iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Leveraging Data Analytics in the Assessment of Water Quality Parameters on Salmon Mortality in Aquaculture: A Case Study from a Fish Farm in Norway

Ana Rita Duarte Pires

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL - University Institute of Lisbon

Co-supervisor:
PhD Øystein Klakegg, Assistant Professor,
Molde University College

June, 2023

iscte

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Leveraging Data Analytics in the Assessment of Water Quality Parameters on Salmon Mortality in Aquaculture: A Case Study from a Fish Farm in Norway

Ana Rita Duarte Pires

Master's in Integrated Business Intelligence Systems

Supervisor:
PhD João Carlos Amaro Ferreira, Assistant Professor with Habilitation,
ISCTE-IUL - University Institute of Lisbon

Co-supervisor:
PhD Øystein Klakegg, Assistant Professor,
Molde University College

June, 2023

ACKNOWLEDGMENTS

Completing this thesis closes a stage in my life of hard work but of immense personal and educational growth. The SIAD master's program granted me an experience I will cherish for the rest of my life. In a few words I want to recognize and thank those who encouraged and supported me throughout this journey.

To Professor João Carlos Amaro Ferreira, Director of this master, my deepest gratitude for the availability and interest he has always shown. The proximity he creates with the students is a lever for our success, and I believe that if he were not so present in our formation, I would not be delivering this document today. Likewise, I thank Luís Elvas for his commitment to all master's students.

I must also acknowledge the role of my co-supervisor, Øystein Klakegg, who always made himself available to follow the development process of this thesis and accompanied us on the visit to the Norwegian fish farms. In this sense, I truly appreciate the opportunity that Fish2Fork 69 project, EEAGRANTS, gave me to visit this fantastic country and directly contact the fish farming industry.

To my parents, Ilídio Pires and Ana Maria Duarte, who, despite their fears, supported me in this switch of academic path, and stood by me. Love is repaid with love, and this thesis is theirs and for them. I cannot forget my sister, Vitória Pires, who is an example of strength and struggle; if anyone deserves to fulfil all dreams, it is her, for the dedication and focus she puts into her goals.

To my friends Andreia Nóbrega, Susana Lopes and Filipa Santos, for being the constants in my life. In this life of changes and challenges, they always encouraged me to follow what my heart wants and to believe in myself. They, too, are incredible forces and independent women who inspired me never to give up. To Hina Givane, who has been with me through every cycle of university study, lifting me in the lows and celebrating the successes as if they were her own.

To my partner in life, work, and studies, Miguel Mira. For all the nights and weekends we spent working at ISCTE, where we suffered and laughed together. My academic success is the merit of your dedication and belief in me. You always found time for me in the midst of your personal goals, and I cannot express properly how much I value your companionship. Thank you for helping me to make the best choices and prove them right.

Finally, I want to thank ISCTE for being an institution that genuinely cares about its students and provides an enriching and welcoming teaching environment.

RESUMO

Esta tese investiga a aplicação da análise de dados na indústria da piscicultura, centrando-se especificamente na identificação dos parâmetros de qualidade da água que influenciam as taxas de mortalidade do salmão.

O foco desta investigação é examinar um conjunto de dados de cinco anos, explorando minuciosamente o impacto de nove variáveis-chave da qualidade da água - pH, redox, salinidade, temperatura, amónio, nitrato, nitrito, alcalinidade e dióxido de carbono - nas taxas de mortalidade. Seguindo a metodologia CRISP-DM, o ênfase é colocado na preparação de dados, análise de correlações e modelação. O objetivo é descobrir padrões e correlações significativos entre estes parâmetros e a mortalidade do salmão, fornecendo informações valiosas sobre a complexa dinâmica entre a qualidade da água e a mortalidade e, em última análise, contribuindo para a gestão sustentável e eficiente das explorações piscícolas.

