the flood prediction models in making predictions of future flood incidents. And with the aid of the generated projection data, predictions of flood incidents were made for the years starting from 2022 to 2050.

Keywords: flood, prediction, climate change, machine learning

Acknowledgement

Firstly, all praise and glory be to the Great Allah thanks to whom our thesis have been completed without any major drawbacks.
Secondly, to our advisor Mr. Arif Shakil sir and also our co-advisor Mr. Tanvir Rahman sir for their kind support and advice in our work.
Thirdly, Md. Golam Rabiul Alam,PhD sir and the whole judging panel of Data Science and Big-Data Processing for their appreciative feedback on our thesis work.
And finally to our parents without whose constant support it may not have been possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

DT Decision Tree

EL Ensemble Learning

$GMSL$ Global Mean Sea Level

kNN K-nearest neighbours

LR Logistic Regression

RF Random Forest

ROC Receiver Operating Characteristic

SVC Support Vector Classifier

SVR Support Vector Regressor

Chapter 1

Introduction

1.1 Background

Natural disasters and extreme weather events have always been one of the biggest threats for human settlements both large and small. Every year, disasters like cyclones, hurricanes, wildfires, floods and tidal waves claim thousands of lives and cause nations to incur losses in the millions of dollars. But in the recent decades, humanity has had to contend with another major threat, which is anthropogenic climate change. And with the advent of this problem, for which human activity is solely to blame, the natural disasters that were already devastating to begin with are turning even more deadly in their intensity and frequency. The global average temperature has been observed to increase by about 0.2° Celsius per decade for the last few decades [1]. And in the last 30 years, the number of calamities associated with the climate has reportedly tripled. Moreover, global sea levels are rising 2.5 times more rapidly than before for the majority of the 20th century [2]. Out of 405 studies about extreme weather events, 70% found anthropogenic climate change to be responsible for making these extreme weather events more probable. And out of 81 studies looking at rainfall or flooding, 58% found that anthropogenic climate change made the cases of flooding and rainfall were more severe and for 69 studies of drought, it was 65% [3]. Predictions show that in the coming times weather extremes would get more ferocious and intense and would also infiltrate over extensive ecosystems and geological areas [4, 5]. The developing and underdeveloped regions of the world are struggling the most to deal with this issue. For example, in the case of flash floods, developing countries were identified to be the bigger victims when compared to developed countries [6]. One of the worst victims of this is Bangladesh. Due to the region's unique hydro-geographic setting and it being located in a subsided flood zone, Bangladesh has to experience floods each year from June to September in the monsoon, during which the loss usually exceeds US$20 million [7]. Africa is another such victim as the already dry and arid conditions of the continent are likely to worsen as specific regions of Africa are anticipated to further face calamities in the form of droughts and heat waves [8].

1.2 Machine Learning

"Machine learning" is a field of computer science where the learning process of humans is mimicked to teach computers to learn and derive new knowledge from

existing information. There are two major categories of learning, which are "supervised" and "unsupervised learning". In the case of "supervised learning", the computer learns to predict the right answer from a given set of answers for some given questions. On the other hand, "unsupervised learning" is the process where the computer looks for patterns in random chunks of data. The goal of machine learning, therefore, is to build mathematical models with the aid of statistics that can make inferences from given samples. The entire process of machine learning methods can be described in a few simple steps. First, a problem has to be defined. Then, a hypothesis of the model has to be formulated for solving the identified problem. After the hypothesis is formulated, we need to collect and organize datasets to use as training sets. After this, a loss function has to be specified for the model. Then comes the part of selecting the appropriate "learning algorithm" that would best serve the problem. Now, the parameters for which the "loss function" fetches the "pole hour" are observed and found. The found parameters are chosen as the "optimal parameters" of the model. Finally, the model with the "optimal parameters" is used to forecast and make estimates related to the identified problem. When it comes to making predictions related to climate induced natural disasters, model-driven and data-driven are some of the most commonly used model types. The convenience and ease of use of ML is very apparent in the sphere of creating personalized educational tools. Incorporating the location of users to predict future occurrences of extreme weather events [9] and tailoring to the learning preferences of the user [10] stands as clear indicators. Furthermore, deep learning, a field of study that falls within the domain of machine learning, has been used quite extensively in predictions and estimates related to natural disasters. For example, in historical datasets, it has been used for the classification of cyclones [11]. It has also been used for the segmentation of cyclones [12]. With the historical datasets, deep learning even contributed to the detection of cyclones [13]. Other than cyclones, deep learning has been used for the classification, detection and segmentation of tornadoes as well [14].

Chapter 2

Research Problems

As the conditions that control anthropogenic climate change continue to worsen and the climate continues to get warmer, climate induced natural disasters will become more destructive. So, it is imperative that we develop better models for estimating and forecasting these disasters. Disasters like storms, droughts, wildfires and floods are becoming stronger and more frequent [15]. Average global sea levels have increased 8 inches in the last 150 years [16], an increase which might contribute to more severe coastal floods. To cope with this ever-increasing risk, we need to understand the repercussions of climate change under different emission scenarios. A simulation comparing strong North Pacific typhoons under high "CO2" ambience found an increment by 28% in close-typhoon rainfall and an increase by 5%-11% in surface wind speeds [17, 18]. We need to look at the past impacts that the climate induced disasters have had on the world and especially the places that are most vulnerable to them, like Bangladesh. The Intergovernmental Panel on Climate Change has identified Bangladesh's low-lying coastal regions to be particularly vulnerable to climate related risks [19]. Extreme weather events are another threat that climate change poses. Although not as destructive, extreme weather events act as a major obstacle in the daily lives of people and sometimes act as an additional contributor to the destruction brought about by climate induced disasters. A study published in hurricane Harvey's aftermath suggested that due to climate change, the rainfall accompanying the hurricane was likely boosted by about 20 to 40% [20]. Moreover, escalating temperatures, shifting precipitation patterns, rising sea levels are negatively affecting agriculture, water resources, human health and the ecosystem as acute weather events are getting more frequent and severe in their intensity [21]. The rapidity at which sea levels are soaring up is hastening. During the years 2006-2015, the yearly escalation in sea levels, which was 0.14 inches per year, was found to be greater than double of that of the 20th century, which was 0.06 inches/year. A low greenhouse gas course would still prove too little, as an escalation of global sea levels by no less than 12 inches is a probable outcome by 2100. The escalation of "global mean sea level" is being driven by Global warming in two different routes. Firstly, the thawing of glaciers and polar ice caps are pouring huge quantities of water into the oceans. And secondly, the volumetric expansion of the waters brought upon by warming is causing the oceans to swell up even further. An insignificant component in this escalation is the depletion of inland water sources such as aquifers, lakes, reservoirs, rivers and soil moisture owing to groundwater pumping [22]. Presupposing a 65 cm sea level escalation by 2080, 40% of arable land is anticipated to get

submerged in southern Bangladesh. Salination of drinking water is already plaguing about 20 million people around the Bangladeshi coasts. Moreover, the poisoning of ground and surface water might be exacerbated by the upswing of strong storms and cyclones as well as the escalation of sea levels [23].

Although there are systems in place to predict or forecast natural disasters, they are not cut out for dealing with the increasingly worsening climate induced ones. And we want to tackle this problem with machine learning. Machine learning has not been fully exploited in this regard. The world has only started to realize the potential of ML in this particular field. Machine learning methodologies derive their outcomes from large amounts of data and data is something we have in abundance. There are satellite systems that are modern and much less expensive that generate petabytes of meteorological monitoring data. There are also massive climate modeling projects that generate equally large amounts of simulated climate data [24]. Not only do we have an abundance of data, we also have proven results from studies indicating that machine learning methods like "Neural Networks" constructively forecast occurrences of flood during acute weather phenomena [25], which is well within the scope of our research.

The scope of our investigation also pertained to the fact that effective usable climate data is largely available for (relatively) more developed countries such as the United States. Furthermore, most climate data do not explicitly examine the links between temperature, average rainfall, sea level rise, and flooding in a single dataset. The element of seeking effective, usable climate data for less developed and more vulnerable to natural disaster countries has been a primary focus of our research. The grounds for focusing on the considerably less developed countries were that the impact of natural catastrophes is magnified in less developed countries due to a lack of efficient post-disaster strategies. Henceforth, the need for prediction systems is more prominent in countries that are still developing. Furthermore, the goal was to organize the datasets in such a way that they were related to the Indian subcontinent's atmospheric conditions and geographical location.

