**Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский ядерный университет «МИФИ»**

**Department of cyber security**

**Direction: 09.04.04 “Software Engineering”**

**Master’s degree**

**Project’s name: Development of a Decision Support System in The Field of Water Reserve Assessment Using Big Data Processing Frameworks**

**Name: Moussa Al Moussa**

**Group: M22-594**

**Supervisor: Dr. Kireev Vasily S.**

**28.06.2024**

Abstract

The explanatory note contains 69 pages, 25 drawings, 2 tables.

Number of sources used: 44

Keywords: water scarcity risk, water shortage, clean water, Arab world, climate change

The purpose of this work is to predict the date when there will be no more accessible clean water in the Arab countries and to help out the scientists and the main stakeholders to have a reliable way in order to solve this crisis.

In the first section, we have an introduction that includes the main problem, the solution, our aim and scope, the definition of water scarcity risk, the explanation and the actuality of our work, the approaches and methods used, the selected software life-cycle, and the task list of our research.

In the second section, we have written our literature review that includes a lot of explanation of several platforms, and these platforms are mWater, AkvoFlow, and Open Water Data.

In the third section, we have explained about fuzzy cognitive maps, and we have drawn the taxonomy diagram.

In the fourth section, we have included a background that includes some methods that are essential to calculate the water scarcity risk in the Arab world.

In the fifth section, we have presented some interesting infographics and graphs that were compiled from reliable sources.

In the sixth section, we have merged our datasets and we have developed our ETL pipeline architecture.

In the seventh section, we have showed a comparative analysis of several prediction models, and we have selected our proposed approach.

In the eight section, we have developed an approach for water scarcity risk prediction, which includes our math task formulation, our proposed approach, some experiments with real data, and our functional requirements.

In the ninth section, we have developed our water-scarcity risk evaluation system, which includes our proposed system's architecture, our data-logical model, and our system requirements.

In the tenth section, we have developed the water-scarcity risk evaluation system (MVP), which includes presenting our selecting tools for development, our development results, and our conclusion.

In the final section, we have written our references, and we have presented our appendix.

Acknowledgments

As an acknowledgment, I would like to thank my tutor “Dr. Kireev Sergeevich” for his guidance and encouragement provided throughout the period of the course. His patience and valuable time have eased the process of making my research work in a more coordinated manner. On top of that, I would like to thank my family and friends for their support. Moreover, without that support, I couldn’t have succeeded in completing this research work.

**Table Of Contents**

[I. Introduction 5](#_Toc167626524)

[1.1. Abstract 5](#_Toc167626525)

[1.2. What is Water-Scarcity Risk? 5](#_Toc167626526)

[1.3. Problem 6](#_Toc167626527)

[1.4. Solution 6](#_Toc167626528)

[1.5. Aim and Scope 7](#_Toc167626529)

[1.6. Explanation 7](#_Toc167626530)

[1.7. Actuality 7](#_Toc167626531)

[1.8. Approaches and Methods Used 8](#_Toc167626532)

[1.9. Selected software life-cycle 8](#_Toc167626533)

[1.10. Open Data ETL Pipeline Preparation 9](#_Toc167626534)

[1.11. Task List for Our Research 9](#_Toc167626535)

[II. Literature review 10](#_Toc167626536)

[2.1. Introduction 10](#_Toc167626537)

[2.2. mWater 10](#_Toc167626538)

[2.3. Akvo Flow 12](#_Toc167626539)

[2.4. Open Water Data 13](#_Toc167626540)

[2.5. Comparative Table 14](#_Toc167626541)

[2.6. Solution 15](#_Toc167626542)

[III. Fuzzy Cognitive Maps 16](#_Toc167626543)

[3.1. Introduction 16](#_Toc167626544)

[3.2. Grey Maps 16](#_Toc167626545)

[3.3. Interval-Valued Fuzzy Cognitive Maps (IV-FCMs) 17](#_Toc167626546)

[3.4. Dynamic Fuzzy Cognitive Maps 18](#_Toc167626547)

[3.4. Taxonomy Diagram 19](#_Toc167626548)

[3.5. Conclusion 19](#_Toc167626549)

[IV. ETL Pipeline -Background 20](#_Toc167626550)

[4.1. Water Quality Index (WQI) 20](#_Toc167626551)

[4.2. Access to Improved Water Resources 21](#_Toc167626552)

[4.3. Sanitation Coverage 22](#_Toc167626553)

[4.4. Waterborne Disease Rates 23](#_Toc167626554)

[4.5. Water Consumption Patterns 24](#_Toc167626555)

[4.6. Geospatial Analysis 25](#_Toc167626556)

[4.7. Water Stress and Scarcity Indicators 26](#_Toc167626557)

[4.8. Community Surveys 27](#_Toc167626558)

[4.9. Conclusion 28](#_Toc167626559)

[V. The Discovered Datasets 29](#_Toc167626560)

[5.1. Introduction 29](#_Toc167626561)

[5.2. Interesting Infographics and Facts 29](#_Toc167626562)

[5.2.1. The number of people without access to an improved water source 29](#_Toc167626563)

[5.2.2. Share of the population without access to an improved water source 31](#_Toc167626564)

[5.2.3. Number of deaths by risk factor in the Middle East & North Africa in 2019 32](#_Toc167626565)

[5.2.4. Share of The Population With Access to Safely Managed Drinking Water 33](#_Toc167626566)

[5.2.5. Death rate from unsafe water sources from 1990 to 2019 34](#_Toc167626567)

[5.2.6. Share of deaths attributed to unsafe water sources from 1990 to 2019 35](#_Toc167626568)

[5.2.7. Number of people without access to safe drinking water 36](#_Toc167626569)

[5.3. ETL Pipeline - The Main Dataset 37](#_Toc167626570)

[5.4. ETL Pipeline - Other Datasets 38](#_Toc167626571)

[VI. ETL pipeline - Implementation 39](#_Toc167626572)

[6.1. Our Code 39](#_Toc167626573)

[6.2. Merging Datasets 40](#_Toc167626574)

[6.3. ETL Pipeline Architecture 41](#_Toc167626575)

[6.4. Conclusion 41](#_Toc167626576)

[VII. Prediction Approaches of Water Scarcity Risk 42](#_Toc167626577)

[7.1. Comparative Analysis of Prediction Models for This Risk 42](#_Toc167626578)

[7.1.1. Statistical Regression models 42](#_Toc167626579)

[7.1.2. Time Series Models (e.g., LSTM) 42](#_Toc167626580)

[7.1.3. Machine Learning Ensemble Models 42](#_Toc167626581)

[7.1.4. Hybrid Models 43](#_Toc167626582)

[7.1.5. Conclusion 43](#_Toc167626583)

[7.2. Comparative Table of Prediction Models for This Risk 44](#_Toc167626584)

[VIII. Development of Author’s Approach for Water Scarcity Risk Prediction 45](#_Toc167626585)

[8.1. Math Task Formulation -Introduction 45](#_Toc167626586)

[8.2. Math Task Formulation - Determining Weight Values 46](#_Toc167626587)

[8.3. Proposed Approach 47](#_Toc167626588)

[8.4. Experiments With Real Data 47](#_Toc167626589)

[8.5. Functional Requirements 48](#_Toc167626590)

[IX. Development of Water-Scarcity Risk Evaluation System 50](#_Toc167626591)

[9.1. Proposed System’s Architecture 50](#_Toc167626592)

[9.2. Data-Logical Model 52](#_Toc167626593)

[9.2.1. Entities 52](#_Toc167626594)

[9.2.2. Relationships 52](#_Toc167626595)

[9.2.3. ER Diagram 53](#_Toc167626596)

[9.3. System’s Requirements 54](#_Toc167626597)

[9.3.1. Minimum Hardware Requirements: 54](#_Toc167626598)

[9.3.2. Software Requirements: 54](#_Toc167626599)

[X. Water-Scarcity Risk Evaluation System MVP 55](#_Toc167626600)

[10.1. Selected Tools for Development 55](#_Toc167626601)

[10.2. Development Results 56](#_Toc167626602)

[10.3. Conclusion 60](#_Toc167626603)

[XI. References and Appendix 61](#_Toc167626604)

[11.1. References 61](#_Toc167626605)

[11.2. Appendix 65](#_Toc167626606)

**Table Of Figures**

[Figure 1: mwater -number of users 14](#_Toc165637340)

[Figure 2: importance of water on several sectors 15](#_Toc165637341)

[Figure 3: total rainfall in India 16](#_Toc165637342)

[Figure 4: Pros and Cons of Fuzzy Cognitive Maps 18](#_Toc165637343)

[Figure 5: Taxonomy Diagram 21](#_Toc165637344)

[Figure 6: Number of People Without Access to An Improved Water Source 32](#_Toc165637345)

[Figure 7: Share of The Population Without Access to An Improved Water Source 33](#_Toc165637346)

[Figure 8: Number of Deaths by Risk Factor in The World in 2019 34](#_Toc165637347)

[Figure 9: Number of Deaths by Risk Factor in The Middle East and North Africa in 2019 34](#_Toc165637348)

[Figure 10: Share of The Population with Access to Safely Managed Drinking Water 35](#_Toc165637349)

[Figure 11: Death Rate from Unsafe Water Sources From 1990 to 2019 36](#_Toc165637350)

[Figure 12: Share of Deaths Attributed to Unsafe Water Sources From 1990 to 2019 37](#_Toc165637351)

[Figure 13: Number of People Without Access to Safe Drinking Water 38](#_Toc165637352)

[Figure 14: Merging Datasets 42](#_Toc165637353)

[Figure 15: ETL pipeline architecture 43](#_Toc165637354)

[Figure 16: UML Component Diagram For our Proposed System's Architecture 51](#_Toc165637355)

[Figure 17: Data-logical model (ER diagram) 53](#_Toc165637356)

[Figure 18: Current water availability by country 57](#_Toc165637357)

[Figure 19: Future Projection of water availability in the Arab World (National) 58](#_Toc165637358)

[Figure 20: Future Projection of Water Availability in The Arab World (Rural) 58](#_Toc165637359)

[Figure 21: Future Projection of Water Availability in the Arab World (Urban) 58](#_Toc165637360)

[Figure 22: Future Projection of Sanitation Availability in the Arab World (National) 59](#_Toc165637361)

[Figure 23: Future Projection of Basic Hygiene Services in the Arab World (National) 59](#_Toc165637362)

[Figure 24: Future Projection of Basic Water Services in Schools in the Arab World (National) 59](#_Toc165637363)

[Figure 25: Availability of Surface Water by Country 60](#_Toc165637364)

**Table Of Tables**

[Table 1: Literature Review -Comparative Table Between Discovered Similar Platforms 16](#_Toc167626657)

[Table 2: Comparative Table of Prediction Models for This Risk 46](#_Toc167626658)

# I. Introduction

## 1.1. Abstract

Several countries nowadays are worrying about the date when there will be no more accessible clean water in the world [1], and specially, in the Arab countries [2]. That’s why, using some modelling and automation techniques, we are planning to predict the date when there will be no more accessible clean water in every country in the Arab world. On top of that, we are planning to provide other important graphs and data regarding this topic, which could benefit the scientists in order to advance and look at reliable data to help their research. Besides that, several other techniques will be discussed later in our report.

## 1.2. What is Water-Scarcity Risk?

Water-scarcity risk encapsulates the complex interplay of factors that can lead to insufficient water availability, posing significant challenges to ecosystems, communities, and economies. This multifaceted concept extends beyond mere physical water scarcity and encompasses the potential consequences of imbalances between water supply and demand. It considers the intricate relationships between climatic variations, population growth, urbanization trends, and unsustainable water management practices. The risk is not only quantitative but also qualitative, taking into account issues such as water pollution and deteriorating water quality [3]. In essence, water-scarcity risk serves as a comprehensive measure of the vulnerability of a region or community to the impending challenges of water scarcity. Understanding this concept requires a holistic examination of environmental, social, and economic variables, making it imperative for our prediction approaches to go beyond statistical models. As we delve into the prediction approaches of water scarcity risk in the Arab world, we aim to unravel the nuances of this multifaceted challenge, offering a comprehensive understanding that can inform proactive and sustainable water resource management strategies.

## 1.3. Problem

As we discussed earlier in our report, the clean water crisis is a major concern for all of us in the Arab world, since it has severe consequences that will affect all of the humans in the future in a way or another. Besides that, according to the United Nations statistics, there are approximately a whopping 2 billion people that don’t have drinking water. Moreover, according to many estimations, this issue is likely to get worse in the future, and it is projected to have 5 billion people that will face water shortages by 2050. Besides that, this issue may affect not only the poor people, but it also can affect many other industries, such as the agriculture industry, the energy industry, the manufacturing industry, along with many others. That’s why, we need to solve this issue as soon as possible in order to prevent all of that from happening.

Aside from that, the assessment of water scarcity risk in the Arab world addresses the challenge of gathering data related to the clean water issue in Arab countries. Through the evaluation of collected data, we aim to assess the water scarcity risk in the future. This information serves as a foundation for main stakeholders to analyze, enabling them to make informed decisions and implement measures to alleviate future water scarcity. Before delving into the assessment, it is essential to consider various factors, including physical water availability, climate change, population growth, urbanization, agricultural practices, water infrastructure, water quality, political stability and governance, technological innovations, and social awareness and education [4]. These factors form the basis of our comprehensive assessment methodology.

## 1.4. Solution

In order to solve the above problems, the best thing to start with is to analyze and present some accurate data regarding this topic in each and every country in the Arab world. That’s why, our solution is to acquire some accurate data from reliable sources from the internet, and to do some modelling and automation systems, that would have a huge role of removing dirty data, analyzing the data, and presenting the data in an easy way so that all the people could have a reliable way to look at accurate statistics that could help them in order to start their research on how to solve the clean water crisis that could happen in the future. Furthermore, the information must be continuously updated to match the high-quality services in this system.

That’s why, after finishing this project, the main stakeholders which are the scientists and the organizations are predicted to have a better way to communicate and to solve this crisis.

## 1.5. Aim and Scope

The main goal from our graduation project is to predict the date when there will be no more accessible clean water in the Arab countries and to help out the scientists and the main stakeholders to have a reliable way in order to solve this crisis. Moreover, as we all know, all the information in our database must be continuously updated in order to match the high-quality services in this system.

## 1.6. Explanation

Assessing the water scarcity risk in the Arab world is a timely and crucial endeavor given the pressing challenges facing the region. The Arab world is characterized by arid and semi-arid climates, making it particularly vulnerable to water scarcity, a threat further exacerbated by factors such as climate change, population growth, and urbanization. With a growing population and expanding urban areas, the demand for water is escalating rapidly. Concurrently, changing climate patterns pose uncertainties in precipitation levels, affecting the reliability of traditional water sources. Agricultural practices, a major water consumer in the region, are also under scrutiny as efficient water use becomes imperative for sustainable development. Against this backdrop, the project aims to assess the multifaceted nature of water scarcity risks in the Arab world, considering physical water availability, climate variability, population dynamics, urban expansion, agricultural demands, water infrastructure, economic factors, and the potential impact of political and social factors [5]. This assessment is crucial for informing policy decisions, fostering international cooperation, and developing innovative solutions to address the impending water scarcity challenges in the region.

## 1.7. Actuality

The actuality of assessing water scarcity risk in the Arab world lies at the intersection of cutting-edge data analysis and technological advancements. Leveraging data-driven methodologies and programming, this project aims to provide a forward-looking perspective on water scarcity risks. By employing predictive modeling and scenario analysis, we seek to forecast potential trajectories of water availability and demand in the Arab region [6]. The integration of advanced technology allows for a more nuanced understanding of climate change impacts, population dynamics, and the effectiveness of potential mitigation strategies. Furthermore, the utilization of Geographic Information System (GIS) tools enables the visualization and spatial analysis of water-related variables, offering a comprehensive view of the geographical distribution of water scarcity risks [7]. The actuality of this assessment is rooted in its ability to harness the power of data, programming, and technological innovation to not only anticipate challenges but also to facilitate evidence-based decision-making for sustainable water resource management in the Arab world.

## 1.8. Approaches and Methods Used

In our system, we will tend to apply several techniques during the life cycle of this system’s development. The first technique is to identify the risks before beginning to implement the project, and after that, we have to monitor it during the implementation to get a high value [8]. Secondly, we are going to use the “iterative and incremental” approach, which is going to facilitate the process of developing our system in case we’ve got any type of errors or if we want to check and edit something at a previous stage in our project [8]. Furthermore, we are going to split our project into several tasks according to the schedule that we will mention in the schedule’s section. These and other techniques, will be used in the implementation of our project in order to get a good product at the end of our research work.

## 1.9. Selected software life-cycle

Speaking about the main approaches that we can use in software engineering during the implementation part, there is the old waterfall approach, and there is the new iterative and incremental approach. However, as we have mentioned earlier, using the iterative and incremental approach is so much better than using the old waterfall approach which has a lot of disadvantages, such as not being able of returning back to a previous step if we feel the need to do so. Therefore, this problem could be solved easily by using the iterative approach, which allow us to return back to a previous part of our implementation if there is a mistake that we need to fix [9]. Speaking of which, there are five main parts/steps that are involved in every software development project, and these steps are: setting up our requirements, designing our system, implementing our system, integrating our system, and testing our system. Therefore, as an example, if we encountered any problem while testing our system which requires us to go back to fix something in the implementation part, we can do so by using this approach.

## 1.10. Open Data ETL Pipeline Preparation

Before delving into the comprehensive water scarcity risk assessment, we recognize the critical importance of establishing a robust Extract, Transform, Load (ETL) pipeline for open data. This pipeline will serve as the backbone of our research efforts, ensuring that our data is not only accessible but also structured in a manner conducive to analysis and interpretation.

Implementing an ETL pipeline offers numerous advantages, particularly in the context of addressing water scarcity challenges in the Arab world. By systematically extracting, transforming, and loading relevant datasets, we pave the way for clear, usable data that can be readily understood and leveraged by stakeholders. This clarity is paramount for facilitating informed decision-making and strategic interventions aimed at enhancing clean water access across Arab countries.

Furthermore, the establishment of an open data ETL pipeline lays the foundation for future scalability and adaptability of our research endeavours. As we progress through our analytical framework, this pipeline will continue to serve as a cornerstone, enabling seamless integration of additional data sources and methodologies.

In essence, prioritizing the development of an open data ETL pipeline not only aligns with our commitment to a holistic and data-driven approach but also positions us for long-term success in our pursuit of sustainable water resource management in the Arab world [10].

