Classifying countries on the basis of stress level for water scarcity

A Report submitted in partial fulfillment of the requirement for the degree of

Bachelor Of Technology

In

INFORMATION TECHNOLOGY

Under the Supervision of

Dr. Jyoti Gautam

By

Sankalp Kumar (1809113090)

Surya Pratap Singh (1809113111)

JSS MAHAVIDYAPEETHA

JSS ACADEMY OF TECHNICAL EDUCATION

Session: - 2020-2021

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Name : Sankalp Kumar

Roll No.: 1809113090

Date : 01-05-2022

Name : Surya Pratap Singh

Roll No.: 1809113111

Date : 01-05-2022

CERTIFICATE

This is to certify that Project Report entitled “Classifying countries on the basis of stress level for water scarcity” which is submitted by Maniza Singh(1809113062), Sankalp Kumar(1809113090), Siddharth Prakash(1809113106) and Surya Pratap Singh(1809113111) in partial fulfillment of the requirement for the award of degree B. Tech. in the Department of Information Technology of U. P. Technical University, is a record of the candidate's own work carried out by him under my/our supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Supervisor: **Dr Jyoti Gautam**

Date: ***01-05-2022***

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Jyoti Gautam, Department of Information Technology, JSS Academy of Technical Education, Sector 62 Noida for his constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Name : Sankalp Kumar

Roll No.: 1809113090

Date : 01-05-2022

Name : Surya Pratap Singh

Roll No.: 1809113111

Date : 01-05-2022

ABSTRACT

Water shortage occurs when communities are unable to meet their water demands due to a lack of supply or inadequate infrastructure. Currently, billions of people are suffering from water scarcity. In the past, countries have frequently collaborated on water management. Even yet, there are a few regions, such as the Nile Basin, where transboundary waters are causing difficulties. Rising temperatures would certainly worsen water stress throughout the world, as more unpredictable weather and extreme weather events, such as floods and droughts, result from climate change. With the information of availability and usage of water the water stress level with respect to that country can be predicted which can give us the possible year in which the country will have water stress level at critical point. Also the clusters of countries are made with respect to raw stress values of the countries in a particular year. The goal is to look at the groundwater stress levels in different nations through time and anticipate the stress levels in different countries in the future so that we can see how early we need to pay attention and what steps we need to take to avert a situation of critical water stress level and to adopt more sustainable and creative methods, as well as to increase international water management cooperation.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

1.1 Problem Introduction

When the demand for safe, usable water in a specific region exceeds the availability, water stress or shortage arises. On the demand side, agriculture consumes around 70% of the world's freshwater, with the balance going to industrial (19%) and residential (11%) usage, including drinking. Surface waters such as rivers, lakes, and reservoirs, as well as groundwater obtained through aquifers, are sources of supply. Water stress varies greatly from place to location, and in certain situations, it can have far-reaching consequences for public health, economic growth, and global trade. It has the potential to cause widespread migration and violence. Pressure is rising on governments to adopt more sustainable and creative methods, as well as to increase international water management cooperation.

A case of significant water shortage is described as a critical water stress level, which Cape Town came dangerously close to experiencing last year. They have so far avoided a full water crisis, but it serves as a stark message to the rest of the globe. Water shortages occur in many places across the world throughout the year, creating major disruptions in residents' daily lives, companies, and the economy. The capacity to keep a watch on cities that are approaching critical or alert levels of water stress.

This crucial[15] water stress level might be looked at as a "Day Zero" occurrence, which could be extremely valuable in not only properly delivering relief, but also in taking precautionary measures early on so that local governments are not forced to impose excessive water consumption restrictions on short notice. To assist raise awareness about this issue, we utilized machine learning to create a model that predicts a country's water stress level based on a set of criteria. We split the raw stress levels based on the water resources availability and usage to get a country's likely year for the stress level to reach a critical and alert level. Also clusters of countries are formed based on the water stress level in the present year and predicted using machine learning algorithms for the future years.

1.1.1 Motivation

Various nations throughout the world are seeing a loss in renewable water supplies while increasing their use. Our goal is to look at the groundwater stress levels in different nations through time and anticipate the stress levels in different countries in the future so that we can see how early [11]we need to pay attention and what steps we need to take to avert Day Zero.

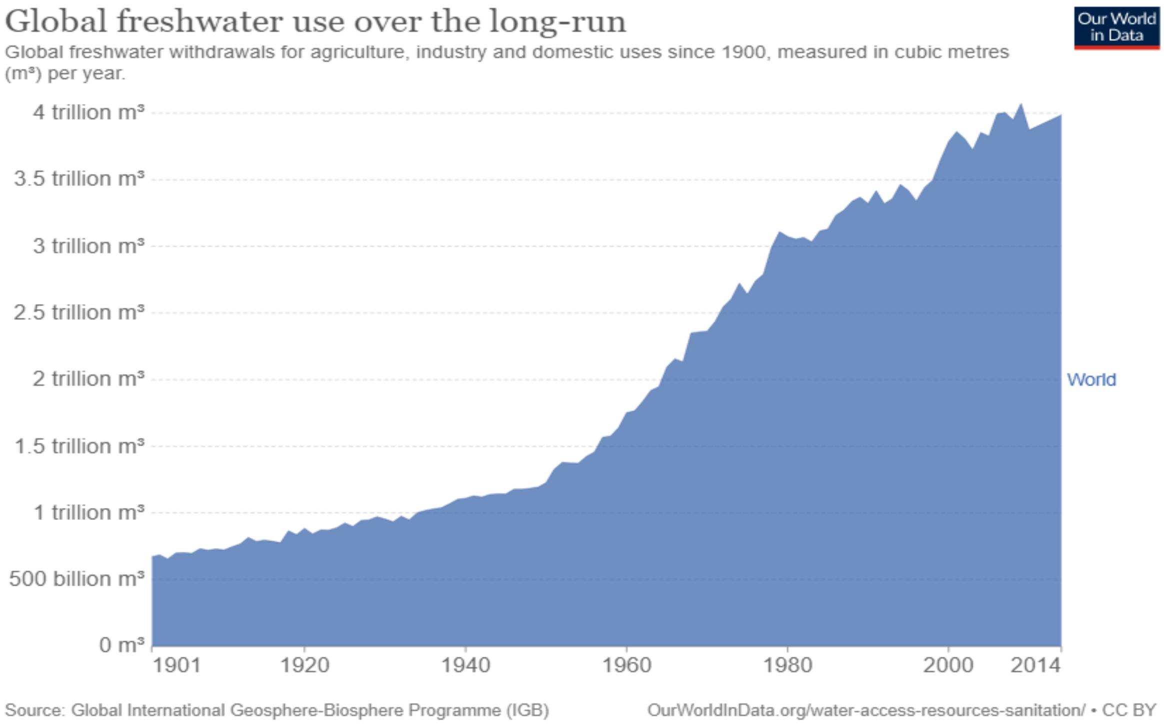


FIG. 1

Water availability per person has decreased rapidly in most regions of the world, indicating that the country is already experiencing a drinking water crisis.