Chapter 3

Research Objectives

The main aim of our study is to create a model that can correctly forecast the occurrences of climate-related natural catastrophes in the future. However, tackling climate-related natural catastrophes would be too broad of a scope to base our research on, hence we will be primarily focusing on floods. The objectives of this research are as follows:

1. To determine the prominent variables that cause floods.
2. To forecast future floods caused by climate change.
3. Research already used machine learning algorithms in predicting climate induced natural disasters.
4. Compile new datasets by dissecting and melding existing datasets as well as available data.
5. Short listing suitable machine learning algorithms for our scope by assessing accuracies.
6. Predicting future flood occurrences using our machine learning approach.
7. To realistically identify limitations of the models and existing data to ensure scope of further developments.

Chapter 4

Related Works

The research paper [26] addresses the problem of coastal flooding. The scope of this specific research was based on South Korea. As the coastal settlements grow in population and expand in its complexity, the cost of damages from coastal flooding will continue to increase. This analysis assessed the genuine risk prospect using a coastal flooding risk analysis that takes into account the yearly rainfall events and tidal levels since the risk of coastal flooding depends on tides. Estimates for the threat of coastal flooding in the future were made by taking the real rising rate of tides and forecasted rainfall data that are in accordance with local climate representations as well as multiple climate change scenarios concerning illustrative absorption routes. For this purpose, machine learning algorithms previously used in probabilistic slants, have been used for computing the flooding threat in coastal areas in a probabilistic manner. The study indicated that probabilistic approaches for analyzing coastal flooding risks are more effective when it comes to analyzing appropriate "coastal zone management". Three different "Machine learning algorithms" were used in this study [26] namely "K-Nearest Neighbor (kNN)", "Random Forest (RF)" and "Support Vector Machine (SVM)". And six variables were used for evaluating the future probability of coastal flooding. All the data regarding tides, rainfall and elevation, that were used in the model, had been converted into a 1 square km data format. The risk probability was calculated using the machine learning classifiers method and the results of the "Receiver Operating Characteristic (ROC)" curves were compared. Comparatively better outcomes were generated by the "kNN model" among all the algorithms that were used, since the "ROC accuracy" for that particular algorithm was higher than the others.

The research paper [27] deals with the problem of predicting the amount of greenhouse gases and forecasting the average temperature in 10 years, next to come, from values previously measured over India. For this, the study [27] used and evaluated the performances of several machine learning algorithms to see which best fit its particular use case. The primary challenge for the research was the creation of a statistical data model on large datasets that is dependable and well organized, one that could correctly correlate between the yearly average temperature and some of the factors that can affect it, like, the concentration of "Carbon Dioxide (CO2)", "Nitrous Oxide (N2O)" and "Methane (CH4)". The study found CO2 to be the largest contributor for temperature fluctuation, with CH4 being the second largest and N2O being the smallest. The machine learning algorithms used in this study

[27] were "Support Vector Regression", "Multi- Regression Tree", "Linear Regression" and "Lasso". Among all these, "Linear Regression" was found to be the most accurate at predicting greenhouse gases and temperatures. Predicting the temperature and concentration of greenhouse gases for the next 10 years was one of the two main purposes of this study. And the other was to make a "graphical interface" established on the prediction that could present the findings in a way that is easy to understand. The findings of the study indicated that the conditions responsible for the increase in global temperatures will likely continue to exacerbate, meaning that global temperatures will keep on increasing in a linear fashion which could prove to be lethal.

Furthermore, in another paper [28], datasets of existing climate model simulations have been used to understand the "short-term" and "long-term" changes in temperature in response to the contrasting anthropogenic emission framework. A surrogate model was used to map short-term response patterns to "long-term" ones contained in a given "Global Climate Model (GCM)". Then in the following steps, the surrogate model was used to accelerate the prediction of short-term scenario outputs. The surrogate model dealt with parameters including Carbon Dioxide (CO2), Methane (CH4), Sulphate (SO4) and black Carbon particles. Here [28], two different machine learning algorithms were compared against traditional pattern scaling approaches. The machine learning methods used were Ridge Regression and Gaussian Process Regression (GPR). The overall advantage of the mentioned two machine learning methods was prominent in case of capturing regional patterns and diversity compared to pattern scaling. When further comparisons were conducted, it was found in this particular research that GPR errors were lower than Ridge Regression.

Another research paper [6] highlighted flood susceptibility using machine learning models. The paper explored the rise in priority of the flood management strategies in the scope of Bangladesh. Emphasis was given on the aspect of flash floods occurring every year throughout the country for the research materials. While four prominent types of floods were addressed including "flash floods", "riverine floods", "coastal floods" and "urban floods"; "flash floods" were identified as the most devastating type. A "Geographic Information System (GIS)" environment was used on 413 current and former flooding points. A "GIS" is a computer system that is used to encapsulate, record, examine and represent data related to coordinates on Earth's facet. In this study, the "machine learning algorithms" used are "Dagging" and "Random Subspace (RS)" combined with "Artificial Neural Network (ANN)", "Random Forest (RF)" and "Support Vector Machine (SVM)". The findings of the particular study [6] showed that the highest class consisted of the smallest area (14.03%) and this was followed by the "low (14.89%)", "high (15.61%)", "moderate (23.78%)" and "very low" classes. For the "SVM model", the area percentages are 30.20%, 16.22%, 13.32%, 16.88% and 23.39% for the "very low", "low", "moderate", "high" and "very high" classes respectively. For the "RS", "RF" and "ANN" models, the area percentages are 21.70%, 29.93% and 33% respectively for "very-high-flood-susceptible" classes. In the case of the "Dagging hybrid model", the highest area was taken by the class of "very-high-susceptibility" and the "very-low-susceptibility" class occupied the smallest area. For the "SVM model", the maximum area belongs to the very "high susceptibility" class whereas the minimum area came under the

"moderate-susceptibility" class. For "RS", "RF" and "ANN" models, 1020, 1090 and 90 square km of the total area have been found belonging to the class of "very-high-susceptibility" while 780, 1130 and 1230 square km have been found belonging to the "low susceptibility" class. When it comes to the proportion of area belonging to the "high" and "very high" classes, all the models seemed to have similar spatial disposition patterns for areas with flood susceptibility. The only difference being that for the "Dagging model", the area percentages are relatively minor. This means that the "Dagging model" can likely generate more practical outcomes with greater reliability and accuracy. "ROC" has been used a metric for judging the authenticity of the models. And "Dagging" performed best in the "ROC curve".

The research paper [29] did a comparative study that focused on the development and comparison of several machine learning models that could be used to forecast rainfall in different scenarios and time horizons. The study took into consideration the apparent link between climate change and the frequency and occurrence of extreme rainfall. The paper also talked about the need for accurate ML forecasting models for preventing natural disasters. The primary goal of this analysis [29] was to differentiate and implement multiple machine learning models like "Neural Network Regression (NNR)", "Boosted Decision Tree Regression (BDTR)", "Bayesian Linear Regression (BLR)" and "Decision Forecast Regression (DFR)" for predicting rainfall in the Terengganu region of Malaysia. Two methods were used to anticipate the rainfall. They are: "forecasting rainfall using Autocorrelation Function" and "forecasting rainfall using projected error". Among all the machine learning models used, the findings of the research identified "BDTR" to be the most accurate since it best predicted "weekly average errors" to correct "weekly average projected rainfall".

Another research paper [30] discussed the threat that tropical cyclones pose to coastal areas. It pointed out how the combination of heavy rainfall, strong winds and storm surges can generate flooding of massive scales. The study [30] conducted research on a machine learning model to analyze its predictive skill of seasonal estimates of tropical cyclone counts for the "North-Atlantic-region". The competence of the mentioned model was measured in relation to an established tropical cyclone prediction model that was developed and kept by "Colorado State University (CSU)". The main aim of this study was to minimize seasonal errors in forecasting tropical storms and it was achieved by applying the Support Vector Model (SVR) algorithm. Errors from this model were compared with the errors from the already existing model of CSU. The comparison indicated that the CSU model had four large errors whereas the SVR had two. The findings also indicated that the SVR model had an 80% improvement over the CSU model.