## 1.11. Task List for Our Research

In embarking on this comprehensive water scarcity risk assessment in the Arab world, our research follows a meticulously planned task list designed to ensure a systematic and effective approach [11]. First and foremost, we undertake an observation of current approaches for problem solution. This involves an extensive review of existing methodologies, frameworks, and technologies employed in addressing water scarcity challenges globally, allowing us to glean insights and best practices.

Subsequently, our research endeavors to implement a prediction model, a pivotal step in our analytical framework. We consider the application of sophisticated techniques, including logistic regression or Long Short-Term Memory (LSTM) models, to forecast water scarcity trends. This predictive modeling aims to provide stakeholders with a proactive tool for anticipating future challenges, enabling them to formulate strategic interventions.

Following the implementation of the prediction model, our research moves into the crucial phase of evaluation. Real data, reflective of the complex dynamics of water availability and demand in the Arab world, will be rigorously analyzed to assess the model's accuracy, reliability, and effectiveness. This step ensures that our predictive tools align with the region's unique characteristics and complexities.

Lastly, our research seeks to integrate the developed prediction module seamlessly into the current system's architecture. This integration ensures practical applicability, allowing stakeholders to readily incorporate our findings into their decision-making processes. The task list outlined herein underscores our commitment to a holistic and data-driven approach, aimed at providing actionable insights for sustainable water resource management in the Arab world.

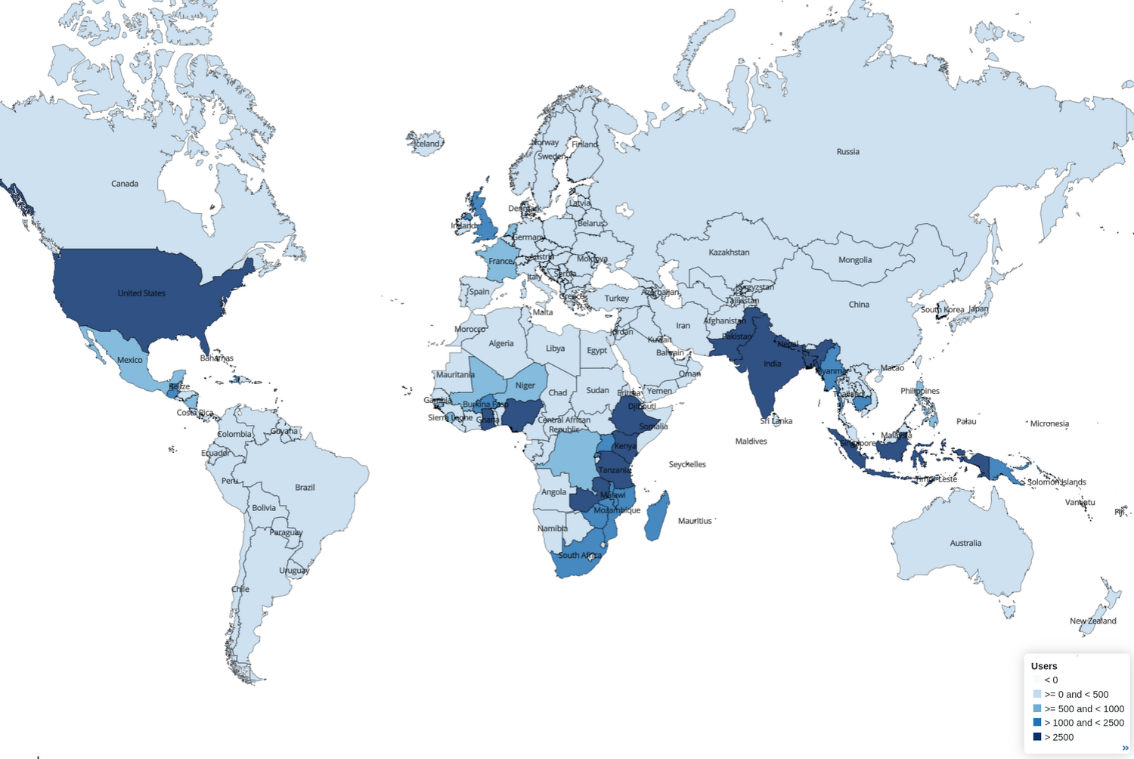
# II. Literature review

## 2.1. Introduction

In this section, we will show some previous applications and websites that are related to this topic (clean water shortage), and we will do a simple summary of them by describing what were their problems, and what they did to solve them. Beside from that, we will discuss also about their designs, how they work, their advantages and disadvantages, and their features [12]. Speaking of which, the main applications that we are going to talk about are “mWater”, “Akvo Flow”, and “Open Water data”.

## 2.2. mWater

“Mwater” is a non-profit company that was established in 2010 for the purpose of resolving problems regarding the lack of clean water. Speaking of which, John Feighery and his wife developed mWater, which has a purpose of recording the precise clean water quality tests on any mobile device. Moreover, using this application, the users can view the data, add pictures, write notes, and add data from new tests. Aside from that, mWater has a lot of advantages, such as it is free and open to everybody, it supports the ability of collecting offline data, and it has built-in data analysis and visualization tools. Furthermore, it includes several features, such as it helps to collect the data using any phone or browser, and it helps with designing and managing surveys. Moreover, it helps with analyzing and visualizing the data related to the clean water, and it is very customizable [13]. However, on the other hand, it has a lot of disadvantages too, such as it is not easy to learn, and it has a limited scope especially regarding water quality monitoring and hydrological modelling.



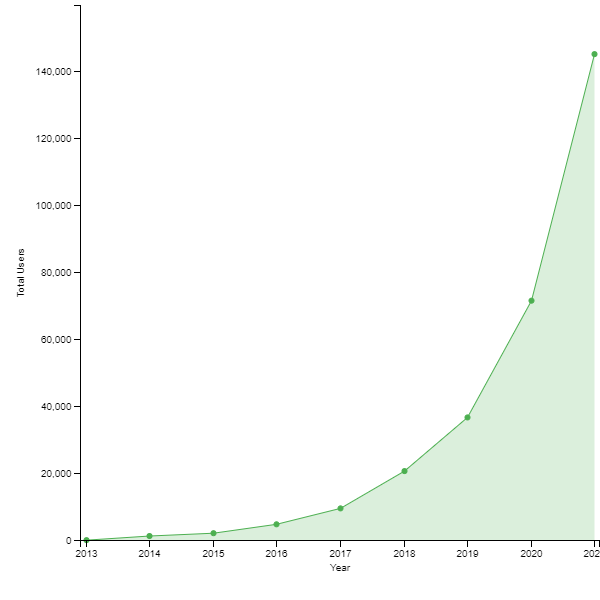


Figure 1: mwater -number of users

As we can see in Figure 1: mwater -number of users, there are a bunch of users who use “mWater” from different parts in the world especially in the US, East Africa, and South East Asia. However, there are no users in a lot of other countries as well, especially in South America, Oceania, and in some parts of Europe. Furthermore, as we can see in the second photo, the number of users who are using mWater is constantly increasing [14].

## 2.3. Akvo Flow

“Akvo Flow” is also a non-profit organization that was established in 2008 by Peter van der Linde and Jeroen van der Sommen in Stockholm. Speaking of which, the main purpose of this application is to provide more accurate data about clean water. On top of that, by having reliable data, all the governments can benefit from that in order to solve the water shortage crisis that could happen in the future. Moreover, after 14 years of establishment, “Akvo Flow” is now available in more than 70 countries across the globe. Aside from that, their main abstract is to “bring data-driven decision making and innovative technology to the development sector and contribute to a better world” [15]. However, while talking about its advantages, it is necessary to mention that it is easy to collect data, it supports the ability of collecting offline data, and it has a real-time monitoring feature. Furthermore, it includes several other features, such as it has customizable forms, and facilitates quality control processes in order to ensure data reliability and accuracy. However, on the other hand, its disadvantages include that it is not easy to learn, and it has limited integrations with other databases and systems.

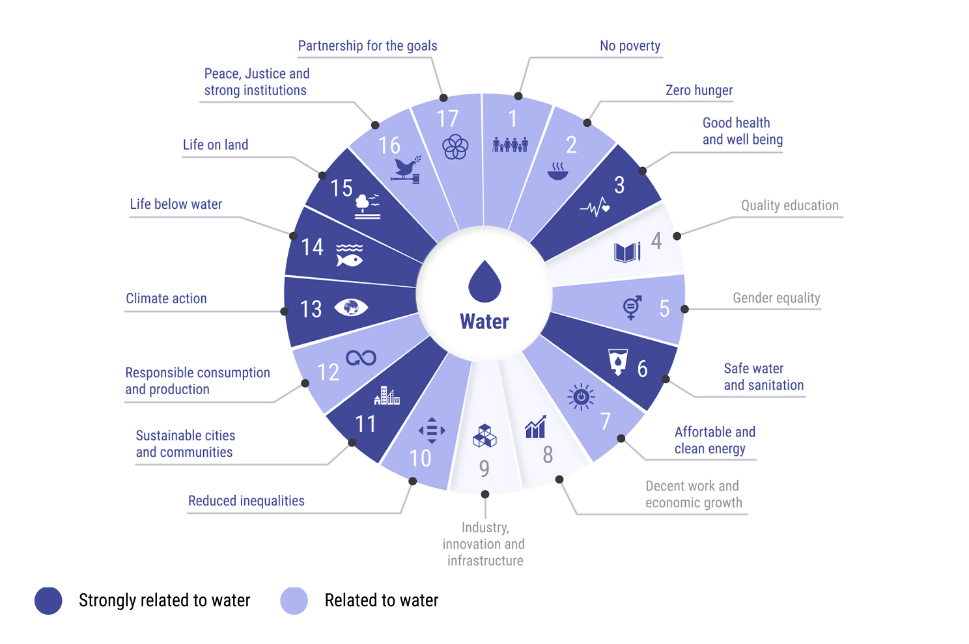


Figure 2: importance of water on several sectors

As we can see in Figure 2: importance of water on several sectors, this figure shows how much several sectors are related to water. Speaking of which, we can conclude that the darker the color, the more related the sector is to solve the water crisis [16].

## 2.4. Open Water Data

“Open Water Data” is also a non-profit organization that was established in 2017 by some software developers from Datameet. Speaking of which, this platform works only in India. Aside from that, its aim was to provide open water data for all of the Indians in order to help them and to help the government to access reliable data regarding clean water in India to solve the clean water crisis. That’s why, the main developers created a map where they show all the data related to this topic. Besides that, it main advantages is that it has comprehensive water data regarding clean water in India, it offers data analysis and visualization tools, and it assists in decision making process. Furthermore, it includes several features, such as it provides access to diverse datasets, and it has some integrations opportunities with other applications and systems. However, on the other hand, its main disadvantage is that it is only available in India, and it is not easy to use.

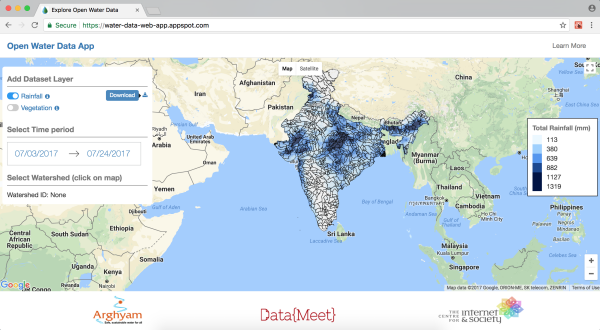


Figure 3: total rainfall in India

As we can see in Figure 3: total rainfall in India, this is a map of India which shows the total rainfall in each part of India [17].

## 2.5. Comparative Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Existing Solutions** | **Advantages** | | **Disadvantages** | **Features** |
| **mWater** | -Easy Data Collection -Available offline -Visualizations -Customizability | -Not easy to use -Limited Scope -Technical requirements -Scalability -Integration -Cost | | -Data collection -Visualizations -Real-time monitoring -Customizability -Mobile and Web Compatibility |
| **AkvoFlow** | -Easy Data collection -Available offline -Customizable forms -Real-time monitoring -Visualizations | -Limited integration -Not easy to use -Not available in all countries | | -Mobile support -Customizable forms -Offline data collection -Real-time monitoring -Visualizations |
| **Open Water Data** | -Comprehensive data -Visualizations -Decision support -Integration opportunities | -Data availability -Technical expertise | | -Diverse datasets -Visualizations -Decision support -Open-Data approach |

Table 1: Literature Review -Comparative Table Between Discovered Similar Platforms

## 2.6. Solution

As we can see clearly, all of the above-mentioned platforms have lots of disadvantages that motivated us to do our project which will be available for each and every country in the Arab world. On top of that, we will be trying to have a system that doesn't have the disadvantages mentioned above, and we will make our system easy to use, free of charge, and scalable, such that it has to have enough integration with other databases and systems, and it must have diverse datasets, offline data collection, and real-time monitoring and visualizations.

# III. Fuzzy Cognitive Maps

## 3.1. Introduction

Fuzzy Cognitive Maps are special maps that are designed for modelling for the complex designed systems. Furthermore, it is designed to show the main relations between its elements, and to compute and show the strength and the degree of impact that these elements have between one another. Moreover, Fuzzy Cognitive Maps are used to describe many aspects in the behavior of complex systems to show its dynamics, and that’s why, there are many types of Fuzzy Cognitive Maps in software engineering [18]. However, the main variants that we are going to mention in this report are: grey maps, Interval-valued Fuzzy Cognitive Maps (IV-FCMs), and dynamic fuzzy cognitive maps.

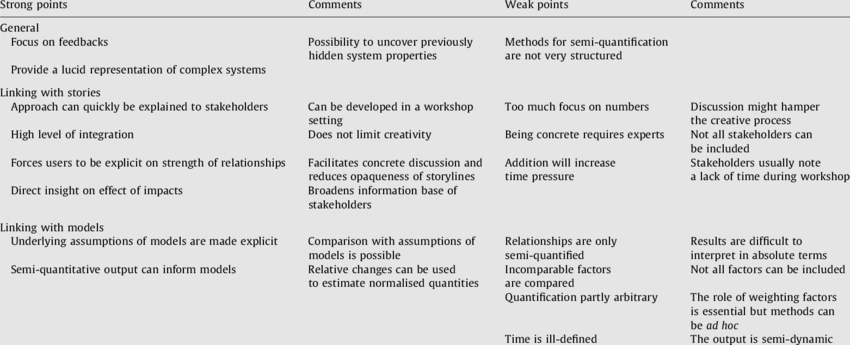


Figure 4: Pros and Cons of Fuzzy Cognitive Maps

As we can see in Figure 4: Pros and Cons of Fuzzy Cognitive Maps, despite of the fact that Fuzzy Cognitive Maps have lots of advantages as discussed earlier, they have some disadvantages too, such as there is too much focus on the numbers, and there is a need for some experts to understand and draw them [19].

## 3.2. Grey Maps

The grey maps, also referred to as “Grey cognitive Maps”, are special types of Fuzzy Cognitive Maps that are used in order to incorporate the grey systems theory. Speaking of which, the latter has a role of handling problems with limited information [20]. Therefore, we can conclude that the grey maps extend the capabilities of the Fuzzy Cognitive maps by including uncertainty. Aside from that, these are some general information about grey maps:

* **Concept’s nodes:** The concepts’ nodes are the general components in grey maps. On top of that, these nodes are connected between one another in order to represent different kinds of interconnections and relations between these nodes, and these nodes could represent various entities.
* **Grey relations**: In grey maps, we have to represent some relationships between our concepts’ nodes in order to represent some values (high, low), and to incorporate grey numbers that are used to express the uncertainties by specifying lower bounds and upper bounds.
* **Uncertainty handling:** Since we know that the grey maps are used to handle problems with limited information, uncertain handling is used when we deal with unknown numbers instead of assigning accurate values, which has much more flexibility than the general fuzzy cognitive maps.
* **Limited information:** As we mentioned earlier, grey maps are very useful regarding dealing with limited information. Moreover, they are also useful when dealing with uncertain values overall.
* **Decision making:** Since we have uncertain information in grey maps, we can use decision makings to accommodate this ambiguity and the uncertain information by considering several possibilities and evaluate the consequences of different actions.
* **Prediction and forecasting:** Since we have a lot of uncertainty in the relationships, grey maps could be used in order to provide more realistic forecasts and predictions.

## 3.3. Interval-Valued Fuzzy Cognitive Maps (IV-FCMs)

Similar to the grey maps, interval-valued fuzzy cognitive maps are also special types of Fuzzy Cognitive Maps that are used to handle problems with limited information. However, the main difference between them is that the grey maps specify upper and lower bounds for the values, while the IV-FCMs represent the uncertainty level using some kinds of intervals that have a role of defining a range of possible values with the degree of their membership within the interval. Another difference is that in grey maps, the relations between the nodes determine the uncertainty level, whereas in the IV-FCMs, the nodes themselves determine the uncertainty level. Moreover, we can say that the grey maps are generally more flexible than the IV-FCMs because they can easily capture the uncertainty level in a non-probabilistic way, whereas the IV-FCMs represent the uncertainty using intervals. On top of that, IV-FCMs focus on the quantity instead of the quality, which is more or less represented better in the grey maps [21]. Aside from that, these are some general information about IV-FCMs:

* **Representation of imprecision:** IV-FCMs capture uncertainty by using intervals, and by that, I mean that these intervals represent some possible values for each concept or node.
* **Interval bounds:** IV-FCMs can have different types of bounds such as the CRISP bounds which have a role of representing precise range of values.
* **Membership degrees:** IV-FCMs assigns memberships degrees using intervals, which have a role of indicating the degree to which each value belongs to its concept.
* **Relationships and connections:** IV-FCMs represents the main relations between the concepts using some connections that can have some weights to represent the strength of influence of the connection.
* **Decision making and analysis:** Since we have some degrees of uncertainty, IV-FCMs helps with providing a framework to analyze complex systems.
* **Handling limited information:** As mentioned earlier, IV-FCMs are useful with uncertain values by allowing a broader range of possible values to be represented using some intervals.

## 3.4. Dynamic Fuzzy Cognitive Maps

Dynamic Fuzzy Cognitive Maps, also referred to as “DFCMs” are also special types of Fuzzy Cognitive Maps that are used to capture the temporal dimension and to model dynamic systems. Just like the previous two variants, DFCMs can also be used to handle problems with limited information, but their primary focus is to represent dynamic systems [22]. Aside from that, these are some general information about grey maps:

* **Temporal modelling:** DFCMs have a role of capturing the temporal dimension of any system by representing the evolution of the concepts over a specific period of time.
* **State transition:** DFCMs have a role of modelling state transition by representing the evolution of the values of concepts over time, and that is done by using some rules and equations in order to describe the influence of the past states on the future states.
* **Time-dependent relationships:** DFCMs have a role of representing the relations between concepts over time, which could be changed and influenced by many factors, such as the evolving causal relationships.
* **Simulation and prediction:** As mentioned earlier, DFCMs have a role of simulating the behavior of dynamic systems in general, and that is done by updating the values of concepts over time, which enables the capability of predicting the future states.
* **Long-term analysis:** DFCMs are very helpful to study the long-term behavior of any system, and that is done by analyzing the resilience and the stability of interventions on the system.
* **Sensitivity analysis:** DFCMs have a role of analyzing sensitivity in order to know the impact level of uncertain values to the behavior of the system, and that can be done by modifying the input parameters within their range.