Os resultados revelam relações intrínsecas entre as variáveis analisadas, sendo validado que o nitrato e o nitrito apresentam um comportamento paralelo e inverso ao redox e ao pH. A taxa de mortalidade apresenta a maior correlação com o nitrato, e a menor com a temperatura. Este estudo mostra o potencial da análise de dados para melhorar a sustentabilidade e a rentabilidade desta indústria, sublinhando a importância do estudo dos dados em piscicultura para serem utilizados como uma ferramenta de gestão proactiva para o bem-estar dos peixes.

Ao ser dada continuidade a este trabalho, os investigadores podem desenvolver soluções práticas e garantir a produtividade e a qualidade das populações de peixes criadas em sistemas de aquacultura.

Palavras-chave: Análise de dados, CRISP-DM, Mortalidade de peixes, Qualidade da água, Explorações piscícolas, *Salmon Salar*

ABSTRACT

This thesis delves into the application of data analytics in the aquaculture industry, specifically focusing on identifying water quality parameters that influence salmon mortality rates in fish farms. Given the significant economic and environmental implications of high salmon mortality rates, understand and mitigate these rates is paramount.

A comprehensive dataset from a Norwegian fish farm is analyzed, encompassing water quality parameters and mortality rates. This research focus is examining a five-year dataset thoroughly exploring the impact of nine key water quality variables - pH, redox, salinity, temperature, ammonium, nitrate, nitrite, alkalinity, and carbon dioxide - on mortality rates. Following the CRISP-DM methodology, emphasis is placed on data preparation, correlation analysis, and modelling. The objective is to uncover meaningful patterns and correlations between these parameters and salmon mortality, providing valuable insights into the complex dynamics between water quality and mortality, and ultimately contributing to fish farms' sustainable and efficient management.

The results reveal intrinsic relationships between analyzed variables, and it is validated that nitrate and nitrite exhibit parallel behaviour and are inverse to redox and pH. The mortality rate presents the highest correlation with nitrate, and the lowest with temperature. It shows the potential of data analytics in enhancing the sustainability and profitability of the aquaculture industry by underscoring the importance of study the data in aquaculture to be used as a proactive management tool for fish welfare.

By continuing this work, researchers can enhance knowledge, develop practical solutions, and ensure productivity and quality of fish populations created in aquaculture systems.

Keywords: Data analysis, CRISP-DM, Fish Mortality, Water Quality, Fish Farms, *Salmon Salar*

INDEX

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AI: Artificial Intelligence

CO_2: Carbon dioxide

CRISP-DM: Cross-Industry Standard Process for Data Mining

DO: Dissolved oxygen

FAO: Food and Agriculture Organization of the United Nation

IoT: Internet of Things

ML: Machine Learning

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analysis

WoSCC: Web of Science

CHARTER 1

Introduction

In this introductory chapter, a comprehensive overview of the research undertaken in this thesis is presented, encompassing the motivation behind the study, the specific objectives pursued, the methodology employed, and the organizational framework of the thesis. It commences by elucidating the significance of the research topic within the realm of fish farming, emphasizing the criticality of effective water quality management. Subsequently, the precise objectives that serve as the guiding principles of the investigation are expounded upon, aimed at unraveling the intricate relationship between water quality parameters and fish mortality rates. Furthermore, an in-depth exposition of the methodological approach adopted is provided, incorporating rigorous data collection procedures, advanced analytical techniques, and rigorous statistical modeling. Lastly, the thesis's organizational structure is meticulously delineated, delineating the key chapters and their respective contributions to the overarching research endeavor. This chapter serves as a cogent foundation, furnishing readers with a comprehensive grasp of the study's scope and design, thereby setting the stage for subsequent chapters that delve into the empirical findings and conclusive insights derived from this research pursuit..