Chapter 5

Methodologies

5.1 Workflow

Firstly, for our research, we have scoped a variety of sources for our datasets including already published papers, “kaggle.com”, “en.climate-data.org”, “wbwaterdata.org”, “datahub.io”, “datacatalog.worldbank.org”, “ourworldindata.org”. Finally, for our Temp_Humidity_Rainfall dataset we have produced the dataset by jointly choosing and arranging data from “en.climate-data.org” and then storing the data into a Temp_Humidity_Rainfall.csv file. All the climatic data of “en.climate-data.org” are based on “ECMWF-Data”. The model utilized by “en.climate-data.org” contains data points with a resolution of 0.1-0.25 grade which are in excess of 1.8 billion. All the tables, graphs and datasets provided on “en.climate-data.org” can be used under the “Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license”. Attribution requires a valid Link to the Website the Data or Graphs are taken: e.g. “Climate Data London” from “Climate-Data.org”. Location related data for all the states/cities are collected from the “OpenStreetMap” project. “OpenStreetMap” is licensed under the “Open Data Commons” along with “Open Database License (ODbL)”, and hence is open data. The map tiles are licensed under the “Creative Commons Attribution-ShareAlike 2.0 license CC BY-SA” according to “https://en.climate-data.org/info/sources/” and “https://en.climate-data.org/info/licensing/”. For our other dataset ‘Kerala.csv’ we have found the prominent dataset from “Kaggle.com”. This particular dataset comprises of records of the monthly rainfall index of Kerala from 1900- 2018.

In the case of our third dataset, we investigated the sea-level-rise dataset and global-mean-temperature dataset from ”datahub.io”; we had to choose the relevant data pertaining to our study purpose from both of the aforementioned chosen datasets. After that, we were able to create the SeaLevelRise_Temp_Rainfall_Flood dataset by assembling and assimilating the rainfall data from kerala.csv and putting it against the respective period of incidence.

The raw data sources for the SeaLevelRise_Temp_Rainfall_Flood dataset from the “datahub.io” are made available under the terms of the ”Public-Domain-Dedication-and-License-v1.0-ODC-PDDL-1.0”. The ”Public Domain Dedication and License” grants individuals the ability to distribute and utilize specified data with change for any purpose and without constraints.

Secondly, we performed data preprocessing. This phase entails performing preliminary analysis of our accumulated datasets to note down outliers, random variables and presence of empty data cells along with activities relating to filtering and resizing the data. The dataset is split into features and labels. This operation would be crucial for the overall cleansing of our datasets. Moreover, the data preprocessing plays a critical role in ensuring more accurate results.

The third key phase comprised the identification to the features and labels for our datasets. In our Temp_Humidity_Rainfall dataset the amount of rainfall in mm was selected as the target variable. On the other hand, in our Kerala dataset the incidence of flood was selected as the target variable.And lastly, for our SeaLevelRise_Temp_Rainfall_Flood dataset we had selected GMSL , GMSL uncertainty , Annual Rainfall , Annual Temperature Mean as our features while the occurrence of Flood as our target variable. After the discovery of the appropriate features and labels for our datasets, train and test splits of our datasets were made.

Now, according to our information obtained from paper evaluation relevant to our research issue and other sources we found some prominent machine learning algorithms suitable for our instance. For our Temp_Humidity_Rainfall dataset we employed "Linear Regression", "Ridge Regression", "Support Vector Regression", "Decision Tree Regression" and "Random Forest Regression". On the other side for both our Kerala and SeaLevelRise_Temp_Rainfall_Flood datasets we utilized "K-Nearest Neighbor Classifier", "Logistic Regression Classifier", "Decision Tree Classification", "Support Vector Machines Classifier", "Random Forest Classifier" and "Ensemble Learning Classifier". After that, we recognized the accuracy of all of our used models and then by comparing the results we arrived at the conclusion of which models would be best suited for additional future research.

With the comparison data in hand, we went on to the next stage of our investigation. The next crucial step was to make out-of-sample-predictions. The predicted outcomes of the models were then observed in contrast to real world flood incidence.

Finally, we prepared a projected dataset by administering an "Auto-Regression" model. Then we used our trained classifiers to make predictions on the projected outcomes. The projected data will be used for flood forecasts in the future as well.

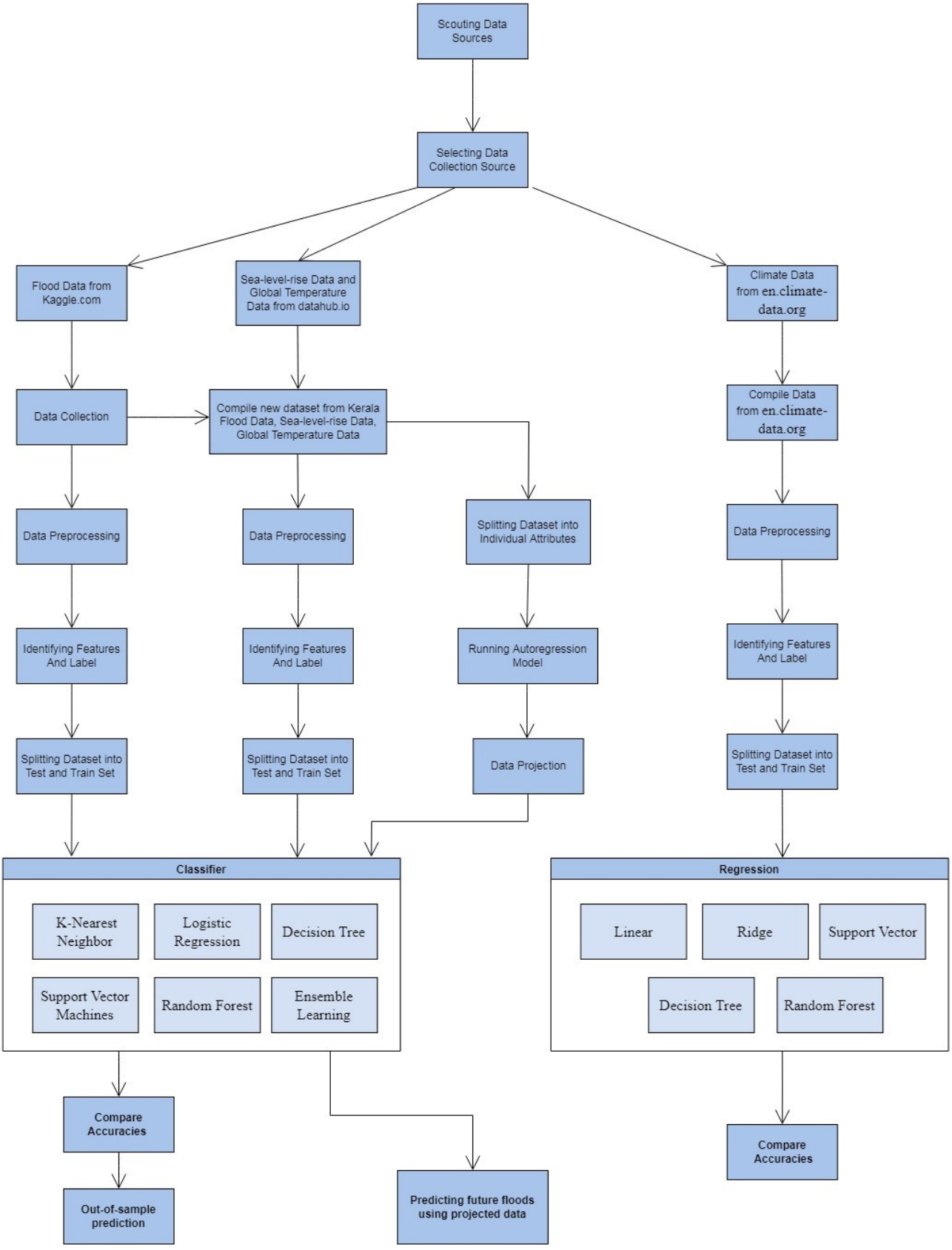

Figure 5.1: Workflow Diagram

5.2 Input Data

There is a scarcity of available datasets when it comes to climate induced natural disasters and their predictions. And as the worsening climate continues to alter the environment by influencing the weather, land and water conditions, people around the world will need open and clear data to help them assess the various climate-caused risks due to issues like sea level rise and flooding [31]. We scoured the internet for any suitable datasets that could be used for our particular use case, which was to train our models, and we only succeeded in obtaining four that could be tweaked or manipulated to suit our goal. All the datasets that we used are publicly available. Some of the publicly available datasets about climate induced natural disasters are: "Urban Flooding Of Greater Dhaka In A Changing Climate - Building Local Resilience To Disaster Risk", "Sea Level Rise Maps (2020s 100-year Floodplain)", "Rainfall in Pakistan", "Historical Rainfall data of Bangladesh (1970-2016)", "Rainfall in India", "Rainfall Data from 1901 to 2017 for India", "Rainfall in Kerala 1901-2017" etc. One of the datasets that we used is from "en.climate-data.org" and the other is from "Kaggle.com".

The dataset from Kaggle contains information about the incidence of floods in relation to the average monthly and yearly rainfall for the state of Kerala in India. We selected this dataset because it had clearly assigned labels about the occurrence of floods which proved to be really helpful for our classification models. There are 118 samples for sixteen attributes, of which twelve are the average monthly rainfalls for each month of the year and the other ones are: state, year, annual rainfall and the incidence of flood.