## 3.4. Taxonomy Diagram

Taxonomy diagram is a diagram that has a role of categorizing the main items in a hierarchical way in order for us to understand and classify our main ideas in a better way. However, in our case, since we are talking about the different types of fuzzy cognitive maps, we decided to do a taxonomy diagram in order to see the main differences between them, and how we can apply them in real life.

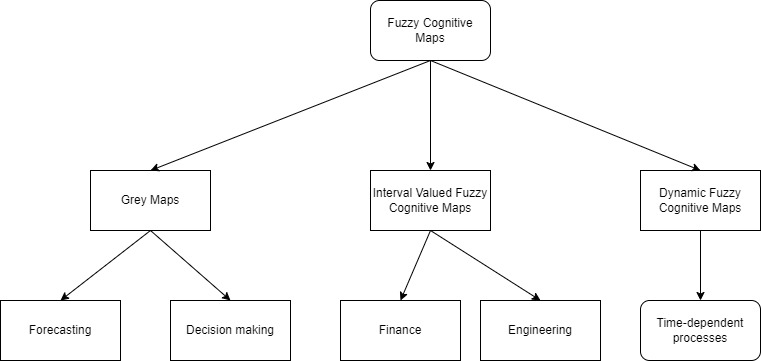


Figure 5: Taxonomy Diagram

As we can see in Figure 5: Taxonomy Diagram, there are different types of fuzzy cognitive maps, and in our case, we are talking about the grey maps, the interval valued fuzzy cognitive maps, and the dynamic fuzzy cognitive maps. On top of that, we decided to do our taxonomy based on the main applications that we can use them in. Speaking of which, we can use the grey maps for forecasting and decision making, we can use the interval valued fuzzy cognitive maps to deal with uncertain data such as finance and engineering, and we can use dynamic fuzzy cognitive maps for modelling dynamic systems such as the time-dependent processes.

## 3.5. Conclusion

As we can see, we can conclude that the fuzzy cognitive maps help us to model the complex designed systems, and it is designed to show the main relations between its elements, and to compute and show the strength and the degree of impact that these elements have between one another. Moreover, it helps us to describe many aspects in the behavior of complex systems to show its dynamics. That’s why, it is very essential to draw Fuzzy Cognitive maps in our system.

# IV. ETL Pipeline -Background

As previously discussed, developing the pipeline of open data in the field of clean water access issue in the Arab countries is very crucial. By doing so, we have the chance to integrate, represent, analyze, and visualize the related data to this field. On top of that, by having a pipeline of open data ETL, we would have a reliable way to represent data that are related to our study to the main users and stakeholders so that they can benefit from it in order to solve the clean water access issue in the Arab countries. Otherwise, the main stakeholders would spend a lot of time in order to gather the data, analyze it, and to visualize it in a professional manner. Even if the information and the data is already presented in some reliable websites, they would spend a lot of time to organize it very well from different resources, and to analyze it, and to visualize it. Therefore, with the data in a single secure source, the main stakeholders could benefit from it drastically.

Since there are a lot of countries that are considered to be Arab, we decided to collect the data for each one of them separately, and then merge all the data into one database [23]. Speaking of which, the database could be large in general, but it would largely benefit all of the Arab counties in order to solve their clean water access issue that are happening right now, or could happen in the future. Speaking of which, we decided to collect the data from a reliable resource such as the world bank, which is considered to be a very reliable way to collect our data.

Aside from that, when collecting our data, it is very essential to keep in mind that we have a lot of countries that are involved in our analysis, and we have to keep in mind that there are many different ways to measure the clean water access issue, such as the water quality index (WQI), the access to improved water sources, the sanitation coverage, the waterborne disease rate, the water consumption patterns, the water stress and scarcity indicators, and the community surveys [24]. Speaking of which, each of these measurements has a unique way to measure the clean water access issue, and therefore, we can conclude that measuring clean water access issue requires a multi-dimensional approach, combining qualitative information, quantitative data, and local context. Thus, combining the methods above would provide a comprehensive understanding of the clean water access situation in the Arab countries, and help inform targeted stakeholders to address this issue effectively.

## 4.1. Water Quality Index (WQI)

The water quality index, also referred to as “WQI” is a metric used to measure and to access the overall quality of water based on several water quality parameters. Speaking of which, it has a huge role of providing a quantitative and a standardized measure of water quality, allowing for comparisons across different time periods or locations [25].

Speaking of which, the water quality index is calculated by assigning weights to different water quality parameters and combining their individual scores into a single index value. It might also be noted that the parameters considered can vary depending on the specific index formula used, however, the most commonly included parameters are the dissolved oxygen, the total suspended solids (TSS), the biochemical oxygen demand (BOD), and the concentrations of various contaminants such as phosphates, nitrates, fecal coliforms, and heavy metals.

Aside from that, in order to calculate the water quality index, there are several steps that are involved. Speaking of which, the first step is the selection of parameters which could be done by identifying the water quality parameters to be included in the index based on their relevance to the regulatory standards or to the specific context. After that, the next step is the data collection, which involves gathering water quality data for the selected parameters, which could be obtained from several resources, including the world bank, the government agencies, the research institutions, or the local water quality assessments. After that, the next step is to normalize the data for each of these parameters to a common range or scale using appropriate transformation methods. After that, the next step is the weight assessment, which involves assigning weights to the normalized parameters based on their regulatory standards or their relative importance. After that, the next step is the sub-index calculation, which involves calculating the sub-indices for each parameter, which could be done by multiplying the normalized value by its assigned weight. After that, the next step is the aggregation step, which involves summing up the weighted sub-indices in order to obtain the water quality index value. After that, we need to define a classification scale or system in order to interpret this WQI value, which could be done by categorizing the water quality as very poor, poor, fair, good, or excellent, which has a huge role of allowing for easy communication of the water quality status between our main stakeholders.

Therefore, we can conclude that the water quality index is considered to be a valuable tool in order to assess the water quality trends, to identify the areas of concern, and to prioritize the water management strategies. On top of that, it has a huge role of helping the researchers, the policymakers, the water resource managers, and all the other main stakeholders to make informed decisions regarding the water quality improvement efforts in the Arab countries.

## 4.2. Access to Improved Water Resources

Access to improved water resources is a term that refers to the availability of reliable and safe water sources that meet certain criteria and standards of reliability, quality, and accessibility. Therefore, we can conclude that that access to improved water resources is a very important indicator in order to ensure the access to sustainable and clean water for several purposes, such as sanitation, drinking, and hygiene [26].

Speaking of which, in order to understand more about “Access to improved water resources”, it is necessary to address some key points regarding this topic. In related manner, improved data sources are defined by the United Nations Children’s Funds and by the World Health Organization, which include protected wells, piped water, boreholes, rainwater collection, protected springs, and water vendors. Secondly, improved water resources are designed to provide safe water for drinking and other domestic uses, which ensure a reduced risk of waterborne diseases. Besides that, access to improved water resources considers several factors such as the reliability and the functionality of the infrastructure, and the distance to the water source. Next, it is essential to note that reporting and monitoring are very essential, which involve monitoring the access to improved water resources through censuses, surveys, and data collection efforts, which is mainly done by the international organizations and by the governments. Next, it is also essential to note that there is a huge difference between urban and rural areas. Speaking of which, the rural areas may not have a good infrastructure, which rely more on springs wells, or other decentralized sources. Apart from that, access to improved water resources has a huge role of tracking progress towards achieving a sustainable development goal, which has a big role of ensuring that the clean water is available for drinking purposes. Besides that, it is worth noting that we have to ensure that the clean water is available for everybody, regardless of their location or their socioeconomic status. However, there are a lot of challenges, which involve the low-income countries, the disputed areas, and the remote areas.

Therefore, we can conclude that access to improved water resources is very essential in order to formulate policies in order to achieve access to clean water in the Arab countries, and in order to assess the impact of interventions and to identify the areas of need. It also has a huge role of contributing a lot to improve public health, and to reduce waterborne diseases.

## 4.3. Sanitation Coverage

Sanitation coverage is a term that refers to the percentage of the population that has access to sanitation services or facilities. Speaking of which, it is considered to be a very important indicator that is used to measure the progress made in order to provide hygienic and safe sanitation conditions for the people, and to measure the level of sanitation access [27].

Speaking of which, in order to understand more about the sanitation coverage, it is necessary to address some key points regarding this topic. In related manner, improved sanitation coverage is defined by the United Nations Children’s Funds and by the World Health Organization, which include ventilated improved pit latrines, pour-flush toilets, composting toilets, or pit latrines. Besides that, the main purpose of the sanitation coverage is to provide hygienic and safe disposal of human excreta. On top of that, it plays a major role to prevent a lot of diseases, protect the public health, and to reduce the environmental pollution. Apart from that, sanitation coverage takes into account several factors, such as the ratio of facilities to pollution, the proximity to these facilities, and the availability of separate facilities for both females and males. Furthermore, another key point to consider is monitoring, which is mainly done by the main government and the international organizations to monitor the sanitation coverage by using some censuses and surveys in order to identify the gaps, track the progress, and inform some policies. Next, it is essential to note that open defecation is very essential, which is basically defined as the lack of access to any toilet facility, which is a very crucial aspect that is related to sanitation coverage. Next, it is also essential to note that the sustainable development goal indicator is very important since it has a huge role of tracking the progress towards achieving the sustainable development goal by ensuring the availability and the sustainable management of water. Apart from that, achieving a high sanitation coverage requires a lot of hygiene promotion efforts and community education, awareness, and behavior change. Lastly, we have to ensure the equitable access to improved sanitation facilities, and we have to ensure that everybody can access improved sanitation facilities regardless of their location or socioeconomic status.

Therefore, we can conclude that sanitation coverage is very essential in order to achieve dignity, public health, and environmental sustainability in the Arab countries, and to reduce waterborne diseases in order to improve the overall hygiene practices, and to promote a healthier and cleaner environment.

## 4.4. Waterborne Disease Rates

Waterborne disease rate is a term that refers to the incidence of diseases that are mainly transmitted through contaminated and polluted water. Speaking of which, there are a lot of diseases that occur when people consume or come into contact with polluted water that is contaminated with bacteria, parasites, or even viruses [28].

Speaking of which, in order to understand more about the waterborne disease rates, it is necessary to address some key points regarding this topic. In related manner, waterborne diseases include various illnesses such as hepatitis, cholera, typhoid fever, along with many others. On top of that, these illnesses vary depending on the sanitation conditions and on the region. Besides that, the waterborne diseases could be transmitted through several routes, such as drinking polluted water, having contact with contaminated water, or even by consuming contaminated food prepared with contaminated water. Aside from that, waterborne diseases have significant impacts on the global health in general. Speaking of which, they can contribute to mortality rates, and several sorts of illnesses. Furthermore, it is essential to note that the elderly, the children, and the people with relatively compromised immune systems are at a very high risk of dying because of these diseases. Therefore, we need to prevent the waterborne diseases, and that could be done by implementing measures in order to ensure that all people have access to safe and clean water, which could be done by implementing water quality monitoring systems, improving water treatment processes, and promoting proper hygiene and sanitation processes. Furthermore, another key point to consider is monitoring, which is mainly done by the main government and the international organizations to monitor the waterborne disease rates by using some epidemiological studies and surveillance systems, which has a huge role of identifying the disease outbreaks, understanding disease patterns, and assessing the effectiveness of interventions. Next, it is essential to note that water safety plans are very essential, which are defined as comprehensive risk management approaches, that have a huge role of addressing potential risks that are associated with water supply systems, which involve identifying the main risks at each stage of the water supply chain, from the source of the water to consumption. Next, it is also essential to note that the climate change has a huge influence on the waterborne diseases, and by that, I mean that extreme weather conditions and rising temperatures can increase the risk of diseases’ transmission. Lastly, we have to improve the water treatments by promoting proper hygiene and sanitation practices, and that could be easily done by conducting some health-education campaigns.

Therefore, we can conclude that reducing the waterborne diseases is very crucial to increase our public health goal, and that could be done by improving the water quality, enhancing sanitation infrastructures, promoting hygiene, and making effective public-health policies.

## 4.5. Water Consumption Patterns

Water consumption patterns refer to the patterns of water usage by people, or even industries. Aside from that, water consumption patterns are very crucial in order to manage the water resources in an effective way, to identify areas of high-water usage, and to implement water conservation measures [29].

Speaking of which, in order to understand more about the water consumption patterns, it is necessary to address some key points regarding this topic. In related manner, water consumption patterns have a huge role of measuring the average water consumptions within a specific region. On top of that, it has a role of providing a baseline for assessing water consumption and demand trends over a specific period of time. Speaking of which, we can measure the per capita water usage, which may vary a lot from country to country. Moreover, we can also measure the household water consumption in order to identify the main water uses, which include cooking, bathing, drinking, toilet flushing, dishwashing, and outdoor irrigation. Besides that, we can also measure the sectoral water consumption, including several sectors such as agricultural, industrial, domestic, and commercial sectors. Speaking of which, each sector has its own water requirements and water usage patterns. Furthermore, another key point to consider is measuring the peak water demand, which refers to the periods when water consumption is highest, which has a huge role of helping in designing better distribution systems and water infrastructure. Next, it is essential to note that seasonal variations can highly influence the water consumption patterns, which are influenced by several factors, such as the climate conditions, the agricultural practices, the tourism, and the irrigation needs. Next, it is also essential to note that water consumption patterns have a huge role of identifying the areas that have high water usage or inefficiencies, which is done by analyzing those patterns, which helps the government and the other main stakeholders to implement measures to improve water use efficiency, which could be done by promoting behavioral changes, adopting water-saving technologies, or even by optimizing irrigation practices. Aside from that, one key to consider is that we can use smart water metering technologies, which have a big role of providing real-time data on water consumption, which could be very beneficial for the main stakeholders to monitor the water usage patterns, promote conservation, and identify leaks, which can help raising awareness about clean water consumption. Lastly, we have to manage the water demand, which involves understanding how and when the water is consumed, and then, the government could implement some measures to manage the water demand, and that could be done by doing some pricing structures or even public awareness campaigns.

Therefore, we can conclude that monitoring the water consumption patterns is very crucial in order to achieve sustainable water resource management in areas facing population growth, water scarcity, or even climate change impacts. Moreover, identifying the areas of conservation is also very crucial to contribute to long-term water sustainability, and to promote water-efficient practices.

## 4.6. Geospatial Analysis

Geospatial analysis is a method of interpreting and analyzing data that has spatial or geographic component. It combines statistical analysis, geographic information system (GIS) and spatial modelling techniques, which involves understanding and examining patterns, trends, and relationships within a geographic region in order to drive meaningful insights [30].

Speaking of which, in order to understand more about the geospatial analysis, it is necessary to address some key points regarding this topic. In related manner, geospatial analysis works with spatial data, which is tied to specific geographic locations. Speaking of which, it includes several types of data such as coordinates of landmarks, road networks, boundaries of regions, and satellite imagery or elevation data. Aside from that, geospatial analysis has a huge role of examining the interactions and the relationships between spatial features, and that is done by exploring how adjacency, proximity, connectivity, containment, or distance impact the data that is being analyzed. Besides that, it has a big role of enabling the representation of complex spatial patterns in a visual way. Speaking of which, we can visualize the data using maps that can display several attributes, such as land use, population density, environmental factors, or even infrastructure. Furthermore, another key point to consider is querying and selecting spatial features based on spatial relationships or other criteria. Next, it is essential to note that there is a very crucial technique called “spatial interpolation”, which is a technique used in geospatial analysis, which is used to analyze and estimate values at unsampled locations based on observed values at nearby locations. Speaking of which, this technique is very useful to create maps or continuous surfaces from scattered point data. Moreover, it is also essential to note that there are a lot of other spatial analysis technique such as butter analysis, overlay analysis, hotspot analysis, spatial clustering, geostatistics, and network analysis, that allow the main stakeholders to examine spatial patterns, make predictions, identify trends, and explore spatial relationships. Next, it is also essential to note that there are several applications in various fields that are related to geospatial analysis, including environmental management, urban planning, transportation, public health, agriculture, natural resource management, business location analysis, and emergency response. Speaking of which, the main stakeholders could benefit from these applications in order to make decisions by understanding spatial patterns in the Arab countries. Lastly, we have to use the geographic information systems (GIS), that have a huge role of providing the capability to analyze, store, and visualize spatial data. On top of that, these systems offer a huge range of spatial analysis functions that could be used to support geospatial analysis workflows.

Therefore, we can conclude that geospatial analysis is a very essential tool that enables us to deeply understand the spatial dimension of several phenomena, which could benefit the main stakeholders to make several decisions in numerous domains in the Arab world.

## 4.7. Water Stress and Scarcity Indicators

Water stress and scarcity indicators are mainly used to assess and measure the demand and the availability of water resources in a given region. Speaking of which, these indicators provide insights into the water supply-demand balance, the potential for water scarcity, and the level of water stress [31].

Speaking of which, in order to understand more about the water stress and scarcity indicators, it is necessary to address some key points regarding this topic. In related manner, water stress is an indicator that refers to the ratio of water extraction to available water resources. On top of that, it has a huge role of measuring the pressure on water resources that is caused by human activities, and then, reflecting the extent to which the water-demand exceeds the available water supply. Speaking of which, this indicator helps identifying the regions where the available water resources are under pressure, and where there is a potential for water scarcity, which occurs when the water demand exceeds the available water supply, which is measured by comparing the renewable fresh water resources to the water withdrawals in a specific region. On the other hand, the water withdrawal refers to the total amount of water that could be extracted from available clean water resources, such as lakes, rivers, and groundwater, which could be used for several purposes such as irrigation, municipal supply, industrial processes, and livestock watering. Therefore, we can conclude that monitoring the water withdrawal could help us to understand the level of water use, and to identify the main sectors that require high water demand. Furthermore, another key point to consider is the availability of renewable water withdrawal, which refers to the water available from rivers and precipitation, which represents the natural water supply that could be replenished over the time. Next, it is essential to note that water use efficiency indicators are other indicators that deal with the effectiveness of water utilization in several sectors, including industries, agriculture, and domestic uses, which have a huge role of assessing how efficiently the water is being used in order to achieve certain outcomes, such as economic output, crop production, or even the domestic water services. Next, it is also essential to note that the population growth and the water demand are huge factors that influence the water scarcity. Speaking of which, when the population increases, the water demand increases. Aside from that, assessing the availability of clean water through dams and reservoirs is very important in order to understand the availability of clean water, and to manage the available water resources. Lastly, it is also important to consider the climate change impacts, which can influence the availability of clean water in a particular region, because it could include changes in temperature, precipitation, and evaporations rates, which could help u to assess the potential impacts in the future in the Arab countries.