People will be more prone to dangerous water-borne infections if they do not have access to clean water. The world's population is increasing as water supplies are diminishing year after year, putting an increasing number of people at risk of running out of water.

We want to build a system that uses machine learning to anticipate ground water stress levels based on the availability and use of water resources in that nation, so that the aforementioned issues may be identified and dealt with efficiently

1.1.2 Project Objective

Groundwater Stress may be measured in a variety of methods. Using the availability and usage of total water resources per capita, different machine learning algorithms to make predictions were employed.

- To analyze the stress level across the different countries.
- To Visualize the trends in declining renewable water resources and increase in total water usage per capita across different countries.
- To predict the day zero year based on the critical water stress value.

1.1.3 Scope of the Project

In our project, we will be taking the Stress Level of Groundwater using data analysis on usage and availability of groundwater can be very useful for prediction of stress level of a country in the future.

Rainfall and Stress level prediction can be helpful for preventing Day Zero Conditions in any region, and can be helpful to farmers for irrigation of crops as it depends on groundwater, drinking water quality.

To help raise awareness to this situation, we used machine learning to develop a model to predict the water stress level of a country given a particular set of attributes. We divided the raw stress values (a scale from 0-100) into 3 categories:

Stress Value of 0-40: Low

Stress Value of 41-80: Medium

Stress Value of >80: High

1.2 Related Previous Work

Data on the total use of renewable water resources is crucial for formulating sustainable groundwater management policy in India. [2]Controlling the pace of groundwater stress requires an understanding of the overall groundwater resources available and how they are used. In India, there are now essentially no limits on groundwater development. India's electricity regulations actually encourage the rapid expansion of groundwater. Agriculture has relied heavily on groundwater, which has been aided by deep irrigation and poverty reduction.

The government is saving groundwater stress rate and rainfall data in order to evaluate it and utilize machine learning techniques to estimate future water demand for [9]better crop development and to assess the falling trend of water levels in water limited places.

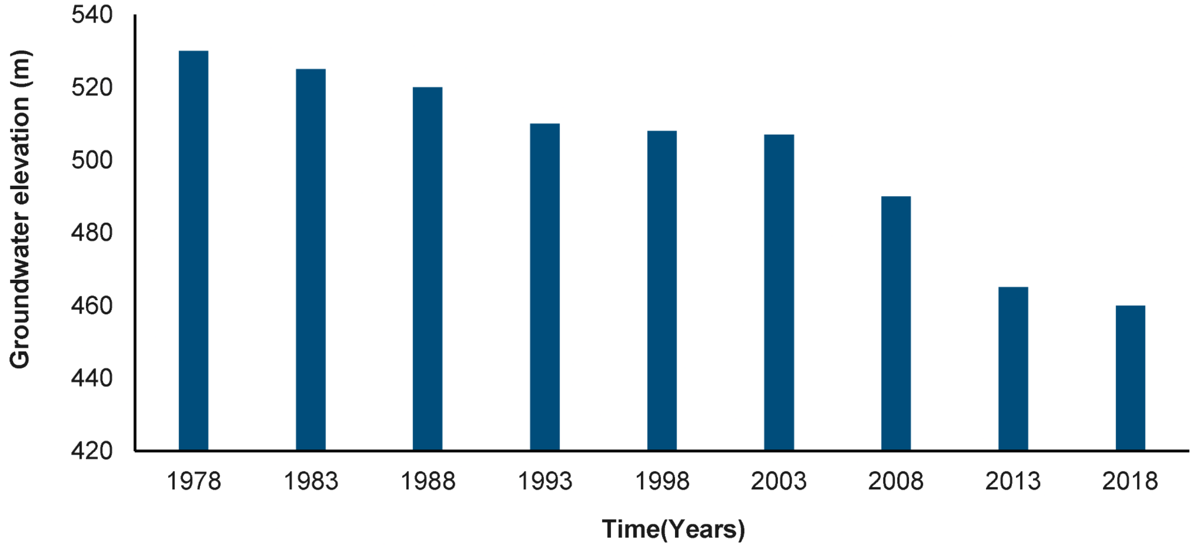


FIG. 2

1.3 Organization Of Report

Report is divided and organized in the following order:

Chapter 1 gives an overview about the problem introduction along with motivation, project objective, scope of the project, previous work.

Chapter 2 presents a quick review of the research and analysis of many approaches in their individual work, as well as their respective methodology.

Chapter 3 describes the system design and architecture along with algorithms and techniques used in our project. A flowchart is given describing the sequence of how various algorithms and techniques are applied throughout the flow of the project.

Chapter 4 specifies the implementation approaches, such as implementation details, features, and ideas, that are utilizedo to obtain outcomes from the implementation.

Chapter 5 Assesses the performance of implemented models, making comparisons to previous states and planning future work.

CHAPTER 2

Literature Survey

Michael B. Richmana, Lance M. Leslieb (2018)- concluded that by detailed application of an ensemble of attribute selection techniques and support vector regression, cross validated prediction of precipitation was made. Adding additional attributes and applying machine learning techniques to winnow the number of predictors can be at least moderately effective in increasing the predictive capability.. As mentioned in the Introduction, this region has been identified as one on which has been identified by the IPCC as being vulnerable to the combined impacts of lower rainfall, increasing temperature, and indirect or non-climate factors, such as rapid population growth and massive land use changes. All these factors should be the subject of future research in leading to extended droughts

P.H. Herbst; D.B. Bredenkamp; H.M.G. Barker (1966)- Proposed a technique for the evaluation of drought from rainfall data. A method, based on monthly rainfall data, is described, whereby it is possible to determine the duration and intensity of droughts and their months of onset and termination; a drought index is also calculated which enables the intensity of droughts to be compared irrespective of their seasonal occurrence. This method has been programmed for art electronic computer to facilitate the investigation of data from a large number of rainfall stations.

Johanna Brühl, Martine Visser (2021)-A study of the combined effectiveness of measures implemented to prevent “Day Zero” **.** They examined the relation of water use reductions and four kinds of drought measures – tariff increases, restrictions, water pressure reduction, and public information and education campaigns – as the drought escalated into a crisis. A water crisis happening in the context of extreme inequality, such as in Cape Town, raises the question of how the various income groups dealt with the water crisis. The more affluent residents found their own private ways to be less reliant on municipal water through investments in alternative water supply such as rainwater tanks, boreholes, well points, and bottled water. Inevitably, this raises fairness and justice issues that could be explored in future research.

G. Thomas LaVanchy & Michael W. Kerwin & James K. Adamson (2019)- They informed about the acute water security vulnerability at regional to local scales is becoming increasingly apparent. In 2017, officials in Cape Town, South Africa, designated the term B Day Zero^ to demark an exact time when the city's taps would be switched off due to critically low reservoir levels. Beyond Day Zero, residents would need to converge at communal water collection points to access a 25-L daily ration of water. The particulars of the crisis and stakeholder responses prove informative to other cities.However, the city must grapple with the fact that reliable, repetitive storms may no longer inundate the region each winter and that climate model projections for the late twenty-first century warn of even less annual precipitation, including wet days and extreme high precipitation events.