The dataset from "en.climate-data.org" contains information about the average rainfall, humidity and temperatures during every single month of the calendar year over a span of states in India, including Kerala, that were collected from the year 1999 to the year 2019. This dataset was selected because we found it to be quite suitable for training our regression models. There are 180 samples for nine attributes: "state", "place", "month", "average temperature", "minimum temperature", "maximum temperature", "humidity", "average sun hours" and "average rainfall".

The rainfall and temperature related dataset from en.climate-data.org were used to correlate between the rise of global temperatures and the occurrence of heavy rainfall events. This can be confirmed by a number of studies and observations. Climate scientists anticipate that as global temperatures rise, substantially more rain will fall in severe storms. The moisture in the air fluctuates due to changes in temperature; heating the air by 1°C, and it can retain roughly 7 percent more water. Moreover, a climate model predicting daily precipitation fluctuations imply that if the world warms by 3° C, most land regions will receive considerably higher rainfall [32]. The dataset that we got from Kaggle about flooding incidents in the state of Kerala was used to correlate between heavy rainfall events and the occurrence of floods. Floods tend to occur on the one calendar day with the greatest precipitation i.e the wettest day of the year. The analysis by "Climate Central" determined how the previously mentioned rainy days are trending over a span of time by examining data for 244 locations across India. As the climate became warmer, rainfall peaks have seen an

increase in most areas. Frequency of severe rainfall events as well as the strength of said events are increasing when compared to the ones in past times. It was found that the reports of top 1% of occurrences of rain were made incommensurately lately for 80% of the examined cities. Climate change caused the maximum precipitation to be 15% greater in intensity and three times more frequent at the very least during "Hurricane Harvey". The record rain in California during February was due to a sensation exacerbated by climate change, which is known as an "atmospheric river" [33].

Through these correlations we want to establish a connection between the rise in global temperatures, global sea level rise and the occurrence of floods. According to "The Royal Society", severely hot and dry zones are created due to the increase in quantity of evaporation, which in turn is caused by the greater worldwide temperatures brought on by climate change. However, the warmer air also allows for a larger capacity for moisture absorption, which results in thicker, rain-soaked clouds, which subsequently unleash all that pent-up precipitation in one fell swoop. Thus we can see, the relationship between floods, droughts, and other extreme weather occurrences is organically related to climate change [34].

5.3 Dataset Assembly and Manipulation

Some of the datasets were assembled and then manipulated because we could not find any dataset with the attributes that aligned with the scope of our research. We built the Temp_Humidity_Rainfall dataset by assembling monthly meteorological data such as "average temperature", "minimum temperature", "maximum temperature", "humidity", "average sun hour" and "average rainfall" from "en.climate-data.org" for every region of the states of Kerala, Karnataka and Tamil Nadu from the year 1999 to the year 2019. All the values for the required attributes were accumulated and saved into a csv file.

We gathered two datasets from "datahub.io", one on sea-level rise and another on global temperature. The dataset on sea-level rise consists of values for "GMSL" and "GMSL" uncertainty from the year 1880 to the year 2013. And the dataset on global temperature contains values for global mean temperature from the year 1880 to the year 2016. The SeaLevelRise_Temp_Rainfall_Flood dataset has the values for "GMSL", "GMSL" uncertainty, annual rainfall, global mean temperature and flood occurrence from the year 1901 to the year 2013. For building the SeaLevelRise_Temp_Rainfall_Flood dataset, we have assembled the values for "GMSL" and "GMSL" uncertainty from the sea-level rise dataset; and we have amassed the values for the global mean temperature from the global temperature dataset which were sourced by "GCAG". The values for annual rainfall and floods, on the other hand, were put together from the Kerala dataset. "GMSL" stands for "Global Mean Sea Level". The "Global Mean Sea Level" is the mean elevation of the ocean's surface obtained from satellite altimeter readings throughout the entire expanse of the earth [35]. "GMSL" uncertainty is the accidental deviations in the values of "GMSL" brought on by the reconstruction of the model used to produce the values. "GCAG" is a reliable tool that yields real-time evaluation of global monthly and annual temperatures with the aid of data extracted from the "Global Historical Climatology

Network-Monthly" and "International Comprehensive Ocean-Atmosphere" datasets [36].

Temp_Humidity_Rainfall Dataset: https://drive.google.com/file/d/1fzXqA8_WpVajZGdsxGANfgPEOUY_2sqi/view?usp=sharing

Kerala Dataset: https://drive.google.com/file/d/1z26U-RNrnyAcowsJQOBT9LXuOTl16Jhc/view?usp=sharing

SeaLevelRise_Temp_Rainfall_Flood Dataset: https://drive.google.com/file/d/1R7Hk43lsu6r09FFD_YXzQqQyFI9CiJ-Z/view?usp=sharing

Each of the datasets served a different purpose in our research. The Temp_Humidity_Rainfall dataset was primarily used to find the correlation between temperature, humidity, sun hours, and the amount of rainfall. This correlation was made to find out how the different attributes influence one another. This was important for furthering our research as we initially didn't know how the various features and components of our regression models would impact their outcomes.

Kerala is an infamous disaster-prone area located at the south-western coast of India. As a result, there is a relative abundance of meteorological data related to the occurrence of climate-induced natural disasters. And since we wanted to keep our focus mostly on the Indian subcontinent, Kerala was the only place we could look into for data on climate-induced natural disasters. Among all the disasters that Kerala is a victim of annually, Floods are one of the most commonly occurring ones. This also served our purpose of wanting to primarily concentrate on floods for our research. The instance of the 2018 Kerala flood, which was considered the most devastating flood incident of the last 100 years span [37] stands as one of the many solid examples of the dire situation. Every year starting from 2018 to the year 2021, the area has been a victim of regular flooding [38]. Incidents of extreme rainfall is a major contributor to this problem. To name a case, the year 2021 had the highest recorded rainfall in Kerala in the last 60 years [39]. The damage and destruction brought on by the repeated floods are also substantial. The human cost, for example, can turn out to be a grievous loss. One of such occurrences was in 2018, when over 400 people lost their lives after the state was flooded by heavy rainfall [40]. Another such occurrence was in the flood of 1924, during which an estimated 1000 people died, and to make things even worse, the rural masses lost arable areas as well as livestock [41]. Taking all of these into account, we inferred that Kerala is an appropriate place to focus our research on. And so, the Kerala dataset, which contains data about flood occurrences and the annual and monthly precipitation, was employed.

The SeaLevelRise_Temp_Rainfall_Flood was a critical dataset for our research purpose. This particular dataset was used to tie up all impacts of the temperature rise, rainfall leading to the rise in sea level correlating to the occurrence of floods. Correlating between these attributes makes sense for a lot of reasons. The impact of heavy rainfall on flood incidents is already quite well known. For example, there is an entire category of floods termed flash floods that occur due to incidents of ex-

treme rainfall events [42]. The steep elevation in global temperatures directly alter sea-levels by pouring large quantities of water, resulting from the thawing of polar sea caps and glaciers, into the oceans of the world [22]. The escalation of sea levels, in turn, influence the occurrence of floods. For instance, escalating sea levels may lead to the flooding of wetlands [43]. Moreover, it has been predicted that approximately 22% of all the coastal wetlands will be drowned due to escalating sea levels [44]. Consequently, the coastal wetland areas will flood at a much higher frequency [45]. This is increasingly relevant in the context of our subject region, Kerala, because almost a fifth of that state comprises wetlands [46]. And of the total area blanketed by the wetlands, 25.45% are coastal wetlands [47]. So, the escalating sea levels would have an immediate impact on the incidents of flood in Kerala.

5.4 Data Pre-processing

When collecting data for training algorithms, it's very unlikely that the data will be properly arranged and processed to be fed directly in. So, for any machine learning application, the data preprocessing phase is one of the most important ones. For instance, there are algorithms that require the data to be in the same scale to maximize their learning efficiency and accuracy [48].

We have used both Regression and Classification algorithms to train our models. And the pre- processing steps for classification and regression are different.

For Classification (on Kerala.csv and SeaLevelRise_Temp_Rainfall_Flood.csv datasets), we have to handle empty cells, encode categorical features and scale the feature values. We handle any case of empty cells or columns by imputing the missing values accordingly. The desired label in each of the aforementioned datasets is the incidence of flood. In our case, the classification is a binary classification problem where the labels are "YES" and "NO". These string values are encoded to 1 and 0 through binary encoding. Scaling is done on the values of the features to account for any kind of large variance disparities, which could negatively impact the prediction capabilities of the algorithms.