Therefore, we can conclude that stress and scarcity indicators are very essential to manage our available water resources and to formulate and make some policies regarding this crisis in the Arab world.

## 4.8. Community Surveys

Lastly, in spite of the fact that community surveys could be less important than the other factors mentioned above, it is very essential to do them anyway in order to gather some data and information regarding this issue from community members in the Arab countries. Moreover, conducting surveys may help us to understand different experiences, perspectives, preferences, and needs [32].

Speaking of which, in order to understand more about the community surveys, it is necessary to address some key points regarding this topic. In related manner, there are several purposes for conducting community surveys, including program evaluations, needs assessments, policy development, and gathering some input to help the main stakeholders to make decisions regarding this issue. Therefore, we can conclude that they help to capture several opinions and to collect data. Aside from that, we have to consider the fact that we need to make a good design for these surveys, and that is mainly done by determining the surveys’ objectives, developing survey questions, selecting survey methods, considering survey format and length, and ensuring that the survey meets the target community linguistically and culturally. On top of that, we need to use some sampling techniques in order to select some users from the survey, which could be done randomly, or by using other techniques such as cluster sampling methods, or even by using the stratified sampling, which depends on the main survey’s objectives. Furthermore, another key point to consider is to develop a survey questionnaire, which involves making concise and clear questions that have a role of addressing the main survey’s objects. Speaking of which, the questions could be open-ended or close-ended, and they should be designed in order to gather specific information by minimizing the users’ burdens. Apart from that, these surveys could be conducted through several methods, such as telephone interviews, face-to-face interviews, online surveys, or even paper-based surveys. Next, after the responses are collected, it is essential to analyze these responses using some data-analysis techniques, such as qualitative analysis or quantitative analysis depending on the survey. Besides that, it is essential to note that while doing these surveys, it is important to protect the users’ privacies and to ensure the data security and confidentiality. However, conducting these surveys has a huge role of providing an opportunity to engage our community members in several stuffs, such as in data interpretation, survey design, or even action planning, which could impact the survey findings positively. Lastly, the last step to do is to communicate about the survey findings, which is mainly done by the main government or even by private organizations, which has a huge role of ensuring that all the citizens and the community members are well-informed about the survey’s results in order to allow them to be involved in planning, discussion, or even decision-making processes.

Therefore, we can conclude that conducting community surveys is a very essential step to understand the citizens’ perspectives, needs, and preferences, which could help the main stakeholders by providing them with valuable insights for implementing and designing programs and policies in the Arab world.

## 4.9. Conclusion

Finally, we can conclude that measuring clean water access issue requires a multi-dimensional approach, combining qualitative information, quantitative data, and local context. Thus, combining the methods above would provide a comprehensive understanding of the clean water access situation in the Arab countries, and help inform targeted stakeholders to address this issue effectively.

# V. The Discovered Datasets

## 5.1. Introduction

After setting up our measurements by which we want to measure the availability of clean water in the Arab countries, we need to search for reliable datasets in order for us to analyze them and visualize the results accordingly in order to help us with our research work. Speaking of which, in order for the datasets to be reliable, we decided to make our research in four different websites, and these websites were created by UNICEF, GlobalWaters, OurWorldInData, and DataWorld, which are very reliable ways in order to collect our water data. Speaking of which, after entering their main websites, we tried to base our research to be only in the Arab countries, and compare it with the world average and the main improvements that happened regarding this topic in the past few decades. On top of that, not only did we collect valuable excel datasets, but we also collected some valuable diagrams that could help us to identify the main water issues in the Arab world.

## 5.2. Interesting Infographics and Facts

In the OurWorldInData website, we discovered a lot of cool infographics and facts about the access to clean water in the Arab countries [33]. On top of that, these cool facts are presented through several ways such as using some charts or tables which helped us to understand these facts in a better way. Speaking of which, we collected various charts that are related to our topic, and these charts represent the death rate from unsafe water sources from 1990 to 2019, the number of deaths by risk factor in 2019, the number of people without access to an improved water source, the number of people without access to safe drinking water, the share of death attributed to unsafe water resources from 1990 to 2019, the share of the population with access to safely managed drinking water, and the share of the population without access to an improved water source in the Arab countries.

### 5.2.1. The number of people without access to an improved water source

First of all, we would like to represent a chart that represents the number of people without access to an improved water source in the Arab countries from the yar 2000 to 2020, and see whether we are making some improvements regarding this topic or not.

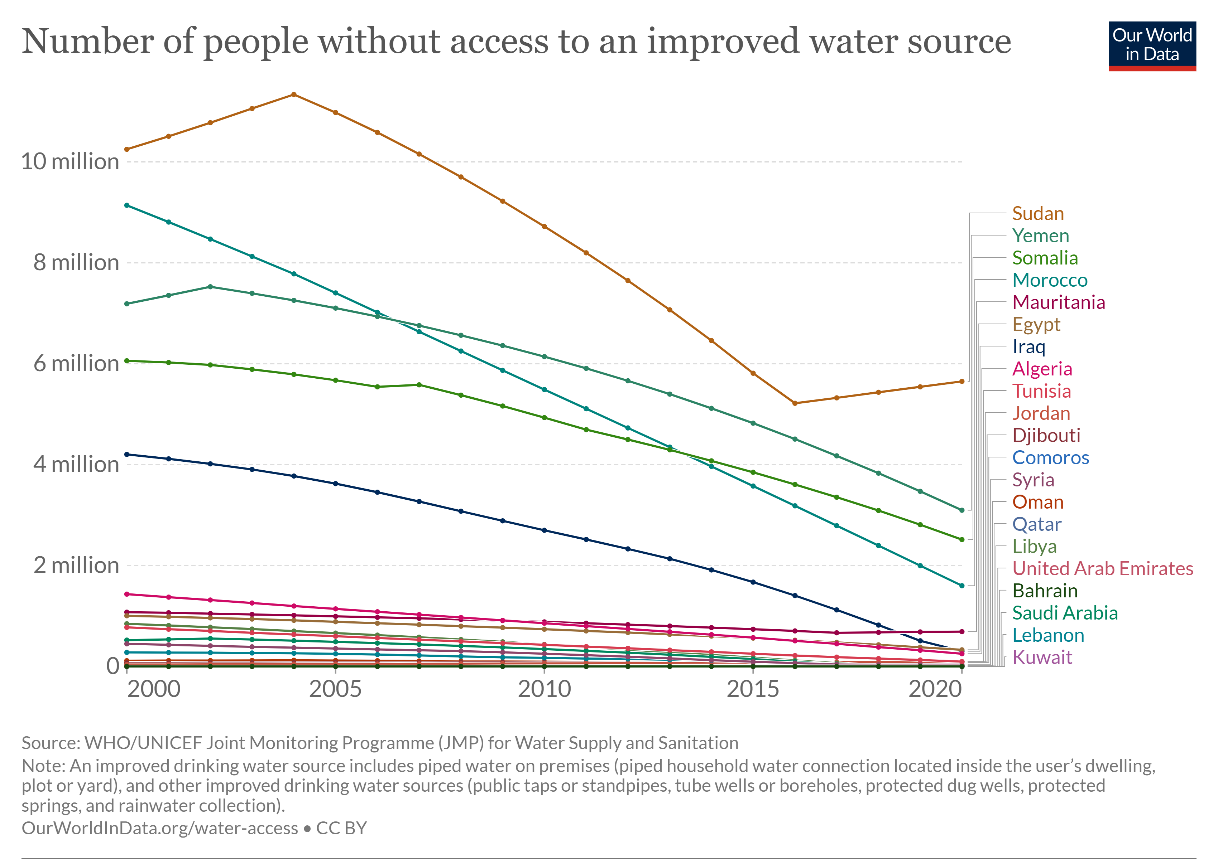


Figure 6: Number of People Without Access to An Improved Water Source

As we can see in Figure 6: Number of People Without Access to An Improved Water Source, this chart represents the number of people without access to an improved water source in the Arab countries. Speaking of which, there are several Arab countries that are doing better than others like Kuwait, Lebanon, and Saudi Arabia, which have less than 2 million people each who have no access to an improved water source, while others are doing very badly, such as Sudan, Yemen, and Somalia, which have over 4 million people each who have no access to an improved water source. However, there are several factors that may contribute to having bad results such as the population of the country, and the climate of that country that may affect our analysis badly. Speaking of which, it is not fair to compare Lebanon, which has a little over 5 million inhabitants and its climate is very mild, with Yemen, which has a little under 33 million inhabitants and its climate is less mild and drier. Nevertheless, apart from Sudan and Mauritania, we can see that most of the Arab countries are having some improvements regarding this issue in the past few decades, because the number of people without access to an improved water source has decreased over the years which is good in general, but there is always a room for more improvement.

### 5.2.2. Share of the population without access to an improved water source

Since we cannot really rely on the first analysis because the number of the population may vary from country to country, we decided to collect another chart from the same source, which has a role of showing the percentage of the population without access to an improved water source, which is considered to be more reliable since it shows the main percentage, regardless of the number of the population of the country.

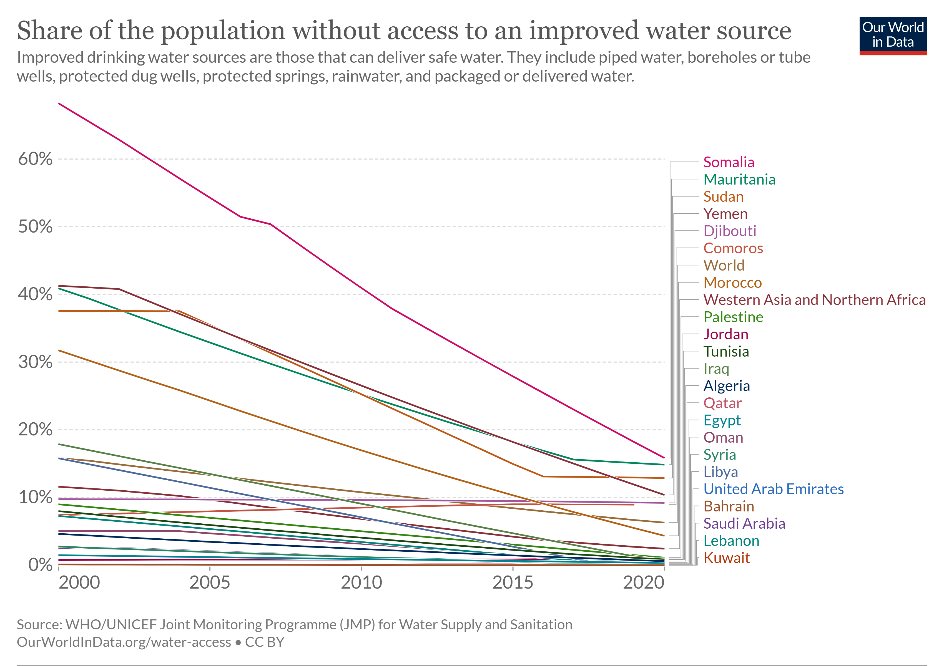


Figure 7: Share of The Population Without Access to An Improved Water Source

As we can see in Figure 7: Share of The Population Without Access to An Improved Water Source, this chart represents the percentage of people without access to an improved water source in the Arab countries. On top of that, we decided to represent also the average percentage of people without access to an improved water source in the world and in western Asia and northern Africa in order to compare the Arab countries with the world regarding this issue. Speaking of which, there are a lot of Arab countries that are doing better than the world average such as Kuwait, Lebanon, Saudi Arabia, along with many others that have a percentage of nearly 1% of the population that don’t have access to an improved water source. However, on the other hand, there are some Arab countries that are doing worse than the world average such as Somalia, Mauritania, Sudan, Yemen, Djibouti, and Comoros, that have a percentage above 9% of the population that don’t have access to an improved water source. However, as we can see, the percentage of the population without access to an improved water source in the Arab countries has decreased drastically over the past few decades, but there is always a room for more improvement.

### 5.2.3. Number of deaths by risk factor in the Middle East & North Africa in 2019

In this section, we are going to represent two figures that show the number of deaths by risk factor in 2019 in both the whole world and in the middle East & Northern Africa which generally speaking, apart from Iran and Turkey, represent the Arab countries.

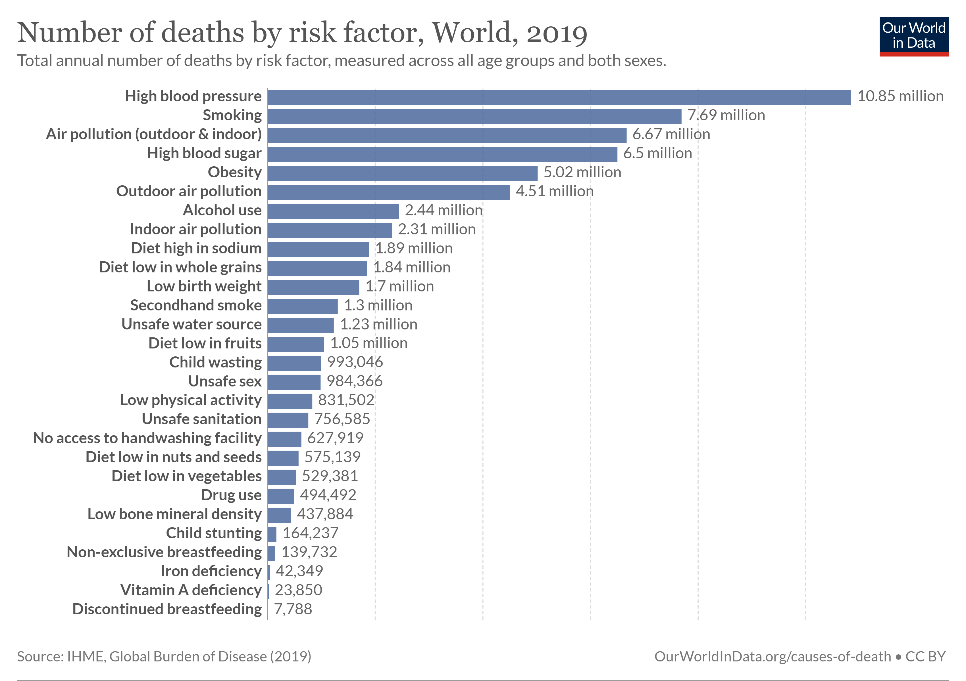


Figure 8: Number of Deaths by Risk Factor in The World in 2019

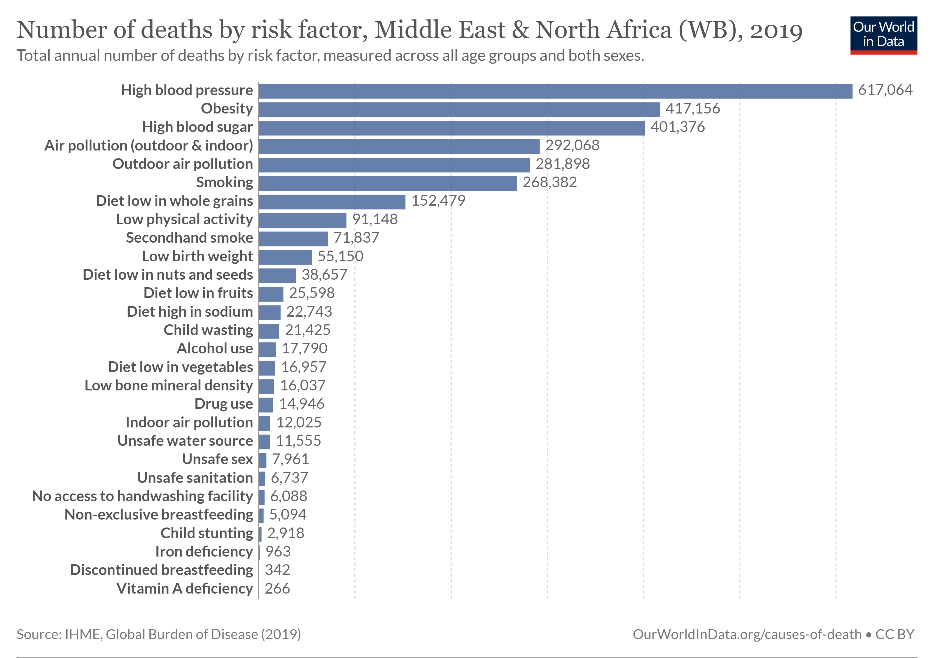


Figure 9: Number of Deaths by Risk Factor in The Middle East and North Africa in 2019

As we can see in Figure 8: Number of Deaths by Risk Factor in The World in 2019, unsafe water source, unsafe sanitation, and no access to handwashing facilities are positioned in the 13th, 18th, and 19th places respectively regarding the number of deaths by risk factor in the world, which are considered to be pretty high in general. However, on the other hand, and as we can see in Figure 9: Number of Deaths by Risk Factor in The Middle East and North Africa in 2019, unsafe water source, unsafe sanitation, and no access to handwashing facilities are positioned in the 20th, 22th, and 23th places respectively regarding the number of deaths by risk factor in the middle East & Northern Africa, which is considered better than the world in general. However, the middle east & Northern Africa is doing worse than the world in other factors, which are not our concerns in our research work.

## 5.2.4. Share of The Population With Access to Safely Managed Drinking Water

Another important factor to consider is the share of the population with access to safely managed drinking water in the Arab countries, because there may be a lot of people that have access to an available water source, but a lot of that water may be not safe to drink in general.