Agana, Norbert A; Homaifar, Abdollah (2017) - A deep learning-based approach was used to investigate the drought prediction problem in this research. For long-term drought prediction, we presented a deep belief network (DBN) and compared its performance to that of typical MLP and SVR models. In comparison to MLP, the DBN model was found to produce higher prediction outcomes with smaller prediction errors, making it more trustworthy and efficient for long-term drought prediction. Its performance over the SVR, on the other hand, was less impressive. This is most likely owing to the lack of big sample sizes required to effectively exploit the deep architecture's capabilities in the DBN model. Future research will look at how well the DBN model predicts drought using global climatic indices in addition to the standardised streamflow index. We may investigate employing the DBN for pre-training and the SVR for final prediction in our future study because the SVR model performed similarly to the DBN model.

Salvatore Pascale, Sarah B. Kapnick, Thomas L. Delworth, and William F. Cooke - Despite huge internal climate variability, the adoption of a high-resolution large ensemble significantly improves the ability to model regional-scale SSA droughts in both current and future situations. With a 95% confidence interval of [4, 8], we find that the rainfall deficit that caused the Day Zero drought was 5.5 times more likely to be caused by anthropogenic climate change. We are thus able to further constrain the risk ratio of SSA drought at and beyond the original [1.4, 6.4] estimate from ref. 5 by using a model with higher resolution and better climatology. This demonstrates the value of high-resolution climate models in predicting future drought risk and provides more information for water management planning to avoid severe drought.Looking ahead, our findings hint to a significant rise in the chance of similar or even more severe meteorological droughts by the end of the twenty-first century. This increased likelihood of meteorological droughts is linked to a significant decline in rainfall, particularly during the shoulder season, similar to what happened in 2015–2017.

M.J.Booysen, M. Visser, R. Burger - The events leading up to Cape Town's alleged "Day Zero" in 2017 and 2018 are described in this study, which links significant events to changes in behaviour of a small sample of consumers with smart metres that monitor cold and hot water, respectively. The largest response was found, not when the limits or tariff increases were enforced, but in response to a three-phased catastrophe plan that warned of devastating effects, according to social media and search phrase studies combined with a time series analysis.Users reacted more strongly to the fear of waterless taps than to the actual water limitations, according to the findings.

Our smart metre data, together with billing data from the city, indicates that citizens have been able to substantially change their consumption patterns in a relatively short amount of time. Furthermore, while inciting some amount of fear-mongering may have been a hazardous tactic for the municipality to pursue, it appears that it may have been the single most successful intervention in producing dramatic behavioural change among individuals.

Amy Maxmen - If water is scarce, Wilkinson adds, maintaining health should be the top focus. "I instantly think about the risk of water-borne infections if people — especially those in poor living situations — are unable to maintain personal and institutional hygiene," Wilkinson adds. He is particularly concerned about the city's economy, which is heavily dependent on tourism and agriculture. Kevin Winter, the director of the university's Water Task Team, which oversees water use on all of its campuses, cites a number of reasons why institutions are unprepared for the situation. As the scenario shifts from hopeful to dire, faculty members are only returning to college following their summer break. Winter explains that while colleges have studied water-saving techniques in the past, such as water-recycling systems, executing those ideas has been challenging due to funding constraints. For example, institutions have agreed to keep their fees constant despite inflation in response to student protests seeking free education in recent years. Winter explains, "It's not as simple as saying, 'Give me two to three million dollars for pipes and pumps and geological surveys.'

Wolski, Piotr (2018) - The best estimate of the meteorological drought's return interval in the WCWSS dams region is 311 years, with a 90% confidence interval of 105 to 1280 years. To put it another way, the recent drought, which was characterised by poor rainfall from 2015 to 2017, was extremely rare and severe.

Capetonians will undoubtedly continue to discuss whether or not poor management contributed to the disaster. However, according to my analysis, the authorities were confronted with a difficult scenario anyway.

Ahmadi, Mohammad Sadeq; SuÅ¡nik, Janez; Veerbeek, William; Zevenbergen, Chris (2020) - The current and anticipated water demand and supply of 12 megacities (Cairo, Delhi, Dhaka, Ho Chi Minh City, Jakarta, Kolkata, Lagos, Lahore, Manila, Mexico City, Mumbai, and Tehran) were examined using a high-level approach. These cities' current water consumption exceeds supply by 5.27 billion m3 yr-1, with this figure anticipated to rise further by 2035.

Physical water losses alone, according to current estimates, are 4.7 billion m3, which is nearly enough to close the overall supply-demand gap. Current physical water losses are enough to provide 135 litres per capita per day to about 100 million people. The enormous demand growth projections in many cities (in the worst-case scenario, nearly six-fold increases) raise important problems and challenges about how to provide sufficient freshwater to these cities' people in a sustainable manner. To make matters worse, present water coverage rates (as low as 10%) are extremely poor, and many settlements are likely to be informal. Water supply sources are under threat from climate change, and socioeconomic trends may lead to increased water demand. It has been demonstrated that lowering demand as well as lowering water loss can result in considerable reductions in the supply-demand gap. As a result, quick and comprehensive action on water loss reduction, overall demand control, and formal infrastructure expansion is required.

Jeroen F. Warner;Richard Meissner; (2021) - When seen from the perspective of disaster risk reduction, the case underlines people's pro-social behaviour and relative lack of panic in the face of (impending) crisis, both of which are extensively documented in disaster sociology research.

Making Day Zero such a high-profile media 'event' faced the risk of oversimplification, overshooting its purpose, or even exploding in the messenger's face, as the former did. While Cape Town's social makeup is unique, disaster (and climate) communicators could learn a thing or two from the city's approach to leveraging people's sense of humour and togetherness even in the face of adversity.

CHAPTER 3

SYSTEM ARCHITECTURE AND METHODOLOGY

3.1 SYSTEM DESIGN

This section explains how the system will be designed, as well as the tools and approaches that will be used throughout the project.