1.1 Overview

Seafood, one of the world's largest industries, accounts for 12% of global livelihoods, sustaining one in ten people on Earth through aquaculture. With the growing global population projected to reach 9.6 billion by 2050, there is a pressing need for 41% more food. The environmental, economic, and social implications of food consumption and waste are significant, with an estimated global food waste rate of at least 30% [1]. Exploratory fishing has reached its limits, demanding the exploration of sustainable alternatives such as fish farms. Innovative technologies can potentially improve aquaculture's sustainability, transforming it into a knowledge-based production regime [2]. Fish farming offers various benefits, including a lower feed conversion rate compared to other species [3], the ability to restore decimated populations, the contribution to rural development, and efficient use of space. However, it also presents challenges such as pollution from waste products and medical treatments, dependence on environmental conditions, and potential environmental impacts [4]. Emerging technologies can enhance fish farm productivity, operations, and outcomes. As fish farms expand and become more complex, and new site types are utilized, advanced technical solutions are required to monitor and control the biological production process. Machine Learning (ML) holds promise in several areas [5], including:

a) Water Quality Monitoring: ML algorithms can interpret sensor data to monitor parameters such as temperature, pH, and oxygen levels, allowing farmers to identify issues and implement necessary changes to maintain optimal conditions.

b) Disease Detection: ML models trained on data from cameras or acoustic sensors can recognize disease symptoms in fish, enabling early detection and proactive measures to reduce disease transmission and minimize losses.

c) Yield Optimization: ML algorithms can analyze data from various sources to identify opportunities for improving farming operations, enhancing sustainability, maximizing yields, and reducing costs.

d) Fish Behaviour Prediction: By considering factors such as water quality, feeding habits, and environmental conditions, ML models can predict fish behaviour. This information empowers farmers to make informed decisions regarding management techniques and feeding strategies.

e) Automated Feeding Systems: ML algorithms can optimize feeding systems based on fish size, weight, and environmental factors, resulting in improved feed efficiency and reduced waste.

Artificial Intelligence (AI) complements ML, allowing systems to learn from experience and progress without explicit programming. Its ability to efficiently process massive amounts of data, detect patterns and generate predictive models will help optimize various aspects of fish farm operations, including water quality management, disease detection, and feed optimization [6].

Given the challenges posed by directly observing millions of underwater creatures in fish farms, AI algorithms can be developed to access and analyze historical data, providing recommendations for operations such as delousing processes.

AI also offers practical applications in various areas, including:

a) Water Quality Monitoring: AI algorithms can provide real-time data analysis for monitoring temperature, pH, and oxygen levels, enabling informed decisions on water management practices.

b) Disease Detection: ML models trained on data from cameras and acoustic sensors can identify disease signals in salmon, empowering farmers to take preemptive actions to prevent disease spread and mitigate losses.

c) Harvesting Optimization: AI systems can analyze data from multiple sources to enhance harvesting operations, increase yields, reduce costs, and improve the overall sustainability of salmon farming.

d) Market Predictions: AI systems can be trained to forecast market trends and consumer preferences, assisting farmers in making informed decisions about production and marketing strategies.

e) Feed Management: By considering factors such as fish size, growth rate, and water temperature, AI algorithms can optimize feeding systems, reducing waste and improving feed efficiency.

In recent years, advanced technologies, such as sensor networks [7], data analytics, and ML, have provided new opportunities for real-time monitoring and analysis of water quality in fish farms. These technologies enable continuous data collection, allowing farmers to detect potential issues early on and take prompt corrective actions [6] [8]. This study uses rigorous data analysis and modelling to identify the critical factors contributing to fish mortality, providing valuable insights to fish farm operators, which will empower them to focus their efforts, implement effective interventions, and provide better solutions to daily challenges, revolutionizing the industry.

The application of the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology in this thesis aims to leverage the available dataset from a Norwegian fish farm to analyze the influence of nine key water quality parameters, including pH, redox, salinity, temperature, ammonium, nitrate, nitrite, alkalinity and carbon dioxide, on fish mortality. With a deeper understanding of the relationship between water quality parameters and fish mortality, farmers can implement targeted interventions, optimize feeding strategies, and create an environment promoting healthier fish populations.

Technological advancements enhance productivity and profitability, and contribute to the sustainability and resilience of fish farming practices. By shedding light on the complex interplay between water quality parameters and fish mortality, this research aims to contribute to the scientific knowledge base in aquaculture and provide practical insights for fish farm operators. Ultimately, it is anticipated that the findings will facilitate the development of data-driven strategies that enhance fish farms' efficiency, productivity, and sustainability, ensuring a reliable and sustainable seafood supply to meet the growing global demand [9].