For Regression (on Temp_Humidity_Rainfall.csv), we only have to handle missing values and encode categorical features. The missing values here are also handled through inputting. The encoding step differs in the sense that it's not a binary encoding and the values for the features month, state and place are encoded manually by assigning integer values of 1-12 for month, 1-3 for state and 1-15 for place.

Dataset Before Preprocessing (raw data) on Kerala.csv

	SUBDIVISION	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	ANNUAL RAINFALL	FLOODS
0	KERALA	1901	28.7	44.7	51.6	160.0	174.7	824.6	743.0	357.5	197.7	266.9	350.8	48.4	3248.6	YES
1	KERALA	1902	6.7	2.6	57.3	83.9	134.5	390.9	1205.0	315.8	491.6	358.4	158.3	121.5	3326.6	YES
2	KERALA	1903	3.2	18.6	3.1	83.6	249.7	558.6	1022.5	420.2	341.8	354.1	157.0	59.0	3271.2	YES
3	KERALA	1904	23.7	3.0	32.2	71.5	235.7	1098.2	725.5	351.8	222.7	328.1	33.9	3.3	3129.7	YES
4	KERALA	1905	1.2	22.3	9.4	105.9	263.3	850.2	520.5	293.6	217.2	383.5	74.4	0.2	2741.6	NO

Figure 5.2: Dataset Before Preprocessing

Dataset After Preprocessing on Kerala.csv

	YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
0	0.000000	0.343713	0.565823	0.237218	0.653179	0.176977	0.696472	0.423473	0.146603	0.322208	0.397277	0.955702	0.238872
1	0.008547	0.080240	0.032911	0.263473	0.314807	0.118325	0.215332	0.763429	0.112431	0.927689	0.580497	0.379527	0.600396
2	0.017094	0.038323	0.235443	0.013819	0.313473	0.286402	0.401376	0.629139	0.197984	0.619077	0.571886	0.375636	0.291296
3	0.025641	0.283832	0.037975	0.147858	0.259671	0.265976	1.000000	0.410596	0.141932	0.373712	0.519824	0.007183	0.015826
4	0.034188	0.014371	0.282278	0.042837	0.412628	0.306245	0.724872	0.259750	0.094239	0.362382	0.630757	0.128405	0.000495

Figure 5.3: Dataset After Preprocessing

Chapter 6

Selected Algorithms

Decision Trees aim to figure out the classification or value of a variable and it does so by learning some uncomplicated decision laws extrapolated from a given set of training data. It can be used for solving both "classification" and "regression problems".

"Random Forest" is a "supervised learning algorithm" that uses an ensemble of "decision trees" to make inferences from a given dataset. It makes trees that are dependent on each other by penalizing accuracy. Similar to "Decision Tree", It works with both "classification" and "regression" models.

"K-Nearest neighbor" is a "supervised learning algorithm" that is applied to both "regression" and "classification" assignments. It works by employing the principle assuming that similar values will be closer in comparison to different ones when observed from a particular node.

"Support Vector Machine" is a "supervised linear learning algorithm" that is applicable for both "classification" and "regression" tasks. It can solve linear as well as non-linear problems. "SVM" generates a line or hyperplane which is used to divide the data into categories. It produces a line as output to the given input data that distinguishes the available categories. The nearest points to the line are discovered by "SVM" from all categories. These points are known as "support vectors".

"Ensemble learning" is an approach that tries to maximize predictive performance by combining the results from multiple predictive models. Some well-known strategies include "Majority Voting", "Simple Averaging", "Weighted Averaging" etc. In our study, we employed an "Ensemble Voting Classifier" while implementing the flood prediction models.

"Autoregression" is a distinct form of "Regression model" that uses values of a single attribute from different points in time to build a linear function that predicts the values for that same attribute. This is dissimilar to regular "Regression models" with the nuance that they rely on dependent features to map a linear function to govern the values of a single quantity.

"Decision Tree", "KNN", "SVM", "Random Forest", and "Ensemble Learning" all

being non-linear in nature, proved to be useful in the final flood prediction model due to the label having non-linear association with most of the features. “SVM”, “DT” and “Random Forest” algorithms were also employed in the regression models that were developed to judge the effect of temperature on rainfall. Simple linear algorithms like “Logistic Regression”, “Linear Regression” and “Ridge Regression” were also used in their respective use cases: “Logistic Regression” in the “Classification” models; “Linear Regression” and “Ridge Regression” in the “Regression” models.

Chapter 7

Implementation and Analysis

7.1 Data Correlation

For the Temp_Humidity_Rainfall.csv dataset, we used a built in dataframe method from the pandas framework to generate the correlation matrix. The correlation matrix shows how strongly one variable influences the other. The result is as follows:

	Avg Temp In Celsius	Humidity In Percentage	Avg Sun Hours	Rainfall In mm
Avg Temp In Celsius	1.000000	-0.439684	0.624362	0.054884
Humidity In Percentage	-0.439684	1.000000	-0.923984	0.845306
Avg Sun Hours	0.624362	-0.923984	1.000000	-0.663181
Rainfall In mm	0.054884	0.845306	-0.663181	1.000000

Figure 7.1: Correlation Matrix of Temp_Humidity_Rainfall.csv Dataset

Here, we see that rainfall has a strong direct correlation with humidity and a strong inverse relationship with average sun hours. On the other hand, it has a weak direct correlation with average local temperature. We wanted to see if these correlated variables could be used to predict the monthly average rainfall. Hence, we plugged these features into a selected array of regression algorithms to see how they perform.

With the SeaLevelRise_Temp_Rainfall_Flood dataset, we figured out the correlation between the "global mean temperature" & the "Global Mean Sea Level". Similar to the previous dataset, a correlation matrix was found for the two attributes. The matrix is as follows:

	GMSL	Mean
GMSL	1.000000	0.915166
Mean	0.915166	1.000000

Figure 7.2: Correlation Matrix of SeaLevelRise_Temp_Rainfall_Flood.csv Dataset

This was graphically represented by plotting a curve of the two quantities.

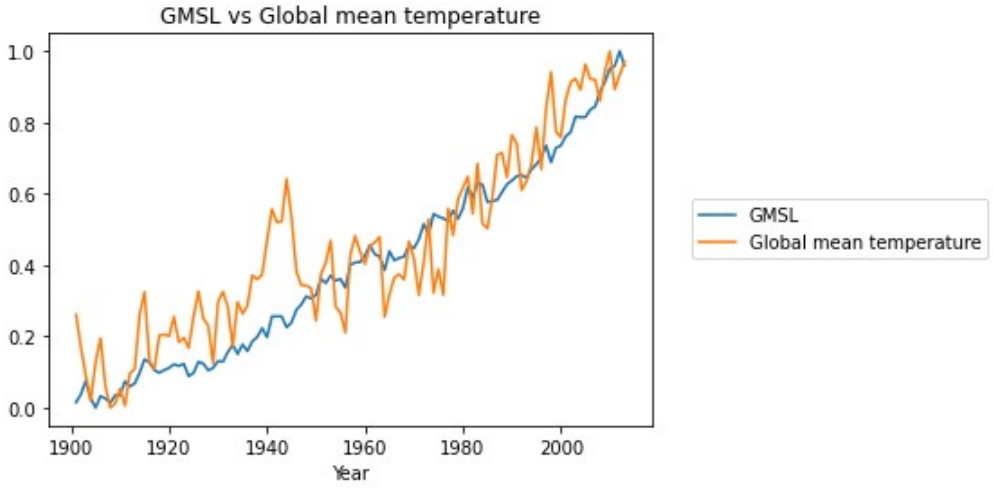

Figure 7.3: Correlation Graph of SeaLevelRise_Temp_Rainfall_Flood.csv Dataset

We noticed that there is a very strong correlation between the two quantities. This correlation was used later to make the flood prediction classification model work.

7.2 Model Implementation and Accuracy

As the field of AI and machine learning has grown, a lot of different algorithms have been developed for tasks across multiple disciplines. With so many different learning methods at our disposal, it is important that we select the proper method. No one algorithm has any inherent advantage over all the others in its functional domain. For problem specific applications, it's important that multiple different algorithms are compared and the best one is selected. "Classification accuracy" is used as the main yardstick for such identifications [48].

In a study [27] of forecasting greenhouse gases and temperature, Linear Regression was shown to be the most accurate among the tested "Linear Regression", "Multi Regression Tree", "Support Vector Regression" and "Lasso". Furthermore, in coastal flooding prediction [26] utilizing "K- Nearest Neighbor (KNN)", "Random Forest (RF)" and "Support Vector Machine (SVM)"; "KNN" fared the best. Moreover, in another case [30] comparing CSU model versus SVR model in forecasting tropical storm numbers for the North Atlantic area revealed "SVR" to easily beat CSU.