3.1.1 Architecture diagram

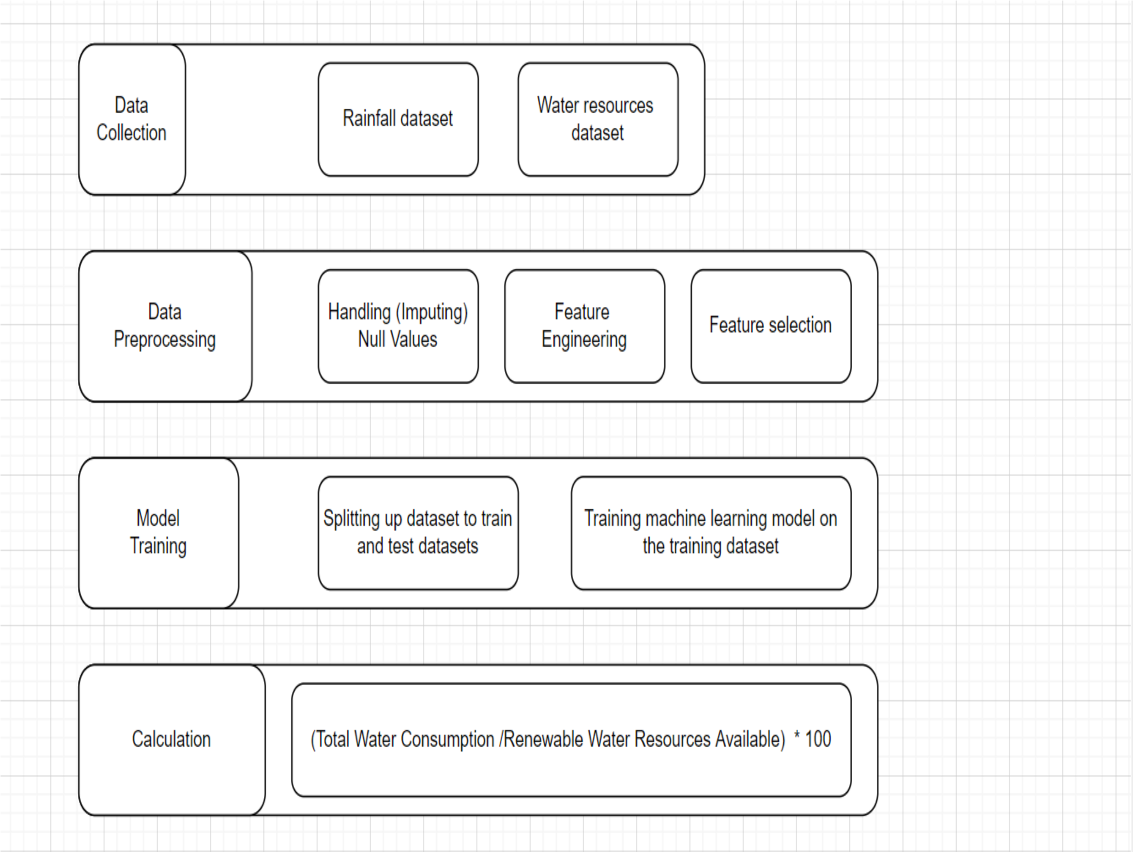

Data Collection
Rainfall dataset
Water resources dataset
Data Preprocessing
Handling (Imputing) Null Values
Feature Engineering
Feature selection
Model Training
Splitting up dataset to train and test datasets
Training machine learning model on the training dataset
Calculation
(Total Water Consumption /Renewable Water Resources Available) * 100