1.2 Motivation

The aquaculture industry is vital in meeting the increasing global demand for seafood and alleviating the pressure on wild fish populations. However, fish farms face numerous challenges, including water quality management, disease outbreaks, and high fish mortality rates. Understanding the factors contributing to fish mortality is crucial for improving fish farming operations' efficiency, sustainability, and profitability.

Water quality parameters, such as pH, redox, salinity, temperature, ammonium, nitrate, nitrite, alkalinity and carbon dioxide, directly impact fish's health and well-being. Fluctuations or imbalances in these parameters can lead to stress, weakened immune systems and increased susceptibility to diseases. Consequently, fish disease and mortality rates can rise, resulting in reduced animal welfare and substantial financial losses for farm operators [10]. Addressing the issue of fish mortality requires a comprehensive analysis of the complex relationships between water quality parameters and their impact on fish health. Traditional methods of monitoring and managing water quality rely on periodic sampling and manual observation, which may not capture the dynamic nature of the aquatic environment. Moreover, the sheer volume of data generated in fish farms necessitates advanced analytical techniques to extract meaningful insights and facilitate informed decision-making [2].

By applying CRISP-DM methodology and leveraging a comprehensive dataset from a Norwegian fish farm, this research seeks to contribute to the existing knowledge base, uncovering the correlations between water quality parameters and fish mortality. The insights gained from this study will empower fish farm operators to optimize their operations, improve disease management strategies, and ultimately reduce fish mortality rates. Furthermore, the findings have broader implications for the sustainability of the aquaculture industry. Understanding the relationships between water quality parameters and fish mortality enables fish farms to adopt proactive measures to minimize the environmental impact, conserve resources, and enhance the overall resilience of their operations.

Overall, this research is motivated by the pressing need to address the challenge of fish mortality in fish farms. There is a dearth of research publications describing the main causes of death in fish farms and potential solutions. Most papers reflect that the water quality may be one of the main reasons for diseases and death in this type of production. This leads us to conclude that there is a knowledge gap in fully understanding the origin of the problems and how to solve them. There are already some data analyzes being done, but it is important to compare the results between farms, and not to make an isolated study. This study aims to provide practical insights and actionable recommendations for fish farm operators, contributing to developing sustainable and efficient practices in the aquaculture industry.

1.3 Objectives

Data analytics offers a proactive approach to death management in the aquaculture industry. A wealth of information can be unveiled by harnessing the power of continuous monitoring and meticulous analysis of diverse data sources, including water quality sensors, feed intake monitors, and fish behaviour tracking systems [11]. This data trove is critical to identifying elusive patterns and subtle changes that may serve as harbingers of a looming disease outbreak.

For instance, the sophisticated algorithms of advanced ML can diligently scrutinize video footage, unveiling the delicate nuances that might be early indicators of disease, enabling swift intervention before the affliction takes hold [12]. Simultaneously, vigilant analysis of water quality parameters can expose telltale shifts, unmasking environmental conditions that may foster the proliferation of specific diseases. A potential devastation can be held at bay through diligent isolation of affected populations, precision adjustments to environmental conditions, and administration of treatments.

The invaluable insights gleaned from this analytical odyssey hold the promise of formulating enduring strategies for disease prevention and long-term management. Fish farmers can engineer a proactive tapestry of preventive measures by peering into the records of past outbreaks and meticulously dissecting the conditions and factors that contributed to their emergence. Data analysis fortifies the aquaculture industry against future disease incidents, promoting resilience and sustainability. Moreover, the wisdom gained from these data-driven endeavours shapes a path towards a future where preventive measures dominate, fortifying fish farms against the ravages of disease [2].

The goal with this thesis is to apply the CRISP-DM methodology to a fish farm dataset. In this regard, we aim to:

1. Clean and transform the data, making it appropriate for analysis and reducing bias. This dataset treatment allows the conclusions to be reliable and enables future studies to be made on them.