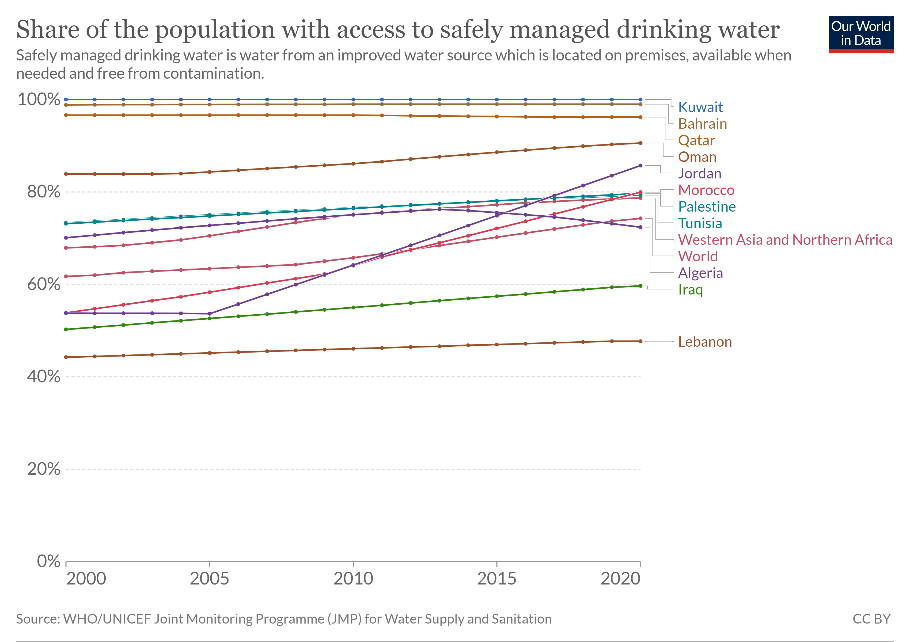


Figure 10: Share of The Population with Access to Safely Managed Drinking Water

As we can see in Figure 10: Share of The Population with Access to Safely Managed Drinking Water, the world has an average of about 70% of the whole population that has access to safety managed drinking water, and of course, there are a lot of Arab countries that are doing better than the average such as Kuwait, Bahrain, and Qatar that have an average of more than 80% of the population that has access to safely managed drinking water. However, on the other hand, there are some Arab countries that are doing worse than the average such as Algeria, Iraq, and Lebanon that have an average of less than 72% of the population that has access to safely managed drinking water. Therefore, we can conclude that Lebanon, despite having a lot of people who have access to an improved water source, most of them don’t have access to safely managed drinking water due to various reasons. Aside from that, apart from Algeria, we can see that most of the Arab countries are having some improvements regarding this issue in the past few decades, but there is always a room for more improvement.

### 5.2.5. Death rate from unsafe water sources from 1990 to 2019

In the main website of OurWorldInData, we have found another interesting chart that represents the number of deaths from unsafe water sources from 1990 to 2019 in the Arab countries per 100000 people, which is fair enough in order to do our comparison.

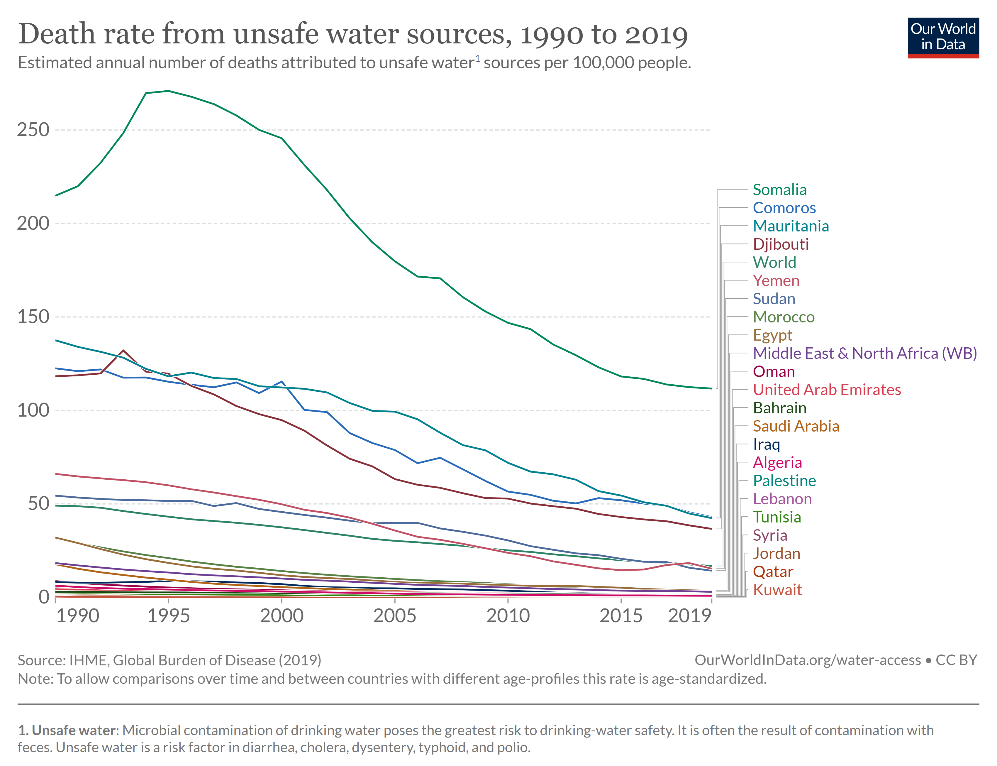


Figure 11: Death Rate from Unsafe Water Sources From 1990 to 2019

As we can see in Figure 11: Death Rate from Unsafe Water Sources From 1990 to 2019, this chart shows the number of deaths from unsafe water sources from 1990 to 2019 in the Arab countries per 100000 people. Speaking of which, there are a lot of Arab countries that are doing better than the world average such as Kuwait, Qatar, Jordan, along with many others that have less than 45 deaths per 100000 people with respect to the unsafe water sources, which is very good. However, on the other hand, there are few Arab countries that are doing worse than the world average such as Somalia, Comoros, Mauritania, and Djibouti, which have more than 45 deaths per 100000 people with respect to the unsafe water sources, which is bad and needs a lot of improvement. However, we can see that the majority of the Arab countries are improving in the past few decades in regards with the number of deaths from unsafe water sources, but there is always a room for more improvement.

### 5.2.6. Share of deaths attributed to unsafe water sources from 1990 to 2019

While in the same topic, in this section, we are more interested to show the percentage of deaths attributed to unsafe water sources in the Arab countries from 1990 to 2019.

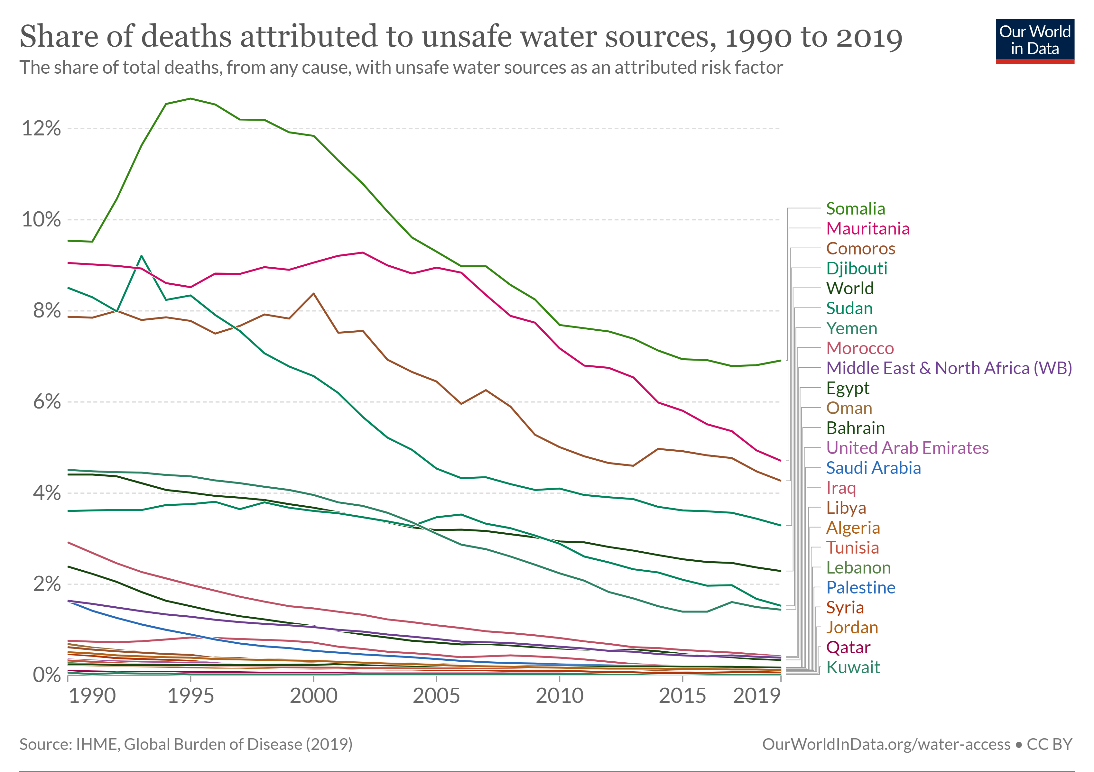


Figure 12: Share of Deaths Attributed to Unsafe Water Sources From 1990 to 2019

As we can see in Figure 12: Share of Deaths Attributed to Unsafe Water Sources From 1990 to 2019, this chart shows the percentage of deaths attributed to unsafe water sources in the Arab countries from 1990 to 2019. Speaking of which, there are a lot of Arab countries that are doing better than the world average regarding this topic such as Kuwait, Qatar, and Jordan, that have approximately 0% of the population that died from unsafe water sources, which is considered to be very good in general. However, on the other hand, there are few Arab countries that are doing worse than the world average such as Somalia, Mauritania, Comoros, and Djibouti, that have between 3% and 8% of the population that died from unsafe water sources, which is considered to be very bad, and there are a lot of decisions that need to be taken in these countries regarding this issue. Aside from that, mostly all of the Arab countries are improving in the past few decades regarding this issue, which is good, but there is always a room for more improvement.

### 5.2.7. Number of people without access to safe drinking water

In this section, we are going to represent a chart that represents the number of people without access to safe drinking water in some Arab countries from the year 2000 to 2020, and see whether we are making some improvements regarding this topic or not.

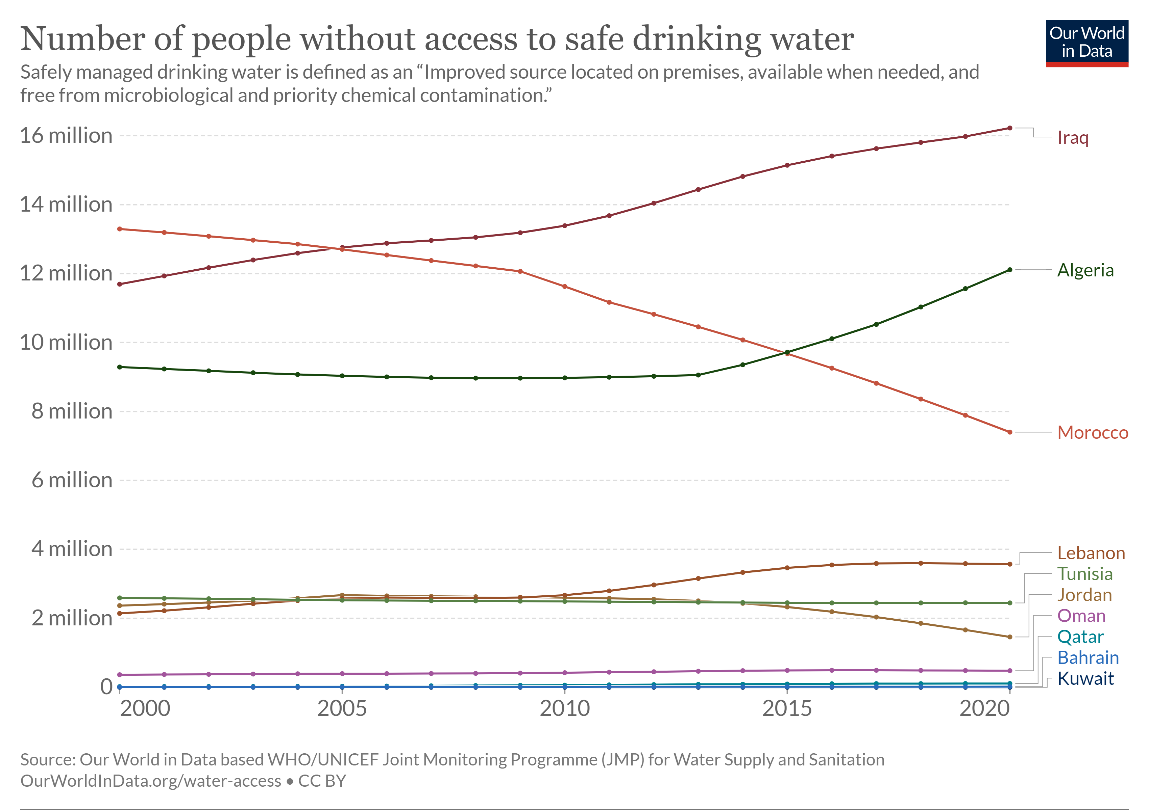


Figure 13: Number of People Without Access to Safe Drinking Water

As we can see in Figure 13: Number of People Without Access to Safe Drinking Water, this chart shows the number of people without access to safe drinking water in some Arab countries from the year 2000 to 2020. Speaking of which, there are a lot of Arab countries such as Kuwait, Bahrain, and Qatar, that have close to 0 people without access to safe drinking water which is very great. However, on the other hand, there are few Arab countries such as Iraq, Algeria, and Morocco, that have over 6 million people each without access to safe drinking water, which is a pretty high number, and there are a lot of decisions that need to be taken in order to make this number lower in these countries in the future. Aside from that, we can see that Algeria, Lebanon, and a lot of other Arab countries are facing some stagnation regarding this issue, which is very concerning and there are a lot of decisions that need to be taken in order to make the situation better in the future regarding this topic.

## 5.3. ETL Pipeline - The Main Dataset

In order to develop the ETL pipeline, it is essential to gather some datasets regarding this topic. Speaking of which, using the UNICEF dataset, we were capable of gathering 1 main dataset and 6 other datasets that are excel files in order to do this task [34]. Besides that, it is quite important to note that after extracting our datasets, it is essential to transform it into a usable format, and then loading it into a target destination in order to store and analyze our data. On top of that, in order to build a great ETL pipeline, we need to understand the content, the structure, and the quality of the data that we are currently working with, and having some datasets will definitely allow us to better understand the data structure, test and validate ETL processes, measure the data quality, monitor performance, along with many others.

First of all, speaking about our main dataset, we tried to gather information about all of the Arab countries that are currently part of the Arab league. Speaking of which, these countries are: Algeria, Bahrain, Comoros, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Qatar, Saudi Arabia, Somalia, Sudan, Syria (also known as the Syrian Arab Republic), Tunisia, the United Arab Emirates, and Yemen. Next, we tried to gather information between the year 2013 and 2020. After that, we tried to put some indicators that we want to make our research based on. Speaking of which, our indicators include the proportion of population with a handwashing facility with soap and water available at home, a limited handwashing facility, and no handwashing facility at home, and the proportion of the population using at least basic sanitation services, basic sanitation services, on-site sanitation facilities with human waste disposed in situ and offsite, improved sanitation facilities, limited sanitation services, improved latrines and other improved facilities, sanitation facilities connected to septic tanks, and sanitation facilities connected to sewer networks. Moreover, it includes the proportion of the population using safely managed sanitation services, sanitation facilities connected to sewer networks and with sewage treated to at least secondary levels, at least basic drinking water services, improved drinking water sources available when needed, basic drinking water services, improved drinking water sources, limited drinking water services, non-piped improved drinking water sources, piped drinking water sources, improved drinking water sources located on premises, improved drinking water sources free from fecal and priority chemical contamination, safely managed drinking water services, surface water, and unimproved drinking water sources. Lastly, it includes the proportion of schools with basic hygiene services, with basic sanitation services, with limited sanitation services, with no sanitation service, with basic drinking water services, with limited drinking water services, and with no drinking water service. Aside from that, it is essential to note that all of these proportions are mainly displayed as percentages out of 100. However, while opening the dataset using the following link: <https://docs.google.com/spreadsheets/d/1O83V3lxbnfMQGNHFmu-EZMl3r1pphxLb/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true> , we can see clearly that there are some values that are displayed as “-“, and this symbol means that there is no data related to that indicator in a particular country in that year. Therefore, this is considered to be dirty data, and we need to handle it.

## 5.4. ETL Pipeline - Other Datasets

While searching in the main website of UNICEF, we have found several other important datasets regarding this topic [34]. Speaking of which, there are two datasets that represent some estimations on the use of water, sanitation, and hygiene per country and per region. There are also two datasets that represent some estimations on water, sanitation, hygiene, waste management, and environmental cleaning in health-care facilities by country and by region. And finally, there are 2 datasets that represent some estimations on water, sanitation, and hygiene in schools by country and by region. Moreover, in order to access these datasets, kindly follow the links in the references down below [35]. Aside from that, it is essential to note that all of these proportions are mainly displayed as percentages out of 100. However, while opening these datasets, we can see clearly that there are some values that are displayed as “- “, and this symbol means that there is no data related to that indicator in a particular country in that year. On top of that, we can see that all of the countries and the regions in the world are represented in these datasets, thus, we need to filter the data to see only the main data that is related to the Arab countries. Besides that, the datasets related to health-care and schools were only made in 2021, and the two datasets that represent some estimations on the use of water, sanitation, and hygiene per country and per region were mainly conducted between 2015 and 2020. Furthermore, it is essential to note that all of these datasets have several sheets in order to separate the indicators which is good.

Aside from that, as you will see in the next section that is related to the ETL pipeline, we decided to merge these six datasets into 1 mega database.

# VI. ETL pipeline - Implementation

## 6.1. Our Code

Speaking about the ETL, we successfully managed to extract our data from a trusted resource. Therefore, we have completed the first step of ETL. However, there are two steps left which are transformation and loading. Speaking of which, using google collab, we managed to create a file called “Developing the pipeline of open data ETL in the field of clean water access issue in the Arab countries.ipynb” which could be found in this link: <https://colab.research.google.com/drive/1dksNRU-2am_UtMIOECmBRFr16h8Pnu-6?usp=sharing>. Speaking of which, we firstly imported our required libraries, and then, with the help of google drive, we were able of importing our seven datasets successfully. After that, we decided to check whether there is dirty data in them, and it turned out that the only dirty data that could impact our analysis is the presence of the “- “symbol, which as discussed above, means that there is no data related to that indicator in a particular country in that year. However, since we don’t know exactly the main values of these symbols, we decided to handle this dirty data by simply leaving it as it is. After that, we decided to start merging our data by firstly merge the two datasets that represent some estimations on the use of water, sanitation, and hygiene per country and per region into a new merged dataset called “general data -merged.xlsx”, which was mainly done by merging the three sheets that represent the water, the sanitation, and the hygiene in both of these datasets to represent both the datasets per country and per region in the same sheet. Speaking of which, as an example, the water sheet in the first dataset that represents the data per country and the water sheet in the second dataset that represents the data per region were merged into one sheet called water that represents both the data per country and per region, and the same thing was done for the other two sheets. Next, we decided to merge the two datasets that represent some estimations on water, sanitation, hygiene, waste management, and environmental cleaning in health-care facilities by country and by region into a new merged dataset called “health care facilities -merged.xlsx”, which was mainly done by merging the five sheets that represent the water, the sanitation, the hygiene, the waste management, and the environmental cleaning in both of these datasets to represent both the datasets per country and per region in the same sheet. Speaking of which, as an example, the water sheet in the first dataset that represents the data per country and the water sheet in the second dataset that represents the data per region were merged into one sheet called water that represents both the data per country and per region, and the same thing was done for the other four sheets. After that, we decided to merge the two datasets that represent some estimations on water, sanitation, and hygiene in schools by country and by region into a new merged dataset called “merged\_school\_data.xlsx”, which was done by simply adding the one sheet of the first dataset and the three sheets of the second dataset, and load them into a new spreadsheet which consists of four sheets. After that, we managed to upload these three new merged datasets into google drive, and then, we decided to merge these three merged datasets into one mega dataset called “merged\_data.xlsx”, which was done by simply adding the three sheets of the first merged dataset and the five sheets of the second dataset, and the four sheets of the third dataset, and load them into a new spreadsheet which consists of twelve sheets, and we managed to upload it to google drive as well, which could be found in this link: <https://docs.google.com/spreadsheets/d/12F2YGkzgFDjN26f5bedBKWhcQnqn5o41/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>.