FIG. 3

3.1.2 Flowchart

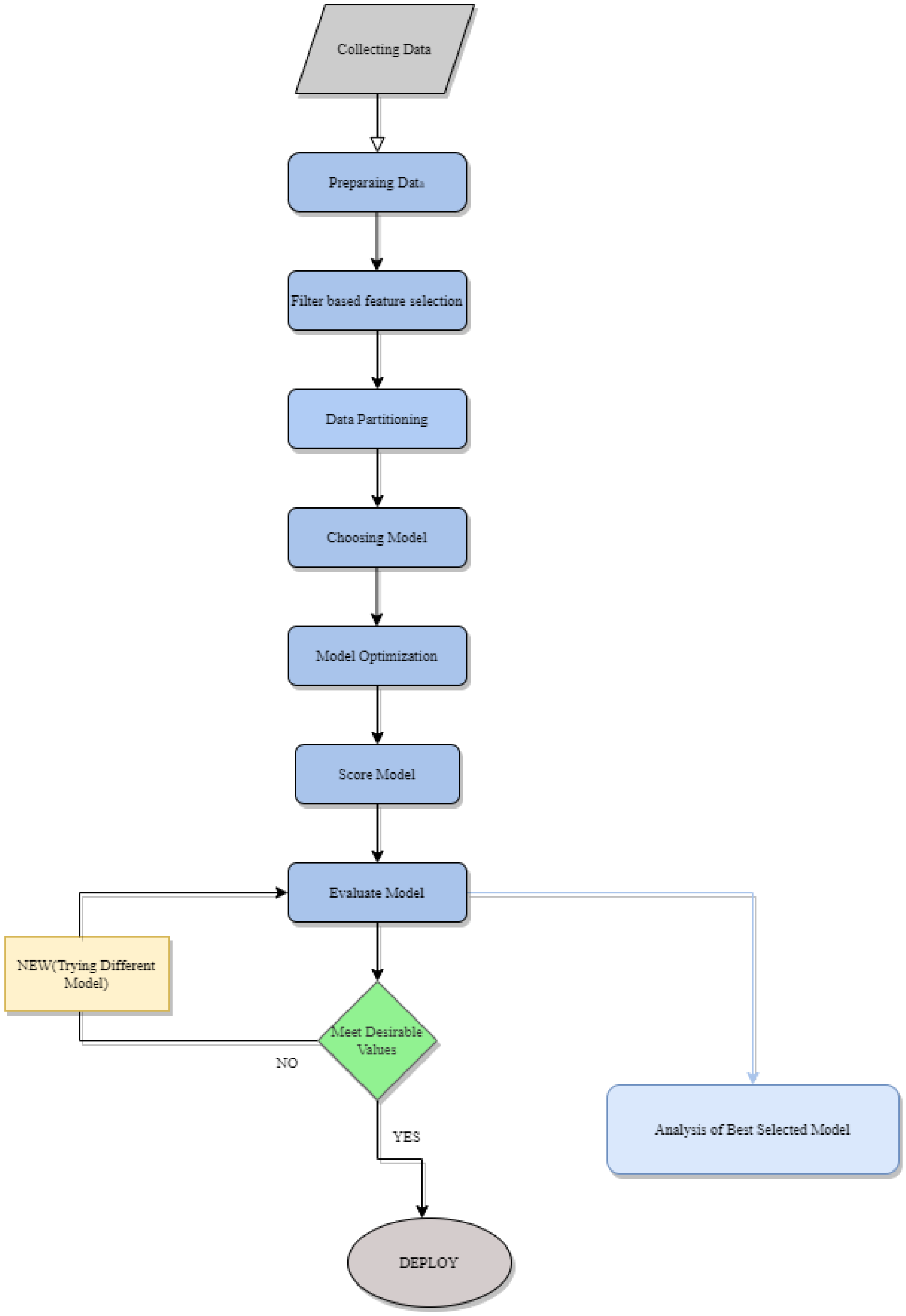


FIG. 4

3.1.3 ER Diagram

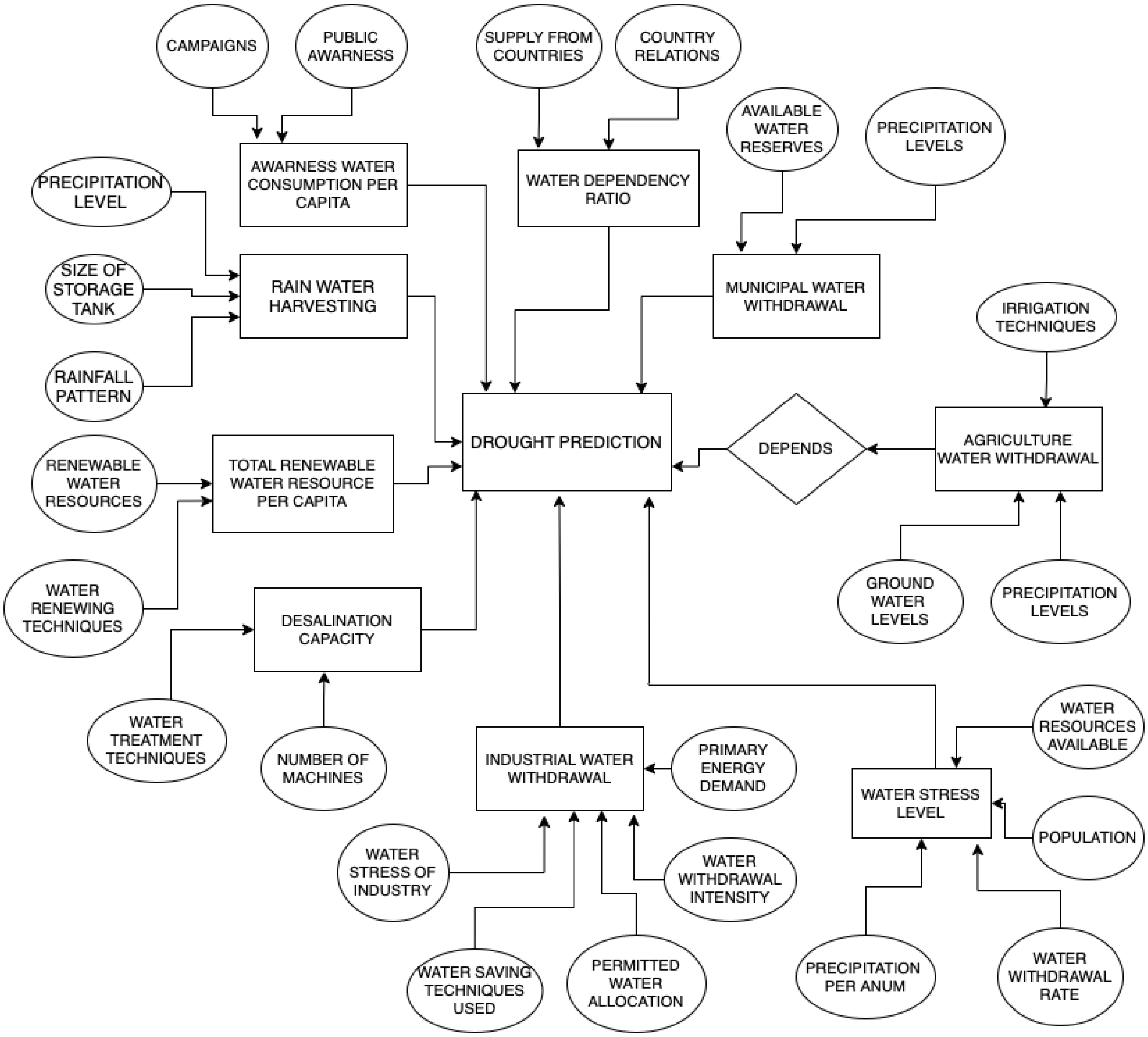


FIG. 5

1. **Data Collection:**

a. The dataset of Rainfall and Total water resources availability and consumption of water resources across different countries from 1962-2017 is collected from the official website of AQUASTAT DATABASE.

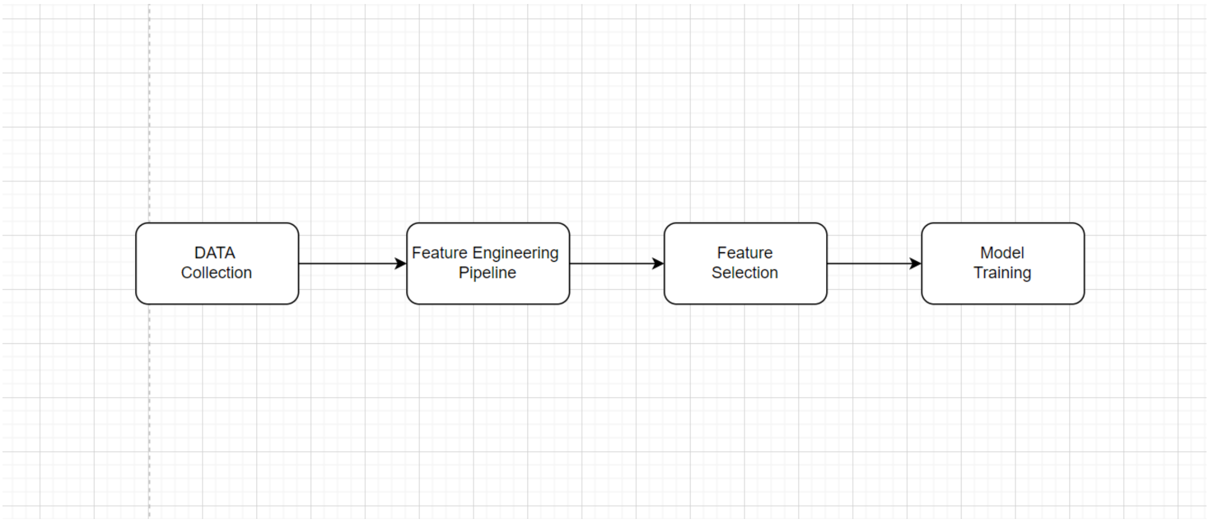


FIG. 6

2. **Data Preprocessing:**

b. Preprocessing is done to clean the data and to make it ready for analysis, we have proposed our aggregated data of Rainfall and Groundwater.

c. The data is sorted accordingly date wise and only the required columns are analyzed.

d. Groundwater availability and consumption values of different countries are available in a 5 years gap.

3. **Calculation**

e. Stress level for any specific year and country, is calculated using this formula:

Groundwater Stress Rate = (Total water resources usage per capita / Total renewable water resources available per capita) * 100

3.3 Algorithms

Decision Tree- Decision Tree Algorithm

The Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too.

The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data (training data)

Algorithm step-by-step:

1. Begin the tree with the root node, says S, which contains the complete dataset.
2. Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
3. Divide the S into subsets that contain possible values for the best attributes.
4. Generate the decision tree node, which contains the best attribute.
5. Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and call the final node as a leaf node.

Isolation Forest- Isolation Forests are nothing but an ensemble of binary decision trees. And each tree in an Isolation Forest is called an Isolation Tree (iTree). The algorithm starts with the training of the data, by generating Isolation Trees.

Complete algorithm step by step:

1. When given a dataset, a random subsample of the data is selected and assigned to a binary tree.
2. Branching of the tree starts by selecting a random feature (from the set of all N features) first. And then branching is done on a random threshold (any value in the range of minimum and maximum values of the selected feature).
3. If the value of a data point is less than the selected threshold, it goes to the left branch else to the right. And thus a node is split into left and right branches.
4. This process from step 2 is continued recursively till each data point is completely isolated or till max depth (if defined) is reached.
5. The above steps are repeated to construct random binary trees.