Thus, armed with the knowledge acquired from observing studies conducted and papers published we focused on the implementation of certain prominent classification and regression models with closely ties in with our study. The classification algorithms that we investigated are "KNN", "Logistic Regression", "Decision Tree Classifier", "Random Forest Classifier", "Support Vector Classifier" and "Ensemble Learning". And the Regression methods that we explored are "Linear Regression", "Ridge Regression", "Support Vector Regression", "Decision Tree" and "Random Forest Regression".

We used the sklearn library of Python to implement the models. After pre-processing, the dataset was split at a ratio of 8:2 into "training" and "testing data". The "training data" was then fed into each of the algorithms to train the models and their performance was measured by an accuracy score in case of both the regression and classification models. Accuracy scores are usually measured by comparing the predicted values with the actual values. The accuracy scores for all the algorithms were then plotted against each other on a bar chart. We also measured the "ROC scores" and plotted an "ROC curve" for the classification models. A "ROC curve" is defined to be a representation of the "true positive rate (Sensitivity)" as a function of the "false positive rate (100-Specificity)" for contrasting "cut-off points" of a specification. The "sensitivity/specificity pair" is highlighted by every single point on the "ROC curve". The "sensitivity/specificity pair" correlates to a particular decision "threshold". The "Area Under the ROC curve (AUC)" is an estimation of how effectively a parameter can discriminate between two feature sets. In other words, the "ROC curve" is a basic apparatus for feature test interpretation [49]. We have seen that the "ROC curve" plots two variables: The "True Positive Rate" and the "False Positive Rate".

True Positive Rate(TPR) is specified below:

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate(FPR) is specified below:

$$FPR = \frac{FP}{FP + TN}$$

TP = "number of true positives"
FP = "number of false positives"
TN = "number of true negatives"
FN = "number of false negatives"

7.2.1 Comparison of Regression Models

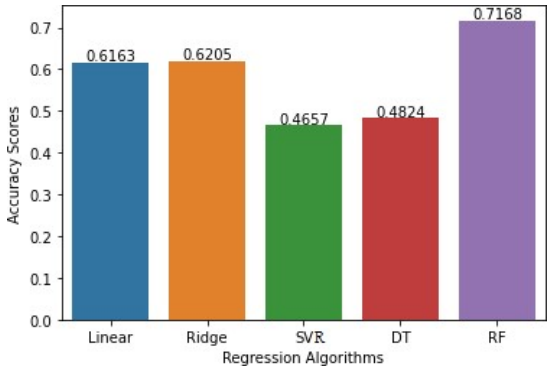

Figure 7.4: Accuracy Score of Regression Model

7.2.2 Comparison of Classification Models on Kerala.csv dataset

Algorithm	ROC Scores
"kNN Classifier"	81.818182
"Logistic Regression"	86.363636
"Decision Tree Classifier"	61.188811
"Random Forest Classifier"	87.762238
"Ensemble Learning"	90.909091
"Support Vector Classifier"	95.454545

Table 7.1: Comparison of Classification Models on Kerala.csv dataset

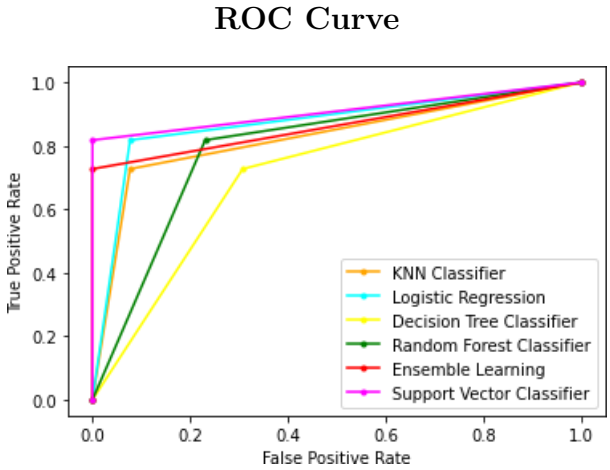

Figure 7.5: ROC Curve for Kerala.csv Dataset

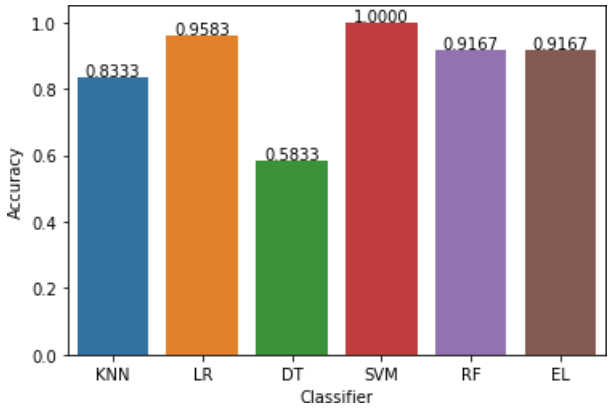

Figure 7.6: Accuracy Scores of Classification Models for Kerala.csv Dataset

7.2.3 Comparison of Classification Models on SeaLevelRise_Temp_Rainfall_Flood.csv dataset

Algorithm	ROC Scores
"kNN Classifier"	86.742424
"Logistic Regression"	91.287879
"Decision Tree Classifier"	95.833333
"Random Forest Classifier"	95.833333
"Ensemble Learning"	86.742424
"Support Vector Classifier"	86.742424

Table 7.2: Comparison for SeaLevelRise_Temp_Rainfall_Flood.csv dataset

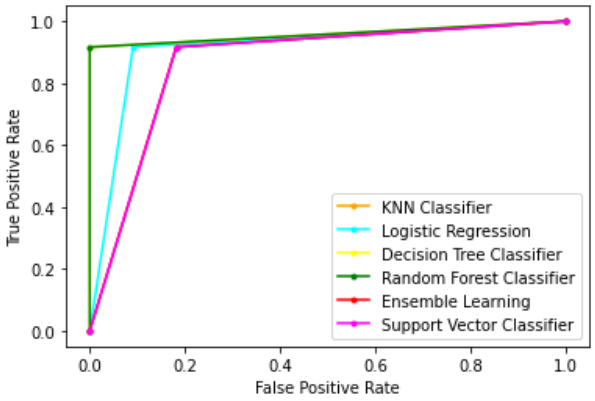

Figure 7.7: ROC Curve for SeaLevelRise_Temp_Rainfall_Flood.csv Dataset

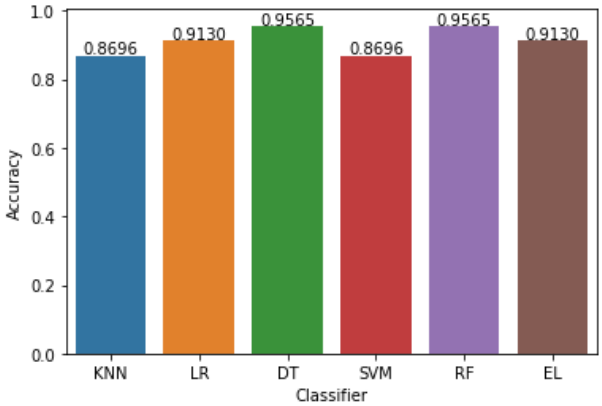

Figure 7.8: Accuracy Scores for SeaLevelRise_Temp_Rainfall_Flood.csv Dataset

7.2.4 Accuracy of Projection Models

Considering that our aim was to predict climate-induced flood incidence, it would have been a challenge to find projected data to run our model on in the future. To rectify that, we built and trained our own models for generating projected data.

These models were then used to populate the testing dataset that we used to evaluate the prediction outcomes in future scenarios. The models in question were trained using an "Autoregression" approach. The working principle of these models are such that they can predict future outcomes for a variable for its values from earlier points in time. In our case, projections were made for "GMSL", "GMSL uncertainty", annual rainfall, and "Global mean temperature". These projected values were then put into the classification model trained for flood prediction to get an idea as to what estimate our model comes up with.