## 6.2. Merging Datasets

Considering the merging stage, we have explained how it is done thoroughly in the previous section. However, we’ve decided to draw a simple diagram to show how it is mainly done [36].

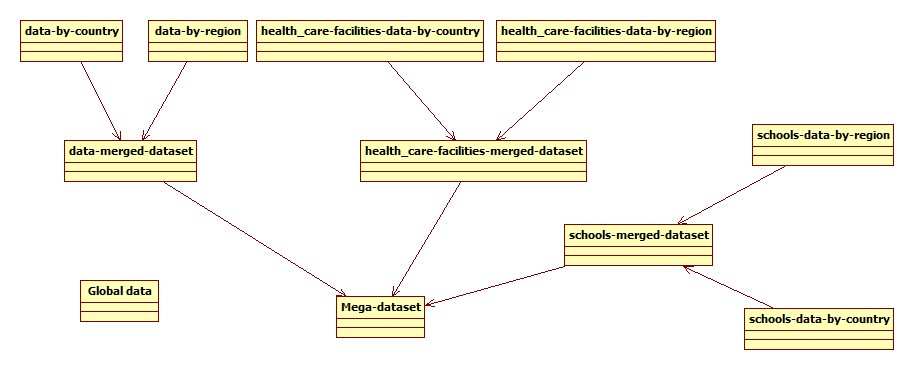


Figure 14: Merging Datasets

As we can see in Figure 14: Merging Datasets, this figure shows how our datasets are merged in google collab. Speaking of which, the global dataset is our main dataset and it contains everything related to the clean water in the Arab countries, and therefore, there is no need to merge it with other datasets. However, we decided to merge the data-by-country dataset and the data-by-region dataset into a dataset called “data-merged-dataset”, and the healthcare-facilities-data-by-country dataset and the healthcare-facilities-data-by-region dataset into a dataset called “healthcare-facilities-merged-dataset”, and the schools-data-by-country dataset and schools-data-by-region dataset into a dataset called “schools-merged-dataset”. After that, we decided to merge the data-merged dataset, the healthcare-facilities-merged dataset, and the schools-merged dataset into a mega dataset called “Mega-dataset”.

## 6.3. ETL Pipeline Architecture

In this section, we are going to represent our main ETL pipeline architecture which I created using draw.io. Speaking of which, the ETL pipeline architecture has to represent the three main steps of every ETL architecture, which consists of the extract part, the transform part, and the load part [37].

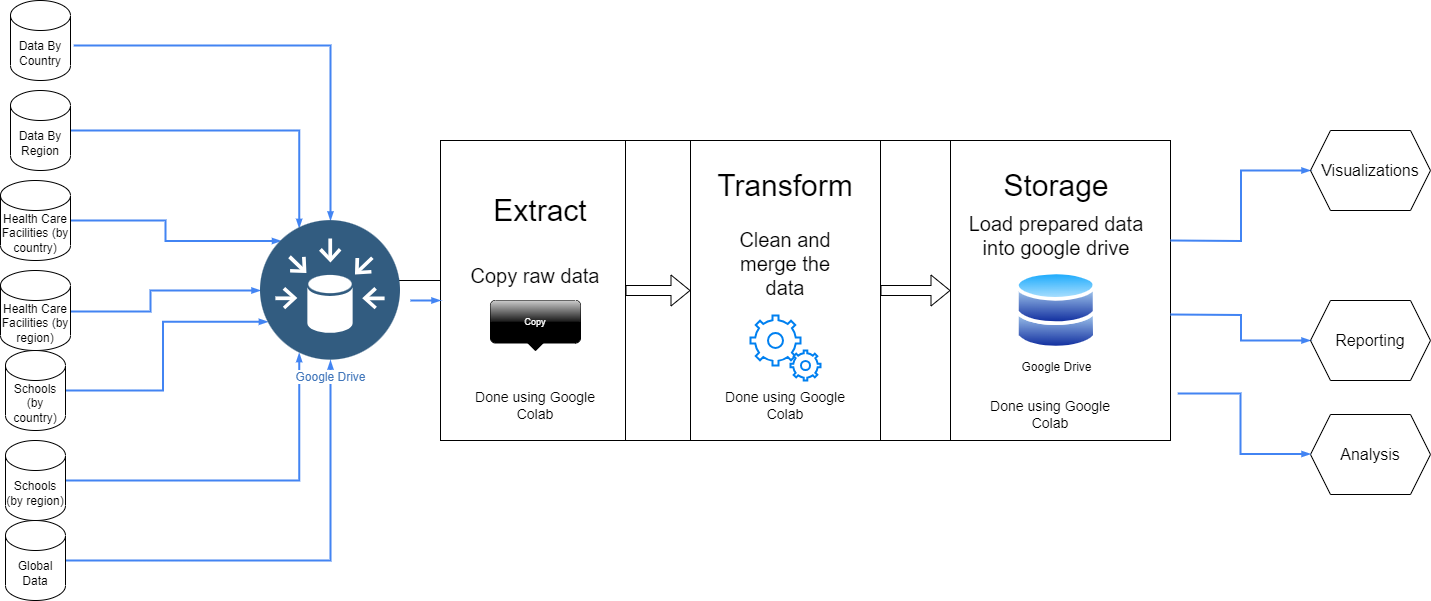


Figure 15: ETL pipeline architecture

As we can see in Figure 15: ETL pipeline architecture, firstly, we gathered our seven datasets and we put them in Google Drive. After that, using Google Collab, we extracted these datasets from Google Drive using pandas and other frameworks. Then, we transformed our datasets by trying to figure out if there is any dirty data and we tried to handle the dirty data, and then, we merged our datasets as mentioned in the previous sections. Lastly, after doing all of these steps using python and Google collab, we loaded and stored our two big merged datasets into Google Drive. Besides that, after we loaded our merged datasets, we can use them to do further visualizations, reporting, and analyses.

## 6.4. Conclusion

As we can see, we can conclude that designing the ETL pipeline architecture has a huge role of understanding how our ETL is going to work and how it can benefit us in the future. Furthermore, it helps us to manage our project in a profound way so that we can see the broad picture of our project.

# VII. Prediction Approaches of Water Scarcity Risk

## 7.1. Comparative Analysis of Prediction Models for This Risk

In this part of our report, we are going to outline some prediction models that can be considered for assessing water-scarcity risk. Each of these models has its own strengths and may be suitable for different contexts to conduct our water scarcity risk assessment [38].

### 7.1.1. Statistical Regression models

* **Description:** Statistical regression models, including linear regression, logistic regression, and other variants, use historical data to identify relationships between various factors and water scarcity. These models are based on statistical patterns and correlations, providing a quantitative basis for prediction.
* **Strengths:** Simplicity, interpretability, and ease of implementation make statistical regression models valuable for initial insights into water-scarcity risk.
* **Limitations:** These models assume a linear relationship between variables, which may oversimplify the complex dynamics of water scarcity. They may also struggle to capture non-linear patterns.

### 7.1.2. Time Series Models (e.g., LSTM)

* **Description:** Time series models, such as Long Short-Term Memory (LSTM) networks, are particularly suited for capturing temporal patterns in data. LSTM, a type of recurrent neural network (RNN), is adept at handling sequential information, making it valuable for predicting water scarcity trends over time.
* **Strengths:** LSTM models excel in capturing long-term dependencies in time series data, making them suitable for predicting water scarcity trends influenced by historical patterns.
* **Limitations:** Complex to train, may require substantial computational resources, and may be sensitive to the quality and quantity of available data.

### 7.1.3. Machine Learning Ensemble Models

* **Description:** Ensemble models, such as Random Forests or Gradient Boosting Machines (GBM), combine multiple individual models to improve overall predictive performance. These models can handle complex interactions between variables and provide robust predictions.
* **Strengths:** Ensemble models leverage the strengths of multiple algorithms, enhancing prediction accuracy. They are also relatively resilient to overfitting and noise in the data.
* **Limitations:** The complexity of ensemble models may make them less interpretable compared to simpler models. They may also require careful tuning of hyperparameters.

### 7.1.4. Hybrid Models

* **Description:** Hybrid models combine different types of models or methodologies to capitalize on their respective strengths. For instance, integrating statistical models with machine learning algorithms or combining physical models with data-driven approaches.
* **Strengths:** Hybrid models attempt to mitigate the weaknesses of individual models by leveraging the advantages of multiple approaches.
* **Limitations:** Designing effective hybrid models requires a deep understanding of the strengths and limitations of each component, and their performance may be contingent on the quality of data and the appropriateness of the combination.

### 7.1.5. Conclusion

Comparing these prediction models involved assessing their performance metrics, computational requirements, interpretability, and suitability for the specific characteristics of water scarcity in the Arab world. Each model brings its own set of considerations, and a thorough analysis was essential in order to select the most appropriate approach for my research.

## 7.2. Comparative Table of Prediction Models for This Risk

|  |  |  |  |
| --- | --- | --- | --- |
| **Models** | **Description** | **Strengths** | **Limitations** |
| **Statistical Regression model** | Use data to identify relationships between factors | * simplicity * Interpretability * Ease of implementation | * Oversimplifying complex dynamics * Struggling with non-linear patterns |
| **Time Series Models (LSTM)** | Suited for capturing temporal patterns in data | * Capturing long-term dependencies | * Complex to train * Sensitive to quality and quantity of data |
| **Machine Learning Ensemble Models** | Combine individual models to improve predictive performance | * Enhancing prediction accuracy * Resilient to overfitting | * Very complex * May require tuning of hyperparameters. |
| **Hybrid Models (chosen model)** | Combine different types of models | * Mitigating the weakness of individual models | * Contingent performance on the quality of data |

Table 2: Comparative Table of Prediction Models for This Risk

# VIII. Development of Author’s Approach for Water Scarcity Risk Prediction

## 8.1. Math Task Formulation -Introduction

The mathematical task for predicting water-scarcity risk involves defining the problem with precision and clarity. The primary objective is to develop a mathematical model that captures the intricate dynamics of water-scarcity risk in the Arab world. Variables contributing to the prediction model encompass climatic indicators, population demographics, urbanization rates, agricultural water use, and other relevant factors. Formulating the problem entails expressing the relationship between these variables in a mathematical equation that encapsulates the complexities of water-scarcity dynamics. The mathematical model must consider both spatial and temporal dimensions, accounting for the region's unique climatic patterns and potential changes over time. Constraints, such as data availability and computational resources, are also factored into the formulation, ensuring a realistic and feasible mathematical representation of the water-scarcity risk prediction problem. The task at hand requires a nuanced approach, striking a balance between mathematical rigor and practical applicability to effectively address the intricacies of water resource management in the Arab world.

The mathematical task for predicting water-scarcity risk involves establishing a model that quantifies the dynamics inherent in this multifaceted challenge. Let's denote the water-scarcity risk as R.  
 Speaking of which, water-scarcity risk (R) refers to the likelihood of insufficient water availability, encompassing physical scarcity and the potential consequences of imbalances between water supply and demand. It considers factors such as climatic variations, population growth, urbanization trends, unsustainable water management practices, water pollution, and deteriorating water quality. R serves as a comprehensive measure of the vulnerability of a region or community to water scarcity challenges.

Besides that, let’s consider relevant variables:

* **A:** Available water resources.
* **D:** Water demand.
* **Q:** Water quality.
* **P:** Population.
* **U:** Degree of urbanization.
* **T:** Time.

The water-scarcity risk (R) can be formulated as a function of these variables:

R = f (A, D, Q, P, U, T)

Here, f represents the complex relationship between available water resources, demand, quality, population, urbanization, and time. The equation captures the essence of water-scarcity risk, acknowledging that each variable contributes uniquely to the overall risk. The specific nature of the function.

Aside from that, in this function, it is essential to assign weights to factors, which can be done using the following mathematical representation:

R = w1 \* A + w2 \* D + w3 \* Q + w4 \* P + w5 \* U + w6 \* T

Where:

* w1, w2, w3, w4, w5, and w6 are weight coefficients representing the importance of each variable, which can be determined through expert evaluation or stakeholder consultation.
* A, D, Q, P, U, and T are the respective values of available water resources, water demand, water quality, population, degree of urbanization, and time.

## 8.2. Math Task Formulation - Determining Weight Values

* **Available Water Resources (A):** This variable represents the quantity of water available for use in a given region. In regions prone to water scarcity, such as many parts of the Arab world the availability of water resources is often a critical factor. Let's assign a weight of 0.25 to this variable, indicating its moderate importance in determining water-scarcity risk.
* **Water Demand (D):** The demand for water, driven by factors such as population size, economic activities, and agricultural practices, directly affects water scarcity. In water-stressed regions, high water demand exacerbates scarcity issues. Let's assign a weight of 0.3 to this variable, indicating its significant impact on water risk.
* **Water Quality (Q):** Water quality is crucial for human health, ecosystem integrity, and economic activities such as agriculture and industry. Poor water quality can exacerbate scarcity by limiting available resources due to contamination. Let's assign a weight of 0.15 to this variable, reflecting its importance but slightly lower than other factors.
* **Population (P):** Population size influences water demand and resource management practices. In densely populated regions like urban areas in the Middle East, population growth can strain water resources. Let's assign a weight of 0.2 to this variable, indicating its substantial impact on water scarcity.
* **Degree of Urbanization (U):** Urbanization affects water demand patterns, infrastructure requirements, and land-use practices, contributing to water scarcity challenges. In rapidly urbanizing regions, such as many cities in the Arab world, the degree of urbanization is a significant factor. Let's assign a weight of 0.1 to this variable, acknowledging its importance but relatively lower compared to other factors.
* **Time (T):** Time represents temporal dynamics such as climate change, seasonal variations, and long-term trends in water availability and demand. Let's assign a weight of 0.1 to this variable, recognizing its role in shaping water-scarcity risk over time.

These weights sum up to 1, reflecting the relative importance of each variable in the overall assessment of water-scarcity risk. However, it's essential to validate these weight values through expert consultation, empirical analysis, and sensitivity testing to ensure they accurately capture the nuances of water scarcity in the specific context of the Middle East and the Arab world.

🡺 R = 0.25\* A + 0.3 \* D + 0.15 \* Q + 0.2 \* P + 0.1 \* U + 0.1 \* T

## 8.3. Proposed Approach

In selecting an approach for predicting water-scarcity risk in the Arab world, we prioritize a solution that balances accuracy and ease of implementation. Our proposed approach leans towards a hybrid model that integrates statistical and machine learning techniques, leveraging their respective strengths. We opt for this approach due to its relative simplicity in implementation, making it accessible to a broader audience of stakeholders involved in water resource management. The hybrid model will blend statistical regression models, such as linear regression for its interpretability, with machine learning ensemble techniques like Random Forests for their ability to capture non-linear relationships and interactions. This combination allows for a robust prediction model while facilitating straightforward implementation and understanding [39]. By emphasizing ease of implementation, we aim to empower decision-makers, including policymakers and water resource managers, to readily utilize and interpret the insights provided by our predictive model. This pragmatic approach aligns with the urgency of addressing water-scarcity risks in the Arab world, ensuring that the outcomes of our research are not only accurate but also actionable for sustainable water resource management.

## 8.4. Experiments With Real Data

In the pursuit of refining and validating our proposed hybrid model for predicting water-scarcity risk, the next crucial phase involves conducting experiments with real data. This entails collecting historical data from reliable sources encompassing a spectrum of variables, including but not limited to water availability, demand, quality, population demographics, urbanization trends, and temporal patterns. Real data from the past serves as the foundation for training and testing the model, allowing it to discern patterns and relationships that contribute to water-scarcity risk. The data collection process involves meticulous attention to quality, consistency, and relevance to ensure the model's accuracy and applicability to the specific dynamics of the Arab world. Once the dataset is curated, the hybrid model will undergo training, fine-tuning, and validation to optimize its predictive capabilities. Through this empirical approach, we aim to derive actionable insights into the future trajectory of water-scarcity risk in the region, contributing to informed decision-making for sustainable water resource management.

## 8.5. Functional Requirements

In outlining the functional requirements for the development of our water-scarcity risk prediction model using Google Colab and Python, we identify key components and capabilities necessary for the successful implementation of the proposed approach [40].

**-Data Integration and Preprocessing:**

* **Requirement:** The system must enable seamless integration of data from Google Drive, incorporating historical datasets that cover essential variables for water-scarcity risk assessment in the Arab world.
* **Rationale:** This capability is fundamental for the model's training and validation processes, ensuring that real data is effectively utilized.

**-Framework Compatibility:**

* **Requirement:** The system must support the integration of major Python-based frameworks for AI and data science, such as TensorFlow and Scikit-learn.
* **Rationale:** Compatibility with these frameworks provides access to powerful tools and algorithms necessary for building, training, and evaluating the predictive model.

**-Machine Learning Model Implementation:**

* **Requirement:** The system should facilitate the implementation of a hybrid model, incorporating statistical regression models and machine learning ensemble techniques.
* **Rationale:** This capability allows for the integration of diverse approaches, balancing accuracy and ease of implementation in the prediction model.

**-Google Colab Integration:**

* **Requirement:** The system must seamlessly integrate with Google Colab for cloud-based computing resources, allowing for efficient model training and analysis.
* **Rationale:** Utilizing Google Colab ensures access to scalable computational power, crucial for handling large datasets and training complex models.