Random Forest-Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." [2]Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting**.**

Algorithm Step-by-step:

1. Select random K data points from the training set.
2. Build the decision trees associated with the selected data points (Subsets).
3. Choose the number N for decision trees that you want to build.
4. Repeat Step 1 & 2.
5. For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

Multiple linear regression- Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regressions is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regressions are the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

Algorithm step-by-step:

1: Data Pre Processing
Importing the Libraries.
Importing the Data Set.
Encoding the Categorical Data.
Avoiding the Dummy Variable Trap.
Splitting the Data set into Training Set and Test Set.
2: Fitting Multiple Linear Regression to the Training set
3: Predicting the Test set results.

Row Interpolation- Interpolation is a statistical method by which related known values are used to estimate an unknown price or potential yield of a security. Interpolation is achieved by using other established values that are located in sequence with the unknown value. Interpolation is at root a simple mathematical concept. If there is a generally consistent trend across a set of data points, one can reasonably estimate the value of the set at points that haven't been calculated. The code generated by using row-major interpolation algorithm performs with the best speed and memory usage when operating on table data with row-major array layout.

CHAPTER 4

IMPLEMENTATION AND RESULTS

4.1. Software and Hardware Requirements

- **Jupyter Notebook :-**

  Jupyter Notebook (previously IPython Notebooks) is an interactive computing platform for writing notebook papers that is accessible through the web. Jupyter Notebook is compatible with a variety of kernels, allowing you to programme in a variety of languages. A Jupyter kernel is a software that can handle a variety of queries (code execution, code completion, and inspection). Unlike many other Notebook-like interfaces, Jupyter kernels are unaware that they are attached to a specific document and can connect to several clients at the same time.

- **Pandas :-**

  Pandas is primarily used for data analysis and related tabular data manipulation in Data Frames. Pandas supports importing data from comma-separated values (CSV), JSON, Parquet, SQL database tables or queries, and Microsoft Excel. Pandas supports a wide range of data manipulation operations, including merging, reshaping, and selecting, as well as data cleaning and wrangling. Many similar aspects of dealing with Dataframes that were created in the R programming language were incorporated into Python with the creation of pandas. The pandas library is based on NumPy, a Python library geared at effectively working with arrays rather than the characteristics of working with Dataframes.

- **skLearn :-**

  Scikit-learn is mostly built in Python, and it heavily relies on NumPy for high-speed linear algebra and array operations. Scikit-learn (previously scikits.learn, and also known as sklearn) is a free Python machine learning package. [3] It includes support-vector machines, random forests, gradient boosting, k-means, and DBSCAN, among other classification, regression, and clustering techniques, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy.

- **Numpy :-**

  NumPy is a Python library that adds support for huge, multi-dimensional arrays and matrices, as well as a large number of high-level mathematical functions to work on these arrays. Many recent large-scale scientific computing applications have needs that are beyond NumPy arrays' capability. NumPy arrays, for example, are often loaded into a computer's RAM, which may be inadequate for massive dataset processing. NumPy operations are also performed on a single CPU.

- **Matplotlib :-**

  Matplotlib is a graphing package for Python with NumPy, the Python numerical mathematics extension. It provides an object-oriented API for embedding charts into programmes utilising GUI toolkits such as Tkinter, wxPython, Qt, or GTK. There's also a procedural "pylab" interface built on a state machine (like OpenGL) that's meant to look like MATLAB, however it's not recommended. [3] Matplotlib is used by SciPy.

4.1 Assumptions and dependencies

- Handling Missing data w/ the linear regression imputation wrt each country
- Collected data is fair.
- Code works with the dependencies listed in the requirement file.
-

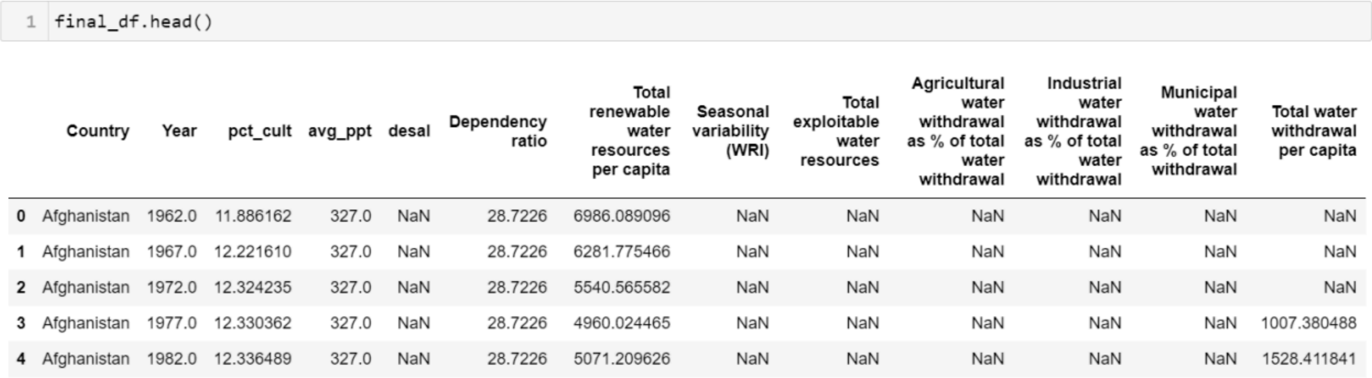
IMPLEMENTATION DETAILS

After retrieving the 5 years dataset of Rainfall and Groundwater for 1962-2017 we performed pre-processing of our data, to further use this data to make a machine learning model.

Steps Involved are:

1. **Data Acquisition:**
    a. The first step of our project involved building a dataset because our goal was to develop a model that incorporates new attributes compared to other models. After doing some initial research,[8] we made a small list of attributes we thought would be useful in classifying a country's stress level. The water consumption and usage database from AQUASTAT proved to be extremely comprehensive, however it needed some transformation (ETL Pipeline) to consolidate multiple tables for the attributes we were interested in

    1. Total land cultivated (%)
    2. Annual precipitation (mm/yr)
    3. Rainwater harvesting awareness
    4. Water consumption per capita (m^3/year/inhabitant)
    5. Total renewable water resources per capita (m^3/year/inhabitant)
    6. Desalination capacity (km^3/year)
    7. Water dependency ratio (%)
    8. Agricultural water withdrawal (%)
    9. Industrial water withdrawal (%)
    10. Municipal water withdrawal (%)
    11. Water stress level (%) [Water stress level measured by dividing total water withdrawal by the total water available minus any water needed for environmental flow.]



```
final_df.head()
```