Accuracy Graphs of Autoregression Models(1901-20)

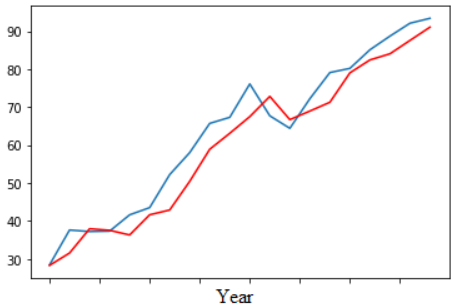

Figure 7.9: GMSL

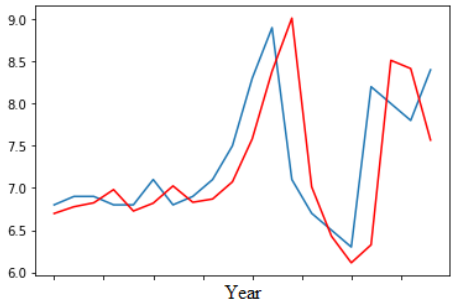

Figure 7.10: GMSL Uncertainty

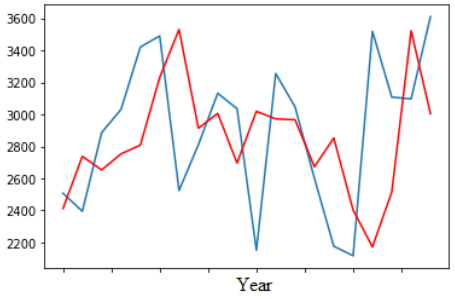

Figure 7.11: Annual Rainfall

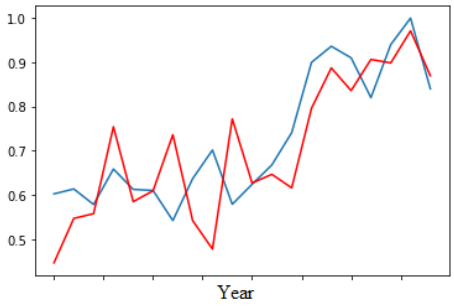

Figure 7.12: Mean Temperature

We observe that the accuracies of the models are decent to be used in future predictions. Due to this high accuracy, we compiled all the projected values into a single dataset.

Projection dataset:
https://drive.google.com/file/d/1LMhUtPlnxfywhJaAsFBXuq72FUBPnqiX/view?usp=sharing

7.3 Out-of-Sample Predictions

To evaluate the efficacy of the trained models, we checked them against real world scenarios from the year 2018 to the year 2021 with values outside of our original test dataset. We collected information about "GMSL" and "GMSL uncertainty" for all the years from the World Meteorological Organization [50]. The global mean temperature, on the other hand, was collected from "GCAG" [51]. We collected the annual rainfall and flood incidence data for the years 2019 and 2020 from a single source [38]. We found out that in 2021, Kerala had the highest recorded rainfall in 60 years, which also yielded the information about the annual rainfall for that year [39]. We also learned that before 2021, 2018 had the highest recorded rainfall in that time period, which gave us the information about that year's annual rainfall [52]. Finally, we discovered that all the years from 2018 to 2021 had some incidence of flooding [53].

Time	**Predicted Outcome**						**Actual Outcome**
	"kNN"	"LR"	"DT"	"RF"	"EL"	"SVC	
2018	Flood	Flood	Flood	Flood	Flood	Flood	Flood
2019	Flood	Flood	Flood	Flood	Flood	Flood	Flood
2020	Flood	Flood	Flood	Flood	Flood	Flood	Flood
2021	Flood	Flood	Flood	Flood	Flood	Flood	Flood

Table 7.3: Out-of-Sample Predictions

7.4 Future Flood Forecast Using Projections

The projected data that we got from the "Autoregression" models were then employed to get flood incidence forecasts from out trained flood prediction model. This gives us some idea about the estimation capability of the model. This can also help with future flood preparedness for calamity-prone areas like Kerala.

Time	**kNN**	**LR**	**DT**	**RF**	**EL**	**SVC**
2022	YES	YES	YES	YES	YES	YES
2023	YES	YES	YES	YES	YES	YES
2024	YES	YES	YES	YES	YES	YES
2025	YES	YES	YES	YES	YES	YES
2026	NO	YES	NO	NO	NO	NO
2027	YES	YES	YES	YES	YES	YES
2028	YES	YES	YES	YES	YES	YES
2029	NO	YES	NO	NO	NO	NO
2030	NO	YES	NO	NO	YES	NO
2031	YES	YES	YES	YES	YES	YES
2032	YES	YES	YES	YES	YES	YES
2033	NO	NO	NO	NO	NO	NO
2034	NO	NO	NO	NO	NO	NO
2035	YES	YES	YES	YES	YES	YES
2036	YES	YES	YES	YES	YES	YES
2037	YES	YES	YES	YES	YES	YES
2038	YES	YES	YES	YES	YES	YES
2039	YES	YES	YES	YES	YES	YES
2040	YES	YES	YES	YES	YES	YES
2041	YES	YES	YES	YES	YES	YES
2042	YES	YES	YES	YES	YES	YES
2043	NO	NO	NO	NO	NO	NO
2044	YES	YES	YES	YES	YES	YES
2045	YES	YES	YES	YES	YES	YES
2046	NO	NO	NO	NO	NO	NO
2047	NO	NO	NO	NO	NO	NO
2048	NO	NO	NO	NO	NO	NO
2049	YES	NO	YES	YES	YES	YES
2050	YES	YES	YES	YES	YES	YES

Table 7.4: Future Flood Forecast Results(2022-2050)

Chapter 8

Limitations

The beautiful factor about scientific research is the fact that the research operations are cumulative pursuits. Moreover, for the comprehensive effectiveness of any study, it is imperative that we highlight the limitations of the particular study in a faultless manner. Thus, we are including the limitations that we faced during this particular research as well.

1. Modern structured publicly accessible climate data relating to flooding, sea-level rise, and rainfall is sparse, especially when comparing emerging nations to developed portions of the globe.

2. To this day, research articles and field studies concentrating on the relationship between sea-level rise, rainfall, and the incidence of floods connected to the scope of machine learning are hard to come by.

3. Publicly available and organized datasets related to the future prediction of "GMSL", Global Annual Temperature, "Annual Rainfall" would boost up the effectiveness of our prediction model to new heights. Unfortunately, there still remains a lack of publicly available future forecasts of the aforesaid mentioned data.

4. Because the majority of this study was undertaken during the peak period of the Corona Outbreak, online data gathering, online resources, and online measurements had to be prioritized above any physical data exploration methods.

Chapter 9

Conclusion

Challenges pertaining to the tackling of the existing rate of occurrence of natural calamities are overwhelming as they are and the lack of action isn't assisting the situation. Despite living in a time when there are encyclopedias worth of information at our fingertips, organized information about climate induced disasters is still relatively scarce. This seems to be at odds with the current situation where we need data now more than ever. So, there is a lot of potential and need for more research in this particular field. But we can still employ machine learning and put the scant information we have to good use in our fight against one of the biggest threats that is plaguing the modern world, which are climate induced natural disasters. As the occurrences of these disasters get more intense and frequent the need for predictions long ahead of time will keep on increasing. And although our project won't be able to help with every possible climate induced natural disaster, it can hopefully do some good in regards to flooding incidents brought about by climate change. Tropical coastal regions like Kerala, which was at the focus of our research, will increasingly be at a risk of flooding due to the escalating sea levels. This bulging risk factor has the potential to overwhelm the fragile defense and recovery mechanisms of an already disaster-prone region. We sought to explore and better understand the factors that contribute to the incidence of climate-induced flooding. And to that end, we would like to imagine that our pursuit has been successful in bringing attention to the research potential in the sphere of climate-induced flood projection using machine learning approaches.

Bibliography

[1] P Vellinga and Willem J Verseveld. *Climate change and extreme weather events*. WWF, 2000.

[2] *5 natural disasters that Beg for Climate Action*. 2020. URL: https://www.oxfam.org/en/5-natural-disasters-beg-climate-action.

[3] R Pidcock and R McSweeney. *Mapped: How climate change affects extreme weather around the world*. 2021. URL: https://www.carbonbrief.org/mapped-how-climate-change-affects-extreme-weather-around-the-world.

[4] Anke Jentsch and Carl Beierkuhnlein. "Research frontiers in climate change: effects of extreme meteorological events on ecosystems". In: *Comptes Rendus Geoscience* 340.9-10 (2008), pp. 621–628.

[5] Susan Solomon et al. *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*. Vol. 4. Cambridge university press, 2007.

[6] Abu Reza Md Towfiqul Islam et al. "Flood susceptibility modelling using advanced ensemble machine learning models". In: *Geoscience Frontiers* 12.3 (2021), p. 101075.

[7] FAO et al. "SOFI 2017 - The State of Food Security and Nutrition in the World". In: *Fao.org* (). URL: http://www.fao.org/state-of-food-securitynutrition/2017/en/.