**-Model Training and Evaluation Tools:**

* **Requirement:** The system should provide tools for training the model with historical data and evaluating its predictive performance.
* **Rationale:** These tools are essential for iteratively refining the model and assessing its accuracy in predicting water-scarcity risk.

**-Visualization Capabilities:**

* **Requirement:** The system must offer visualization tools for interpreting and communicating the model's outcomes effectively.
* **Rationale:** Visualizations aid in conveying insights to stakeholders, facilitating informed decision-making in water resource management.

**-Data Security and Privacy:**

* **Requirement:** The system should implement measures to ensure the security and privacy of sensitive data, especially when dealing with real datasets.
* **Rationale:** Protecting data integrity and privacy is paramount, especially when dealing with information that may be sensitive or subject to regulations.

By delineating these functional requirements, we aim to create a robust and versatile system that empowers the implementation of our water-scarcity risk prediction model within the Google Colab and Python environment.

# IX. Development of Water-Scarcity Risk Evaluation System

## 9.1. Proposed System’s Architecture

The proposed system's architecture for the Water-Scarcity Risk Evaluation System is designed to seamlessly integrate data, analysis tools, and machine learning models within the Google Colab and Python environment [41]. The architecture follows a modular and scalable approach to accommodate the complexity of water-scarcity risk assessment in the Arab world. Key components of the proposed system's architecture include:

**1-Data Integration Layer:**

* **Objective:** This layer focuses on integrating datasets from Google Drive, ensuring the inclusion of historical data spanning variables crucial for water-scarcity risk assessment.
* **Implementation:** Utilizing Python libraries, such as Pandas and Google Colab-specific tools, to efficiently import, preprocess, and merge datasets for comprehensive analysis.

**2-Framework Integration Layer:**

* **Objective:** Facilitate seamless integration with major Python-based frameworks for AI and data science, specifically TensorFlow and Scikit-learn.
* **Implementation:** Leveraging the capabilities of Google Colab to effortlessly incorporate these frameworks, allowing for the implementation and evaluation of machine learning models.

**3-Machine Learning Model Layer:**

* **Objective:** Implement a hybrid model combining statistical regression models and machine learning ensemble techniques for robust water-scarcity risk prediction.
* **Implementation:** Utilizing TensorFlow and Scikit-learn functionalities within Google Colab to build, train, and evaluate the hybrid model with real data.

**4-Google Colab Cloud Computing Layer:**

* **Objective:** Harness the power of cloud-based computing resources provided by Google Colab for efficient model training and analysis.
* **Implementation:** Designing the system to seamlessly interact with Google Colab's cloud infrastructure, ensuring scalability and resource optimization.

**5-Visualization Layer:**

* **Objective:** Enable effective communication of model outcomes through visualization tools.
* **Implementation:** Leveraging Python libraries such as Matplotlib and Seaborn to create visualizations.

The proposed architecture emphasizes flexibility, scalability, and accessibility, aligning with the goal of providing a user-friendly platform for water-scarcity risk assessment. Through this modular structure, stakeholders, including policymakers and water resource managers, can navigate and leverage the system efficiently, fostering informed decision-making for sustainable water resource management in the Arab world.

However, in order to provide a visual representation of the system's components and interactions, we have included a UML component diagram shown below, which was created using “Whitestar UML” software. This diagram complements the earlier textual description, offering a more intuitive understanding of the system's structure.

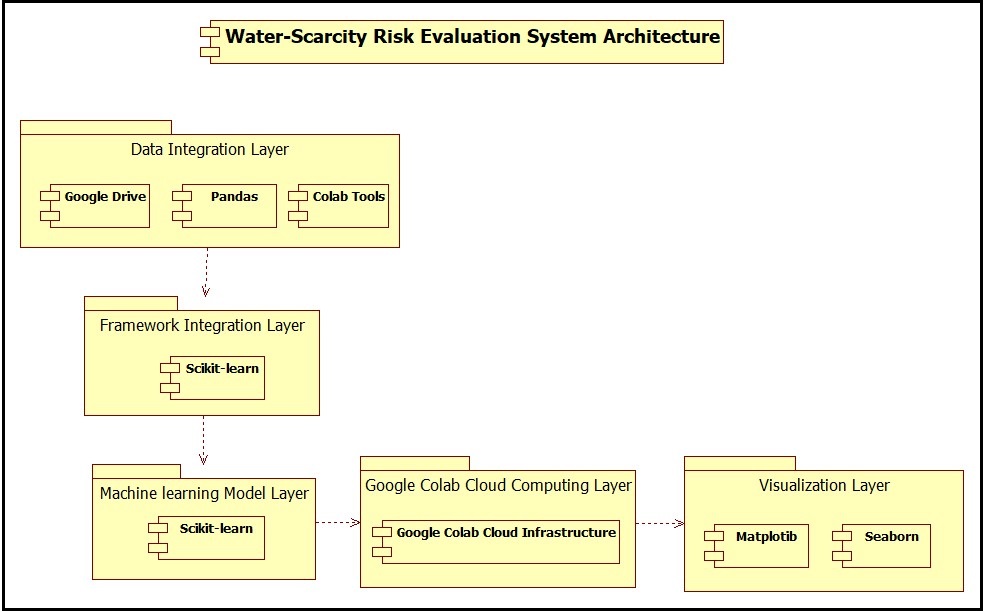


Figure 16: UML Component Diagram For our Proposed System's Architecture

## 9.2. Data-Logical Model

In the following section, we present the datalogical model for our Water-Scarcity Risk Evaluation System. A datalogical model serves as a fundamental representation of how data entities are structured within the system and how they interconnect. This model is crucial for understanding the underlying data architecture, relationships, and the logic behind our risk evaluation approach. The system's datalogical model encapsulates the essential components, their attributes, and the associations that form the backbone of our comprehensive approach to assessing water scarcity risk in the Arab world. The provided code snippet serves as a conceptual representation, outlining the relationships and attributes that define our data entities.

**Here is the code snippet:**

### 9.2.1. Entities

1. **Country:**

- Attributes: CountryID (Primary Key), Name, Region.

1. **WaterAvailability:**

- Attributes: WaterAvailabilityID (Primary Key), CountryID (Foreign Key), Year, NationalAvailability, RuralAvailability, UrbanAvailability.

1. **WaterScarcityRiskIndex:**

- Attributes: RiskIndexID (Primary Key), CountryID (Foreign Key), WaterAvailibilityID (Foreign Key), Year, RiskIndexValue.

### 9.2.2. Relationships

* One-to-Many relationship between **Country** and **WaterAvailability:**

One country can have multiple records in the WaterAvailability table for different years.

* One-to-One relationship between **WaterAvailability** and **WaterScarcityRiskIndex:**

One WaterAvailability record corresponds to one WaterScarcityRiskIndex record for the same country and year.

### 9.2.3. ER Diagram

After representing the data-logical model using a code snippet, we will be representing it using an entity-relationship (ER) model, which is a fundamental representation that illustrates how data entities are organized and interconnected within our Water-Scarcity Risk Evaluation System [42]. Speaking of which, this ER diagram encapsulates essential components such as countries, water availability records, and the calculated water-scarcity risk index. Each entity's attributes and the relationships between them form the backbone of our comprehensive approach to assessing water scarcity risk in the Arab world. Therefore, we decided to draw the following ER diagram for our system:

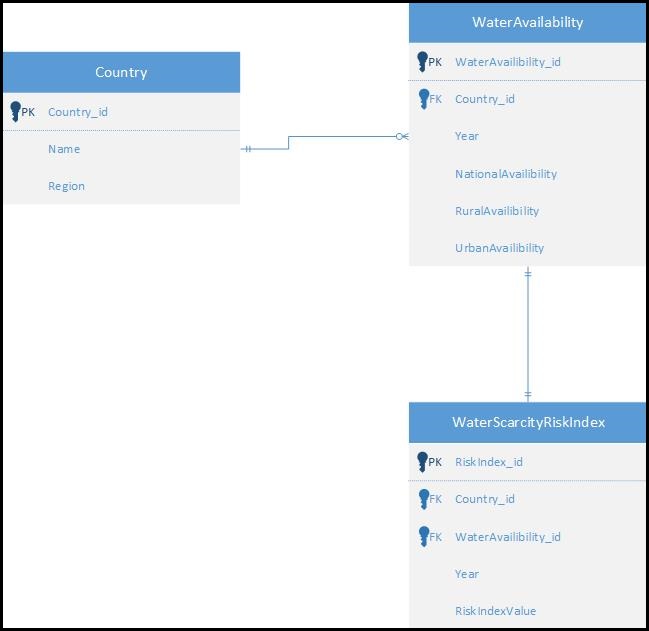


Figure 17: Data-logical model (ER diagram)

## 9.3. System’s Requirements

To use the Water-Scarcity Risk Evaluation System, users need a computer or laptop running Windows 10 operating system with a minimum of 2GB RAM. The application is optimized for 64-bit systems to ensure optimal performance. Additionally, users are required to have access to the necessary datasets, which should be imported into Google Drive. To interact with the provided code and execute the analysis, users are expected to have a Google Colab account. Furthermore, Python must be installed on the user's machine, allowing seamless execution of the code and analysis within the Google Colab environment [43].

### 9.3.1. Minimum Hardware Requirements:

* Computer or laptop with Windows 10
* 2GB of RAM
* 64-bit architecture

### 9.3.2. Software Requirements:

* Access to necessary datasets imported into Google Drive
* Google Colab account for code execution and analysis
* Python installed on the user's machine

# X. Water-Scarcity Risk Evaluation System MVP

## 10.1. Selected Tools for Development

The development of the Water-Scarcity Risk Evaluation System MVP involved a thoughtful selection of tools to ensure efficiency, flexibility, and scalability [44]. The following tools were chosen for various aspects of the development process:

1. **Programming Language: Python**

* Python was selected for its versatility, extensive libraries, and data analysis capabilities. It serves as the primary language for data manipulation, statistical modeling, and machine learning.

1. **Data Analysis and Visualization: Pandas and Seaborn**

* Pandas, a powerful data manipulation library, was used for cleaning and processing datasets. Seaborn was employed for creating insightful visualizations to communicate complex data trends effectively.

1. **Machine Learning: Scikit-learn (sklearn)**

* Scikit-learn was utilized for implementing linear regression models, facilitating predictive analysis of water availability in the Arab world.

1. **Notebook Environment: Google Colab**

* Google Colab, a cloud-based Jupyter notebook environment, was chosen for its collaborative features, access to GPU resources, and seamless integration with Google Drive for dataset management.

1. **Linear Regression: Scikit-learn (sklearn)**

* Scikit-learn's implementation of linear regression models enabled the prediction of future water availability based on historical data.

These selected tools collectively contribute to the development and functionality of the Water-Scarcity Risk Evaluation System MVP, offering a robust foundation for ongoing enhancements and iterations.

## 10.2. Development Results

Speaking about the water risk assessment in the Arab world, we successfully managed to extract our data from a trusted resource and imported them in google drive. After that, using Google Colab, we managed to create a file called “Water Scarcity Risk Assessment in The Arab World.ipynb” which could be found in this link: <https://colab.research.google.com/drive/1cjWLphO8BpU0OVFxlQogq7xN9WCYCL-U?usp=sharing>. Speaking of which, we firstly imported our required libraries, and then, with the help of google drive, we were able of importing our six datasets successfully. After that, we decided to split our work by importing each sheet of these datasets individually. After that, we decided to do some functions in order to filter the countries in these datasets to return only the data that is related to the Arab countries. Then, we decided to use matplotib and seaborn frameworks in order to display a chart about the current water availability in each Arab country. After that, using Scikit-learn (also known as sklearn), we managed to assess the water scarcity risk in the Arab countries in the future. Speaking of which, we based our research on representing charts according to each country’s national level, along with representing charts according to each country’s rural areas and urban areas, which would have a significant role of helping the main stakeholders to make their decisions regarding this issue in a more accurate way by considering all of these factors. However, speaking about the charts that we decided to represent, here is a list of all of them:

* Future Projection of Water Availability in The Arab World (National)
* Future Projection of Water Availability in The Arab World (Rural)
* Future Projection of Water Availability in The Arab World (Urban)
* Future Projection of Sanitation Availability in The Arab World (National)
* Future Projection of Sanitation Availability in The Arab World (Rural)
* Future Projection of Sanitation Availability in The Arab World (Urban)
* Future Projection of Basic Water Services in the Arab World (National)
* Future Projection of Basic Water Services in the Arab World (Rural)
* Future Projection of Basic Water Services in the Arab World (Urban)
* Future Projection of Basic Sanitation Services in the Arab World (National)
* Future Projection of Basic Sanitation Services in the Arab World (Rural)
* Future Projection of Basic Sanitation Services in the Arab World (Urban)
* Future Projection of Basic Hygiene Services in the Arab World (National)
* Future Projection of Basic Hygiene Services in the Arab World (Rural)
* Future Projection of Basic Hygiene Services in the Arab World (Urban)
* Future Projection of Basic Waste Management Services in the Arab World (National)
* Future Projection of Basic Environmental Cleaning Services in the Arab World (National)
* Future Projection of Basic Water Services in Schools in the Arab World (National)
* Future Projection of Basic Water Services in Schools in the Arab World (Rural)
* Future Projection of Basic Water Services in Schools in the Arab World (Urban)

After that, we decided to display some charts that are related to the availability of clean water by country, the unimproved water index by country, the availability of limited water by country, the availability of surface water by country, and the annual rate of change by country. Speaking of which, we decided to represent each of these charts in a different color in order to differentiate them. Moreover, we have successfully presented these charts by defining two weights that represent the rural areas and the urban areas respectively, and we combined them and drew our charts accordingly.

Now let me show you some of these charts. However, in order to see all of them, kindly click on the Google Colab link provided above:

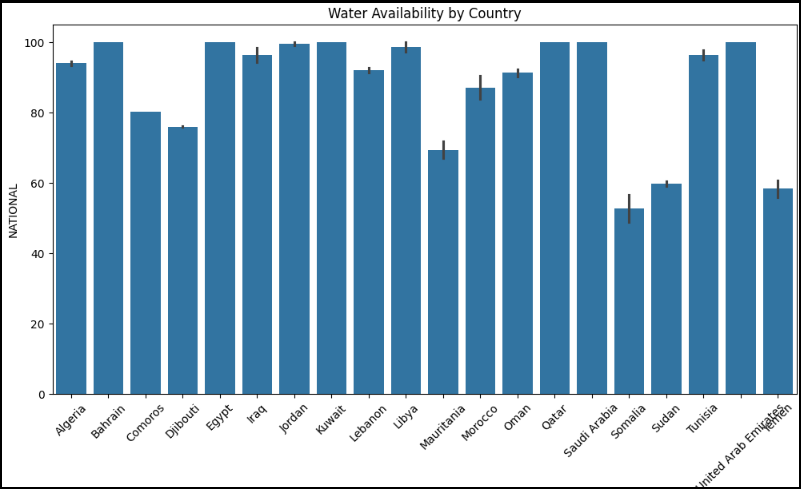


Figure 18: Current water availability by country

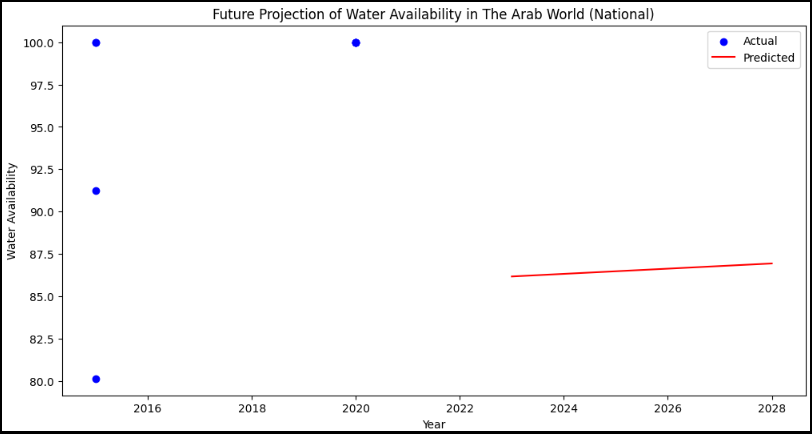


Figure 19: Future Projection of water availability in the Arab World (National)

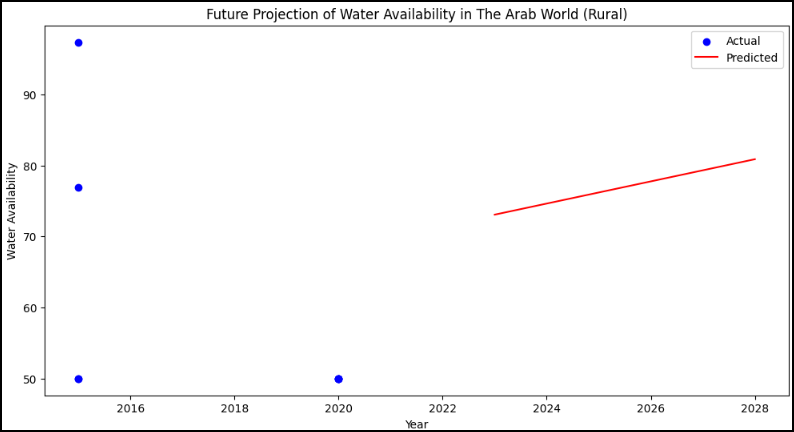


Figure 20: Future Projection of Water Availability in The Arab World (Rural)

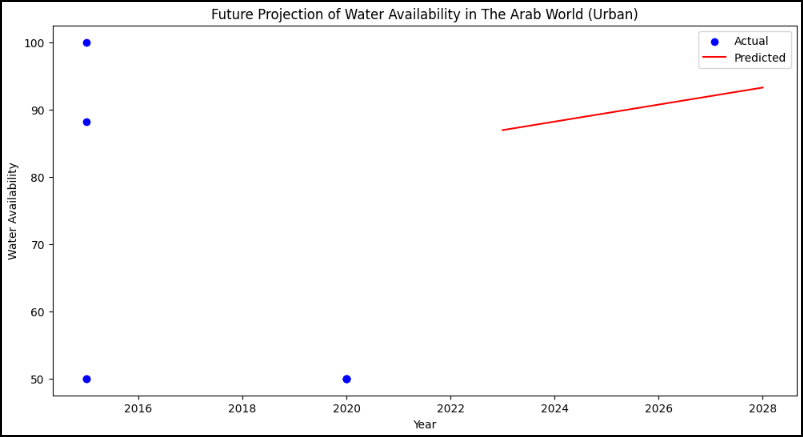


Figure 21: Future Projection of Water Availability in the Arab World (Urban)