2. Utilize modelling techniques to uncover valuable insights regarding the impact of water quality on fish mortality and provide practical insights on the analyzed water parameters. Through a detailed examination of the data related to salmon, the goal is to validate the existence and quantify the strength of the relationship between nine water quality variables and the target variable.

The findings of this thesis contribute to a better understanding of the relationship between water quality and fish mortality and serve as a foundation for further research in this field. Furthermore, the knowledge gained from this study will be instrumental in developing predictive models that can assist fish farm operators in optimizing their production conditions.

1.4 Methodology

Data analytics has yet to realize its full potential in the field of fish farming production. Although Internet of Things (IoT) technologies for control and monitoring are being invested in, their high cost and the scarcity of skilled professionals have hindered their widespread development and adoption. This implies that there is a lack of data available in this industry to analyze [13].

In order to achieve the objective, the CRISP-DM methodology has been chosen. This methodology promotes best practices and allows for project replication. In addition, it provides a solid foundation for activity planning and management, and it can be applied to any Data Science project, regardless of the specific subject matter.

The CRISP-DM methodology consists of six distinct phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. Each phase contributes to the project's overall success and ensures a systematic approach to data analysis [14]. By following the CRISP-DM methodology, we aim to comprehensively understand the business requirements and goals related to fish farming production. We then acquire a deep understanding of the available data, including the nine water quality parameters and fish mortality rates. Follows the data preparation which will ensure its quality and suitability for analysis. Next, we employ various modelling techniques to explore the relationship between the water quality parameters and fish mortality rates. These results are evaluated and refined to ensure their accuracy and reliability. Finally, we provide actionable insights and recommendations for fish farm operators.

Figure 1.1 visually depicts the application of the CRISP-DM methodology to address the specific problem at hand, illustrating the structured and iterative nature of the process.

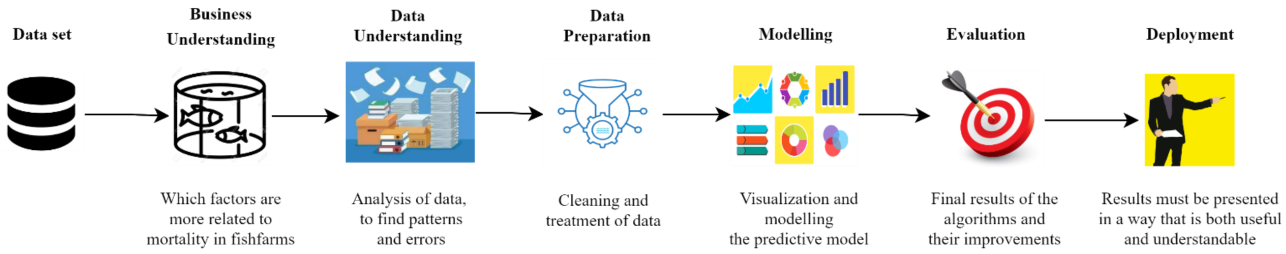
Data set
Business Understanding
Data Understanding
Data Preparation
Modelling
Evaluation
Deployment
Which factors are more related to mortality in fishfarms
Analysis of data, to find patterns and errors
Cleaning and treatment of data
Visualization and modelling the predictive model
Final results of the algorithms and their improvements
Results must be presented in a way that is both useful and understandable

Figure 1.1 -CRISP-DM methodology applied to data analysis in fish farms

By employing the CRISP-DM methodology, the objective is to harness the power of data analytics and contribute to the advancement of fish farming practices. This ultimately leads to improved production efficiency, better decision-making processes, and enhanced sustainability in the aquaculture industry.

Due to the nature of the available data, applying a predictive model in this study is not feasible. The dataset provided for analysis primarily focuses on testing purposes, lacking the information and context required for predictive modelling. Predictive models rely on historical data with known outcomes to establish patterns and relationships for future predictions. However, in this case, the dataset primarily consists of experimental or observational data, which does not provide the necessary historical records or outcome variables required to build a predictive model. The absence of such data limits the ability to train a model and make accurate predictions. Therefore, this study focuses on the exploratory analysis and establishing relationships between the water quality parameters and fish mortality.