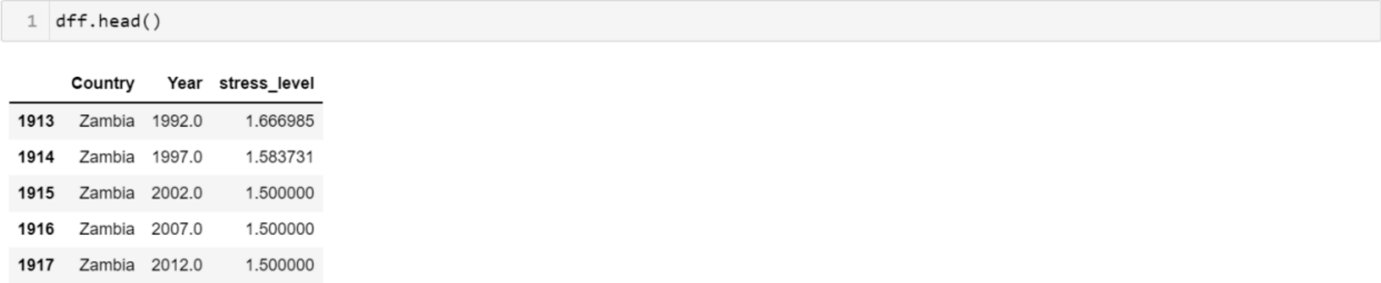
	Country	Year	pct_cult	avg_ppt	desal	Dependency ratio	Total renewable water resources per capita	Seasonal variability (WRI)	Total exploitable water resources	Agricultural water withdrawal as % of total water withdrawal	Industrial water withdrawal as % of total water withdrawal	Municipal water withdrawal as % of total withdrawal	Total water withdrawal per capita
0	Afghanistan	1962.0	11.886162	327.0	NaN	28.7226	6986.089096	NaN	NaN	NaN	NaN	NaN	NaN
1	Afghanistan	1967.0	12.221610	327.0	NaN	28.7226	6281.775466	NaN	NaN	NaN	NaN	NaN	NaN
2	Afghanistan	1972.0	12.324235	327.0	NaN	28.7226	5540.565582	NaN	NaN	NaN	NaN	NaN	NaN
3	Afghanistan	1977.0	12.330362	327.0	NaN	28.7226	4960.024465	NaN	NaN	NaN	NaN	NaN	1007.380488
4	Afghanistan	1982.0	12.336489	327.0	NaN	28.7226	5071.209626	NaN	NaN	NaN	NaN	NaN	1528.411841

FIG. 7

2. Data Set Preprocessing:

Data from our sources were presented in several different formats and contained extra information that was not useful for our application, so we filtered the columns to keep only what was necessary. All the clean datasets were then merged into the final_dataframe. Best-fit model categorizing the output stress levels of each country into one of 3 classes and splits the whole dataset into testing and training subsets.

- Stress Value of 0-40: Low
- Stress Value of 41-80: Medium
- Stress Value of > 81: High

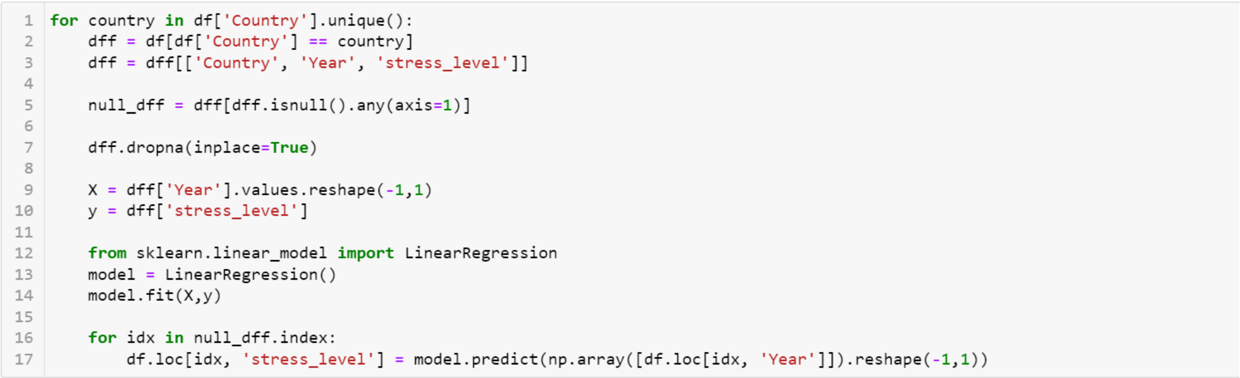


```
dff.head()
```

	Country	Year	stress_level
1913	Zambia	1992.0	1.666985
1914	Zambia	1997.0	1.583731
1915	Zambia	2002.0	1.500000
1916	Zambia	2007.0	1.500000
1917	Zambia	2012.0	1.500000

FIG. 8

Imputing nan values with respect to each country.



```
for country in df['Country'].unique():
    dff = df[df['Country'] == country]
    dff = dff[['Country', 'Year', 'stress_level']]

    null_dff = dff[dff.isnull().any(axis=1)]

    dff.dropna(inplace=True)

    X = dff['Year'].values.reshape(-1,1)
    y = dff['stress_level']

    from sklearn.linear_model import LinearRegression
    model = LinearRegression()
    model.fit(X,y)

    for idx in null_dff.index:
        df.loc[idx, 'stress_level'] = model.predict(np.array([df.loc[idx, 'Year']]).reshape(-1,1))
```

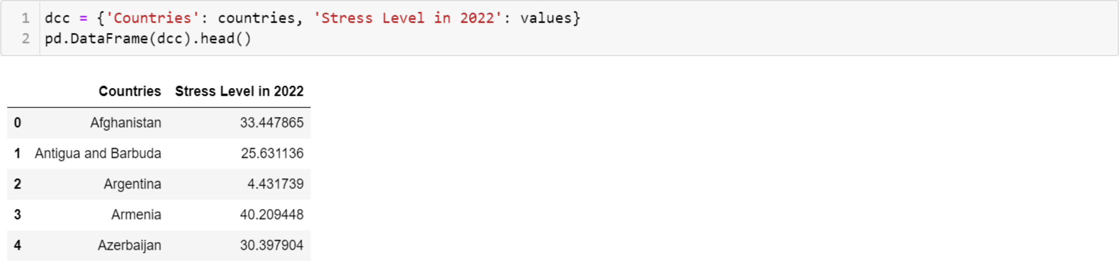
FIG. 9

3. Predictions

The target attribute that is used to train the machine learning model is water stress for a particular country and for a particular year. [4]Using the randomly generated training and testing sets, we ran a series of tests to find the best algorithm that would classify the data. Given the type of inputs and outputs, we suspected using a decision tree or some instance based learner would be effective in classifying the data due to presence of missing attributes.

Once the new set of stress values were generated, we collected this data and used scikit-learn to perform linear regression on each country's stress data over the time period of 1960 to 2014. Using this model,[6] we plotted a graph of the water stress vs. year for each country and used that to find and predict the year when stress crossed a critical level.