[8] Elina Oksanen et al. "Photosynthesis of birch (Betula pendula) is sensitive to springtime frost and ozone". In: *Canadian Journal of Forest Research* 35.3 (2005), pp. 703–712.

[9] Victor Schmidt et al. "Visualizing the consequences of climate change using cycle-consistent adversarial networks". In: *arXiv preprint arXiv:1905.03709* (2019).

[10] Eugene C Cordero, Anne Marie Todd, and Diana Abellera. "Climate change education and the ecological footprint". In: *Bulletin of the American Meteorological Society* 89.6 (2008), pp. 865–872.

[11] Yunjie Liu et al. "Application of deep convolutional neural networks for detecting extreme weather in climate datasets". In: *arXiv preprint arXiv:1605.01156* (2016).

[12] Thorsten Kurth et al. "Exascale deep learning for climate analytics". In: *SC18: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE. 2018, pp. 649–660.

[13] Evan Racah et al. "ExtremeWeather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events". In: *arXiv preprint arXiv:1612.02095* (2016).

[14] Valliappa Lakshmanan and Travis Smith. "An objective method of evaluating and devising storm-tracking algorithms". In: *Weather and Forecasting* 25.2 (2010), pp. 701–709.

[15] Christopher B Field et al. *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.

[16] *How climate change is fueling extreme weather*. 2021. URL: https://earthjustice.org/features/how-climate-change-is-fueling-extreme-weather.

[17] Thomas R Knutson and Syukuro Manabe. "Model assessment of decadal variability and trends in the tropical Pacific Ocean". In: *Journal of Climate* 11.9 (1998), pp. 2273–2296.

[18] Thomas R Knutson and Robert E Tuleya. "Increased hurricane intensities with CO2-induced warming as simulated using the GFDL hurricane prediction system". In: *Climate Dynamics* 15.7 (1999), pp. 503–519.

[19] IPCC Glossary. "Climate Change 2014 Report Fifth Assessment Report". In: *Intergovernmental Panel On Climate Change* (2014).

[20] Stephen Ornes. "Core Concept: How does climate change influence extreme weather? Impact attribution research seeks answers". In: *Proceedings of the National Academy of Sciences* 115.33 (2018), pp. 8232–8235.

[21] Tim Wheeler and Joachim Von Braun. "Climate change impacts on global food security". In: *Science* 341.6145 (2013), pp. 508–513.

[22] Rebecca Lindsey. *Climate change: Global sea level*. 2020. URL: https://www.climate.gov/news-features/understanding-climate/climate-change-global-sea-level.

[23] World Bank Group. *Warming climate to hit Bangladesh hard with sea level rise, more floods and cyclones, World Bank report says*. 2013. URL: https://www.worldbank.org/en/news/press-release/2013/06/19/warming-climate-to-hit-bangladesh-hard-with-sea-level-rise-more-floods-and-cyclones-world-bank-report-says.

[24] David Rolnick et al. *Tackling Climate Change with Machine Learning*. 2020. URL: https://www.microsoft.com/en-us/research/publication/tackling-climate-change-with-machine-learning/.

[25] Muhammed Sit and Ibrahim Demir. "Decentralized flood forecasting using deep neural networks". In: *arXiv preprint arXiv:1902.02308* (2019).

[26] Sang-Jin Park and Dong-Kun Lee. "Prediction of coastal flooding risk under climate change impacts in South Korea using machine learning algorithms". In: *Environmental Research Letters* 15.9 (2020), p. 094052.

[27] Deva Hema et al. "Global Warming Prediction in India using Machine Learning". In: *International Journal of Engineering and Advanced Technology(IJEAT)* 9.1 (2019), 4061–4065. DOI: 10.35940/ijeat.A1301.109119.

[28] Laura A Mansfield et al. "Predicting global patterns of long-term climate change from short-term simulations using machine learning". In: *npj Climate and Atmospheric Science* 3.1 (2020), pp. 1–9.

[29] Wanie M Ridwan et al. "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia". In: *Ain Shams Engineering Journal* 12.2 (2021), pp. 1651–1663.

[30] Michael B Richman et al. "Reducing tropical cyclone prediction errors using machine learning approaches". In: *Procedia computer science* 114 (2017), pp. 314–323.

[31] Betsy Gardner. *The importance of data for forecasting and tracking floods*. 2019. URL: https://datasmart.ash.harvard.edu/news/article/importance-data-forecasting-and-tracking-floods.

[32] Alexandra Witze. *Why extreme rains are gaining strength as the climate warms*. 2018. URL: https://www.nature.com/articles/d41586-018-07447-1.

[33] *Pouring it on: How climate change intensifies heavy rain events*. 2019. URL: https://www.climatecentral.org/news/report-pouring-it-on-climate-change-intensifies-heavy-rain-events.

[34] Andrew Krosofsky. *Flooding and climate change: A mutually-destructive connection*. 2021. URL: https://www.greenmatters.com/p/how-does-climate-change-affect-floods.

[35] *NASA Sea Level Change Portal*. 2021. URL: https://sealevel.nasa.gov/understanding-sea-level/key-indicators/global-mean-sea-level/.

[36] National Oceanic and Atmospheric Administration. *Climate at a glance*. URL: https://www.ncdc.noaa.gov/cag/global/background.

[37] Shreeshan Venkatesh and Rejimon Kuttappan. *This is why Kerala floods were the worst in a century*. 2018. URL: https://www.downtoearth.org.in/coverage/climate-change/this-is-why-kerala-floods-were-the-worst-in-a-century-61491.

[38] *Envis Hub: Kerala state of environment and related issues*. URL: http://www.kerenvis.nic.in/Database/CLIMATE_829.aspx.

[39] Tiki Rajwi. *Kerala records its highest annual rainfall in 60 years in 2021*. 2021. URL: https://www.thehindu.com/news/national/kerala/kerala-records-its-highest-annual-rainfall-in-60-years-in-2021/article38082426.ece.

[40] *Kerala floods: At least 26 killed as rescuers step up efforts*. 2021. URL: https://www.bbc.com/news/world-asia-india-58940880#:~:text=In%202018,%20more%20than%20400,%20from%20vulnerable%20areas%20to%20safety.

[41] Sanchita Sivaraman. *Do you know about the Kerala Flood ofnbsp;1924?nbsp;* 2018. URL: https://www.newindianexpress.com/states/kerala/2018/aug/17/do-you-know-about-the-kerala-flood-of-1924-1859072.html#:~:text=Several%20church%20buildings%20were%20destroyed,killed%20a%20number%20of%20livestock.

[42] *What causes flash floods?* URL: https://www.metoffice.gov.uk/weather/learn-about/weather/types-of-weather/rain/flash-floods.

[43] Christina Nunez. *Sea level rise, facts and information*. 2019. URL: https://www.nationalgeographic.com/environment/article/sea-level-rise-1.

[44] Robert J Nicholls, Frank MJ Hoozemans, and Marcel Marchand. "Increasing flood risk and wetland losses due to global sea-level rise: regional and global analyses". In: *Global Environmental Change* 9 (1999), S69–S87.

[45] Gabriel Popkin. *This ecologist thinks coastal wetlands can outrun rising seas. Not everyone's convinced*. 2021. URL: https://www.science.org/content/article/ecologist-thinks-coastal-wetlands-can-outrun-rising-seas-not-everyone-s-convinced.

[46] K Kokkal, Periya Harinarayanan, and KK Sabu. "Wetlands of Kerala". In: *Proceedings of Taal* (2007), pp. 1889–1893.

[47] *Envis Hub: Kerala state of environment and related issues*. 2021. URL: http://www.kerenvis.nic.in/Database/Wetland_6747.aspx.

[48] Sebastian Raschka and Vahid Mirjalili. "Python Machine Learning: Machine Learning and Deep Learning with Python". In: *Scikit-Learn, and TensorFlow. Second edition ed* (2017).

[49] Frank Schoonjans. *ROC curve analysis*. 2021. URL: https://www.medcalc.org/manual/roc-curves.php.

[50] *State of climate in 2021: Extreme events and major impacts*. 2021. URL: https://public.wmo.int/en/media/press-release/state-of-climate-2021-extreme-events-and-major-impacts.

[51] NOAA National Centers for Environmental information. *Climate at a Glance: Global Time Series*. 2022. URL: https://www.ncdc.noaa.gov/cag/global/time-series/globe/land_ocean/ytd/11/1880-2021.

[52] Sandeep Vellaram. *2021 rainiest year for Kerala in 6 decades with 110rainfall*. 2021. URL: https://www.thenewsminute.com/article/2021-rainiest-year-kerala-6-decades-110-excess-rainfall-158016.

[53] *Kerala*. URL: https://floodlist.com/tag/kerala.