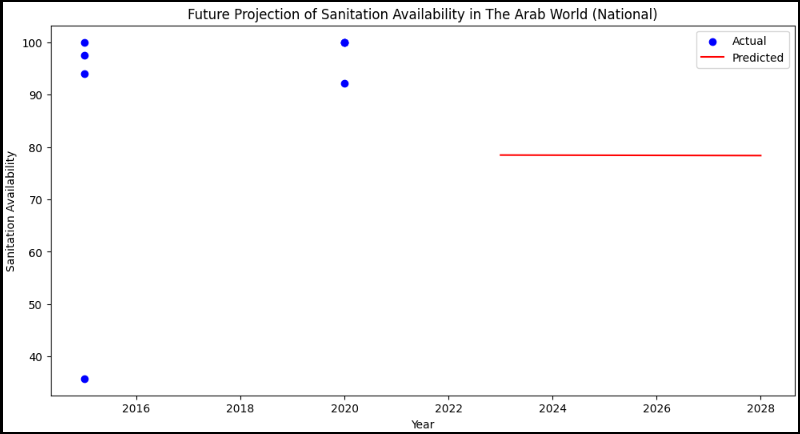


Figure 22: Future Projection of Sanitation Availability in the Arab World (National)

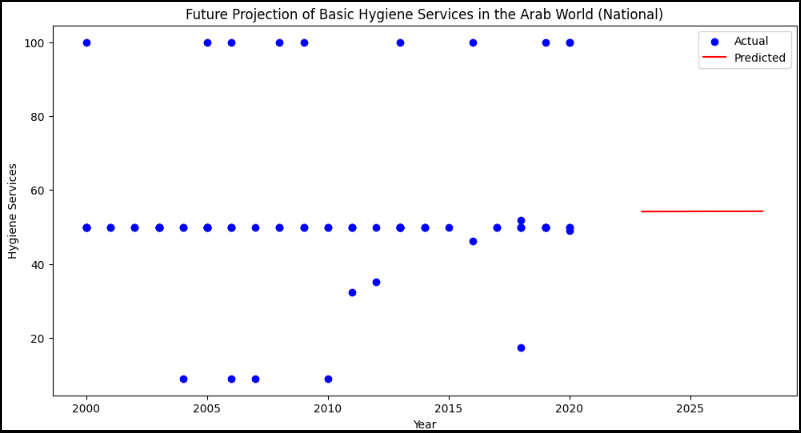


Figure 23: Future Projection of Basic Hygiene Services in the Arab World (National)

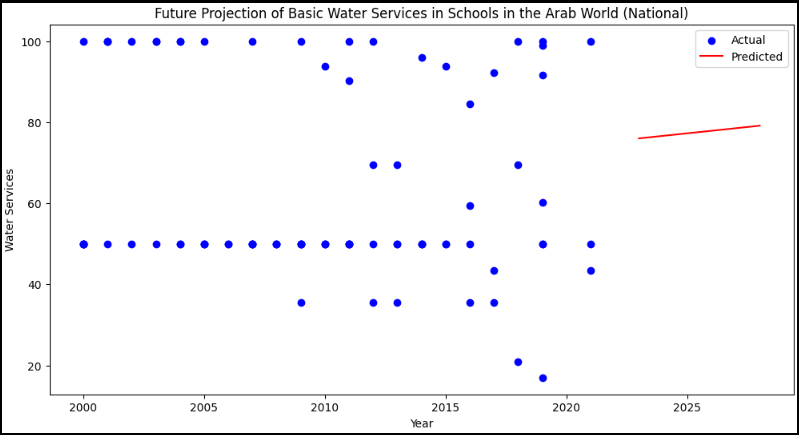


Figure 24: Future Projection of Basic Water Services in Schools in the Arab World (National)

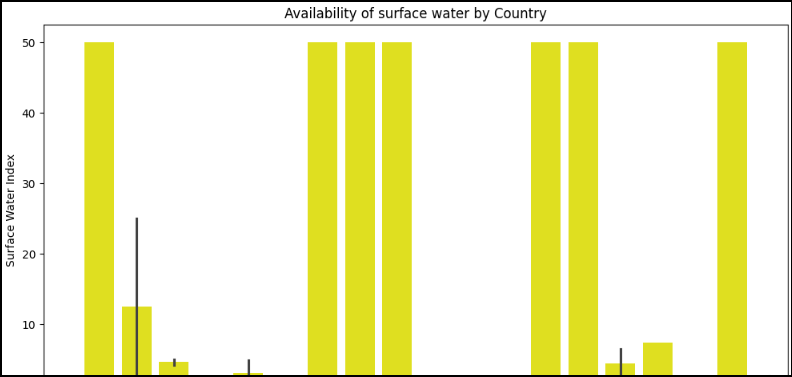


Figure 25: Availability of Surface Water by Country

## 10.3. Conclusion

After drawing these charts using Google Colab, all the main stakeholders can benefit from them by analyzing them and making decisions that could benefit their countries in order to limit the water scarcity risk in the future. On top of that, drawing these charts has a huge role of helping us to manage our project in a profound way so that we can see the broad picture of our project.

# XI. References and Appendix

## 11.1. References

Regarding the references, we put some numbers surrounded by square brackets in our report, in order to allow the teacher to know that the provided information was brought using Google. Thus, we decided to provide all the references to these pieces of information here using “Harvard” style.

1. World Wildlife Fund (2022). *Water Scarcity | Threats | WWF*. [online] World Wildlife Fund. Available at: https://www.worldwildlife.org/threats/water-scarcity [Accessed 10 Apr. 2024].
2. UNICEF (2020). *Water scarcity*. [online] www.unicef.org. Available at: https://www.unicef.org/wash/water-scarcity#:~:text=Half%20of%20the%20world [Accessed 10 Apr. 2024].
3. Schulte, P. (2014). *Defining Water Scarcity, Water Stress, and Water Risk*. [online] Pacific Institute. Available at: https://pacinst.org/water-definitions/ [Accessed 10 Apr. 2024].
4. Maxwell, D. (2015). *The Risks and Opportunities Posed by Water Can’t Be Ignored—by Business, Industries, or Nations | IFAC*. [online] www.ifac.org. Available at: https://www.ifac.org/knowledge-gateway/contributing-global-economy/discussion/risks-and-opportunities-posed-water-cant-be-ignored-business-industries-or-nations#:~:text=Water%20risk%20factors%20include%20water [Accessed 10 Apr. 2024].
5. Impulse, S. (2022). *Solutions to water scarcity: how to prevent water shortages?* [online] solarimpulse.com. Available at: https://solarimpulse.com/water-scarcity-solutions# [Accessed 10 Apr. 2024].
6. Enzo (2023). *The Definitive Guide to Building a Predictive Model in Python*. [online] Graph Database & Analytics. Available at: https://neo4j.com/blog/build-predictive-model-python/ [Accessed 10 Apr. 2024].
7. GIS Geography. (2020). *10 Python Libraries for GIS and Mapping*. [online] Available at: https://gisgeography.com/python-libraries-gis-mapping/ [Accessed 10 Apr. 2024].
8. Wrike. (2023). *What is Iterative and Incremental Development?* [online] Available at: https://www.wrike.com/project-management-guide/faq/what-is-iterative-incremental-development/ [Accessed 10 Apr. 2024].
9. Awati, R. (2024). *What is iterative development? - Definition from WhatIs.com*. [online] SearchSoftwareQuality. Available at: https://www.techtarget.com/searchsoftwarequality/definition/iterative-development#:~:text=Iterative%20development%20is%20a%20way [Accessed 10 Apr. 2024].
10. Pedamkar, P. (2019). *What is ETL? | How it Work | Needs and Advantages | Scope & Career*. [online] EDUCBA. Available at: https://www.educba.com/what-is-etl/ [Accessed 10 Apr. 2023].
11. ProjectManager. (2024). *Task Lists*. [online] Available at: https://www.projectmanager.com/guides/task-lists [Accessed 10 Apr. 2024].
12. Pubrica Academy (2019). *Why is it important to do a literature review in research?* [online] Pubrica. Available at: https://pubrica.com/academy/research/why-is-it-important-to-do-a-literature-review-in-research/ [Accessed 10 Apr. 2023].
13. mWater. (2021). *mWater*. [online] Available at: https://www.mwater.co/ [Accessed 10 Apr. 2023].
14. mWater. (2021). *mWater*. [online] Available at: https://www.mwater.co/ [Accessed 10 Apr. 2023].
15. Akvo Foundation. (2023). *Better data, bigger impact*. [online] Available at: https://akvo.org/ [Accessed 10 Apr. 2023].
16. Emeline, E. (2022). *Data is the new water*. [online] datajourney.akvo.org. Available at: https://datajourney.akvo.org/blog/data-is-the-new-water [Accessed 10 Apr. 2023].
17. www.openwaterdata.com. (2023). *Open Water Data*. [online] Available at: https://www.openwaterdata.com/ [Accessed 10 Apr. 2023].
18. Apostolopoulos, I.D. and Groumpos, P.P. (2023). Fuzzy Cognitive Maps: Their Role in Explainable Artificial Intelligence. *Applied Sciences*, 13(6), p.3412. doi:https://doi.org/10.3390/app13063412.
19. Kok, K. (2023). *Fuzzy Cognitive Maps (pros and cons)*. [online] Research Gate. Available at: https://www.researchgate.net/figure/Strong-and-weak-points-of-Fuzzy-Cognitive-Mapping\_tbl2\_222390094 [Accessed 10 Apr. 2023].
20. Concepcion, L., Nápoles, G., Bello, R. and Vanhoof, K. (2020). On the Behavior of Fuzzy Grey Cognitive Maps. pp.462–476. doi:https://doi.org/10.1007/978-3-030-52705-1\_34.
21. Wang, J. and Guo, Q. (2018). Ensemble Interval-Valued Fuzzy Cognitive Maps. *IEEE Access*, 6, pp.38356–38366. doi:https://doi.org/10.1109/access.2018.2853995.
22. Aguilar, J. (2010). Dynamic Fuzzy Cognitive Maps for the Supervision of Multiagent Systems. pp.307–324. doi:https://doi.org/10.1007/978-3-642-03220-2\_13.
23. The Editors of Encyclopaedia Britannica (2019). Arab League | Britannica. In: *Encyclopædia Britannica*. [online] Available at: https://www.britannica.com/topic/Arab-League [Accessed 10 Apr. 2023].
24. www.swfwmd.state.fl.us. (2018). *Measuring Water Quality | WaterMatters.org*. [online] Available at: https://www.swfwmd.state.fl.us/residents/education/kids/measuring-water-quality#:~:text=Scientists%20use%20many%20different%20instruments [Accessed 10 Apr. 2023].
25. Florida, U.W.I., School of Geosciences, University of South (2023). *Learn More: Water Quality Index (WQI) - Sarasota County Water Atlas - Sarasota.WaterAtlas.org*. [online] sarasota.wateratlas.usf.edu. Available at: https://sarasota.wateratlas.usf.edu/library/learn-more/learnmore.aspx?toolsection=lm\_wqi [Accessed 10 Apr. 2023].
26. Centers for Disease Control and Prevention. (2019). *Assessing Access to Water & Sanitation* . [online] Available at: https://www.cdc.gov/healthywater/global/assessing.html [Accessed 10 Apr. 2023].
27. Bankole, A.O., Abraham Olamilekan James, Enovwo Odjegba, Folashade Racheal Bankole, Emmanuel, B., Fabiana Alves Fiore, Pu, J.H. and Rodrigo Moruzzi (2023). Factors affecting sanitation coverage in three income levels and potential toward achieving SDG 6.2. *Water Policy*, [online] 25(2), pp.146–176. doi:https://doi.org/10.2166/wp.2023.181.
28. CDC. (2022). *Waterborne Disease*. [online] Available at: https://www.cdc.gov/ncezid/what-we-do/2021-highlights/waterborne-disease.html [Accessed 10 Apr. 2023].
29. Avni, N., Fishbain, B. and Shamir, U. (2015). Water consumption patterns as a basis for water demand modeling. *Water Resources Research*, 51(10), pp.8165–8181. doi:https://doi.org/10.1002/2014wr016662.
30. www.heavy.ai. (2022). *What is Geospatial Analytics? Definition and Related FAQs | OmniSci*. [online] Available at: https://www.heavy.ai/technical-glossary/geospatial-analytics [Accessed 10 Apr. 2023].
31. Damkjaer, S. and Taylor, R. (2017). The measurement of water scarcity: Defining a meaningful indicator. *Ambio*, 46(5), pp.513–531. doi:https://doi.org/10.1007/s13280-017-0912-z.
32. panorama.solutions. (2016). *Community involvement in data collection*. [online] Available at: https://panorama.solutions/fr/building-block/community-involvement-data-collection [Accessed 10 Apr. 2023].
33. Ritchie, H. and Roser, M. (2019). *Clean Water Access*. [online] Our World in Data. Available at: https://ourworldindata.org/water-access [Accessed 10 Apr. 2023].
34. UNICEF (2021). *Drinking water - UNICEF DATA*. [online] UNICEF DATA. Available at: https://data.unicef.org/topic/water-and-sanitation/drinking-water/ [Accessed 10 Apr. 2023].
35. A. <https://docs.google.com/spreadsheets/d/1hk1VOr3rXemGffFs4nkyQb_sWFnJxbJY/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

B. <https://docs.google.com/spreadsheets/d/12Qy4_DNBXJ44-Obq0eOTxDU2lHcoNWcw/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

C. <https://docs.google.com/spreadsheets/d/1VipMcVb24phTtGRsynNPBLqZS3cs2cqt/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

D. <https://docs.google.com/spreadsheets/d/1fc333G_uAjPfKeIoqbz7ihkqplnxdL37/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

E. <https://docs.google.com/spreadsheets/d/16zTwGJAU2Np9Uuv9fx_Jy610xc5uu22a/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

F. <https://docs.google.com/spreadsheets/d/1CHUpF7_a8WfAI9Qhc1T3YwrIqickaWxl/edit?usp=sharing&ouid=105535614615338973847&rtpof=true&sd=true>

1. jakevdp.github.io. (2023). *Combining Datasets: Merge and Join | Python Data Science Handbook*. [online] Available at: https://jakevdp.github.io/PythonDataScienceHandbook/03.07-merge-and-join.html [Accessed 10 Apr. 2023].
2. Tobin, D. (2020). *The 6 Parts of ETL Data Pipeline Architecture*. [online] Integrate.io. Available at: https://www.integrate.io/blog/etl-architecture-building-blocks/ [Accessed 10 Apr. 2023].
3. HALTON, C. (2023). *Predictive Analytics: Definition, Model Types, and Uses*. [online] Investopedia. Available at: https://www.investopedia.com/terms/p/predictive-analytics.asp#:~:text=Predictive%20Analytics%20vs.&text=The%20most%20common%20predictive%20models [Accessed 10 Apr. 2024].
4. Das, P. and Kashem, A. (2024). Hybrid machine learning approach to prediction of the compressive and flexural strengths of UHPC and parametric analysis with shapley additive explanations. *Case Studies in Construction Materials*, [online] 20, p.e02723. doi:https://doi.org/10.1016/j.cscm.2023.e02723.
5. Gorbachenko, P. (2021). *Functional vs Non-Functional Requirements [Updated 2021]*. [online] enkonix.com. Available at: https://enkonix.com/blog/functional-requirements-vs-non-functional/ [Accessed 10 Apr. 2024].
6. www.drawio.com. (2023). *Blog - UML component diagrams show the structure of a system*. [online] Available at: https://www.drawio.com/blog/uml-component-diagrams [Accessed 10 Apr. 2024].
7. Fergusson, K. (2018). *Entity Relationship Diagrams with draw.io*. [online] draw.io. Available at: https://drawio-app.com/blog/entity-relationship-diagrams-with-draw-io/ [Accessed 10 Apr. 2024].
8. Rouse, M. (2019). *What are System Requirements? - Definition from Techopedia*. [online] Techopedia.com. Available at: https://www.techopedia.com/definition/4371/system-requirements [Accessed 10 Apr. 2024].
9. seersco.com. (2022). *Why are Software Tools Important For Your Education? Here’s the Answer - Seers | Blogs*. [online] Available at: https://seersco.com/blogs/why-are-software-tools-important-for-your-education-heres-the-answer/ [Accessed 25 Jan. 2024].

## 11.2. Appendix

This appendix provides a comprehensive overview of the methodologies, tools, and findings associated with the Water Scarcity Risk Assessment project conducted using Google Colab. The project aimed to develop a predictive model for assessing water scarcity risks in the Arab world, considering key variables such as water availability, demand, quality, population dynamics, urbanization, and time.

1. **Project Setup:**

The project was initiated by importing and preprocessing datasets related to water, sanitation, hygiene, and associated factors from authoritative sources. The primary datasets included information at both country and regional levels, covering various aspects of water-related indicators.

1. **Data Filtering:**

To focus on the Arab world, data filtering was implemented to extract information pertaining to Arab countries. Lists of Arab countries and our region of interest were defined, and functions were created to filter datasets accordingly. This step ensured that subsequent analyses were region-specific.

1. **Predictive Modeling:**

The core of the project involved predictive modeling to assess water scarcity risks. Linear regression models were initially employed to project future water availability based on historical data. Additionally, the inclusion of machine learning ensemble models, statistical regression models, and potential exploration of Long Short-Term Memory (LSTM) networks were considered for future iterations.

1. **Future Projections:**

The predictive models were utilized to project future water availability, catering to different scenarios such as national, rural, and urban perspectives. Visualization tools were employed to present actual and predicted data, providing a comprehensive view of water scarcity risks over time.

1. **Composite Risk Index:**

To synthesize various indicators into a comprehensive measure, a composite risk index was developed. This involved assigning weights to different indicators based on their significance and combining them into a unified risk assessment metric.

1. **System Requirements:**

The system requirements for replicating and extending this project were outlined. These included hardware specifications (Windows 10, 2GB RAM, 64-bit architecture) and software prerequisites (installation of Python, Google Colab account, and access to relevant datasets).

1. **Recommendations for Future Work:**

In the final section of the project, we want to make some considerations for our future work. Speaking of which, the intention is to consolidate and refine the methodologies employed in different segments of the project, particularly focusing on enhancing predictive models and incorporating advanced techniques for a more accurate assessment of water scarcity risks.

1. **Conclusion:**

This appendix encapsulates the journey of water scarcity risk assessment in the Arab world, showcasing the methodologies, analyses, and projections conducted using Google Colab. As the project evolves, there is a clear trajectory towards refining predictions, adopting advanced models, and consolidating the diverse aspects of water scarcity assessment into a cohesive framework. The future iterations aim to contribute significantly to the understanding and management of water resources in the Arab world.

**Words Count: 19266**