Water Stress Level as of 2022

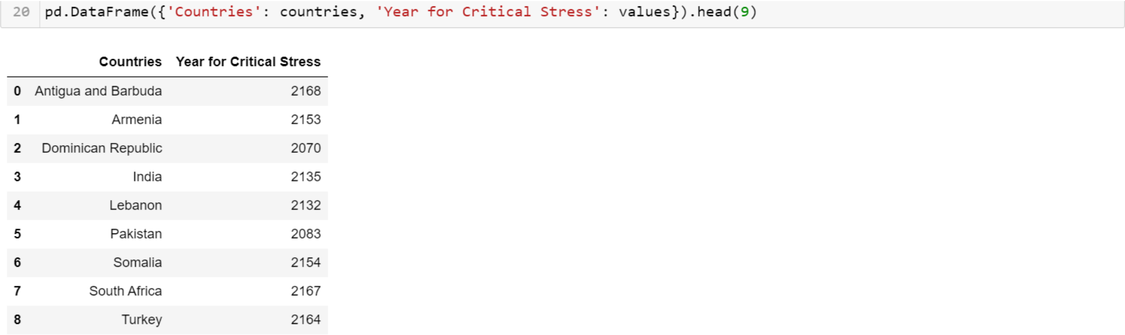


```python
dcc = {'Countries': countries, 'Stress Level in 2022': values}
pd.DataFrame(dcc).head()
```

	Countries	Stress Level in 2022
0	Afghanistan	33.447865
1	Antigua and Barbuda	25.631136
2	Argentina	4.431739
3	Armenia	40.209448
4	Azerbaijan	30.397904

FIG. 10

Countries that may reach a critical water stress level by 2199

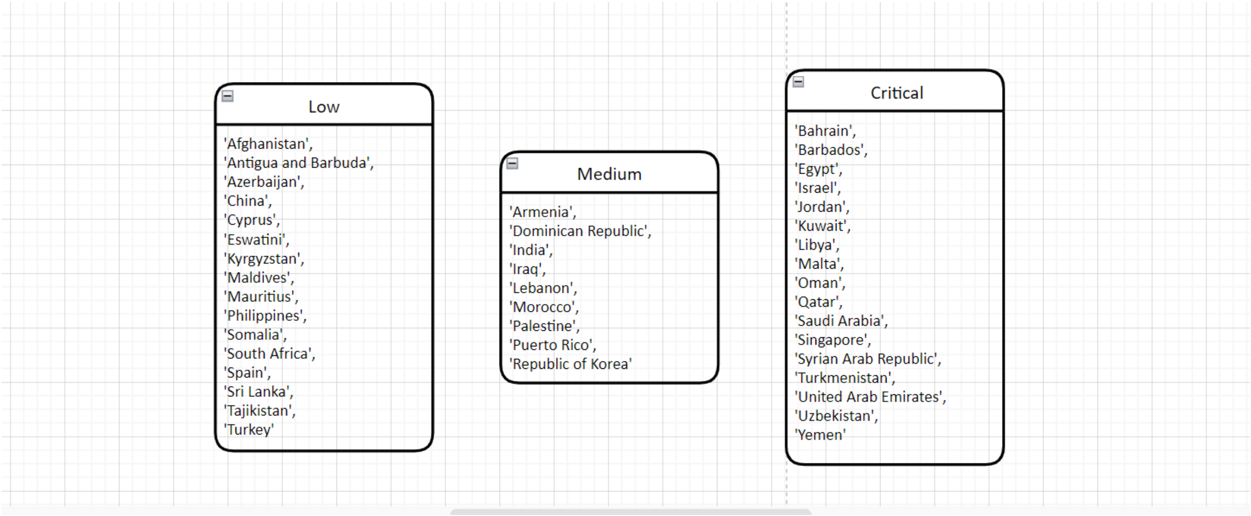


```python
pd.DataFrame({'Countries': countries, 'Year for Critical Stress': values}).head(9)
```

	Countries	Year for Critical Stress
0	Antigua and Barbuda	2168
1	Armenia	2153
2	Dominican Republic	2070
3	India	2135
4	Lebanon	2132
5	Pakistan	2083
6	Somalia	2154
7	South Africa	2167
8	Turkey	2164

FIG. 11

Cluster of countries with respect to water stress



Low

'Afghanistan',
'Antigua and Barbuda',
'Azerbaijan',
'China',
'Cyprus',
'Eswatini',
'Kyrgyzstan',
'Maldives',
'Mauritius',
'Philippines',
'Somalia',
'South Africa',
'Spain',
'Sri Lanka',
'Tajikistan',
'Turkey'

Medium

'Armenia',
'Dominican Republic',
'India',
'Iraq',
'Lebanon',
'Morocco',
'Palestine',
'Puerto Rico',
'Republic of Korea'

Critical

'Bahrain',
'Barbados',
'Egypt',
'Israel',
'Jordan',
'Kuwait',
'Libya',
'Malta',
'Oman',
'Qatar',
'Saudi Arabia',
'Singapore',
'Syrian Arab Republic',
'Turkmenistan',
'United Arab Emirates',
'Uzbekistan',
'Yemen'

FIG. 12

CHAPTER 5

CONCLUSION

The purpose of this study was to identify possible years for day zero water conditions in a country with the help of groundwater stress level value obtained through the water resources available and its consumption across different years in different countries. As observed in the literature review, previously used models used Machine Learning, Predicting Rainfall or Flooding. In addition, they were analyzing the effect of groundwater on storms, floods and rivers. But in this project we used all the information about the impact of availability and consumption of groundwater to calculate the groundwater stress level.

The work done by us not only predicts the stress level of groundwater across different countries but also helps to visualize the increase in water consumption per capita and decrease in the renewable water resources available per capita in the near future.

It is hoped that the analysis used here may provide a basis for predicting water quantity and will lead to a future development agreement.

As expected, the best models were either trees or instance based learners for the training dataset due to presence of missing values. These algorithms did much better than the base test of Zero, showing that a model can certainly be generated with our sample size and attributes. Filled dataset has greater accuracy on the machine learning model in comparison with the original missing data.

Looking at the performance for both datasets, the presence of missing attributes certainly has an effect on how well the algorithm can classify new samples. Hopefully in the near future (and with enough digging around), this dataset can become more complete as more information on each country's water usage is published.

REFERENCES

Journal Article referencing:

[1] Robert A Agana (2017) Deep Learning based approach for long term drought predictions.

[2] Salvatore Pascale, Sarah B. Kapnick, Thomas L. Delworth, and William F. Cooke, “Increasing risk of another Cape Town “Day Zero” drought in the 21st century”.

[3] Dhawan, B. D. (1989) Studies in Irrigation and Water Management (New Delhi, Commonwealth

[4] GEC (1984) Groundwater Estimation Methodology (New Delhi, Ministry of Irrigation).

[5] GEC (1997) Groundwater Estimation Methodology (New Delhi, CGWB).

[6] India Portal Datasets - Water Depth Levels

[7] M.J.Booysen, M. Visser, R. Burger, “Temporal case study of household behavioural response to Cape Town's “Day Zero” using smart meter data”.

[8] Amy Maxmen, “As the Cape Town water crisis deepens, scientists prepare for 'Day Zero”.

[9] Michael B. Richmana, Lance M. Leslieb (2018)

[10] P.H. Herbst; D.B. Bredenkamp; H.M.G. Barker (1966)

[11] Johanna Brühl, Martine Visser (2021)

[12] G. Thomas LaVanchy & Michael W. Kerwin & James K. Adamson (2019)

[13] Wolski, Piotr (2018). How severe is Cape Town's “Day Zero” drought?

[14] Ahmadi, Mohammad Sadeq; SuÅ¡nik, Janez; Veerbeek, William; Zevenbergen, Chris (2020). Towards a global day zero? Assessment of current and future water supply and demand in 12 rapidly developing megacities.

[15] Jeroen F. Warner;Richard Meissner; (2021). Cape Town's “Day Zero” water crisis: A manufactured media event?