1.5 Dissertation Context

The EEAGrant project[2], Fish2Fork[3], provide access to Norwegian fish farm data, allowing us to investigate the impact of water quality on fish mortality in fish farms.

EEA Grants and Fish2Fork are distinct initiatives focused on sustainability and cooperation. The EEA Grants serve as a financial mechanism, fostering solidarity within the European Economic Area by providing grant funding to eligible countries for projects promoting social inclusion, environmental protection, and sustainable development. On the other hand, Fish2Fork is an innovative online platform dedicated to marine conservation, offering consumers, businesses, and policymakers reliable information on seafood sustainability. By encouraging responsible fishing practices and promoting sustainable seafood consumption, Fish2Fork plays a crucial role in preserving marine resources and raising awareness about the urgent need for conservation. These initiatives contribute significantly to building a more cohesive, environmentally conscious, and prosperous world.

This thesis aims to underscore the importance of data analysis by presenting a case study centered on salmon. The availability of data primarily influenced the selection of salmon as the subject of analysis. However, it should be noted that the methodologies employed in this study can be applied to any other similar case. With the data available within the project context, the objective is to analyze the relationship between nine key water quality parameters and fish mortality rates. The dataset includes information on 19 groups of fish in the juvenile stage, with a specific focus on the *Salmo Salar* species.

1.6 Thesis organization

The organization of this thesis is designed to provide a clear and logical structure for presenting the research findings and supporting arguments. It is structured as follows:

Chapter 1 - Introduction: This chapter presents the background, rationale, and objectives of the study, providing a comprehensive overview of the research topics and their significance.

Chapter 2 - State of the Art: This chapter examines the existing literature and research within the field, encompassing the search strategy, inclusion criteria, data extraction, results, and related work.

[2] https://eeagrants.org/

[3] https://fish2fork.eu/

Chapter 3 - Methodology: Crisp-DM: In this chapter, the data collection methods and analytical techniques employed in the study are discussed in detail, ensuring transparency and reliability.

Chapter 4 - Insights: The clean data is analyzed and interpreted, presenting the study's findings by looking at the values by group, variable and globally.

Chapter 5 - Conclusions: The final chapter summarizes the key findings, provides a comprehensive discussion of the results, and critically analyzes their implications and relevance. Additionally, recommendations for future work directions and practical applications are provided.

CHARTER 2

State of the Art

The state of the art in our research field is a complex and dynamic landscape that requires a thorough examination. In this chapter, a meticulous search strategy and well-defined inclusion criteria are employed to gather a diverse range of relevant literature from reputable academic sources. This comprehensive approach ensures the reliability and validity of the study. Following the data extraction phase, the information is carefully analyzed and synthesized, highlighting key findings, methodologies employed, and knowledge gaps within the field. Moreover, related work is explored, delving into the contributions of other scholars and practitioners to contextualize the research and identify areas that warrant further investigation. By thoroughly examining the state of the art, the objective is to build upon existing knowledge, address gaps, and contribute to advancing the research field.

2.1 Search Strategy and Inclusion Criteria

In addition to being a significant component of men's diets worldwide, fish is crucial to the world's food and nutritional security. Therefore, Aquaculture has emerged as a crucial option that can assist people in meeting their needs for animal proteins with a fisheries origin due to the lack of fishery resources and the overexploitation of water bodies. Moreover, water quality is one of the significant elements affecting fish’s growth, reproduction, and production.

In the initial approach practiced to collect information for state of the art, it was found that there are several studies related to fish farms. However, the most common themes found are related to new technologies being developed for monitoring and alerting - mainly in IoT and robotics. These tools allow for real-time monitoring, compare data with history to create alerts, and reduce the manual labour that is currently prevalent [15]; manual labour concerns employees who have the tasks of analysing various parameters in the water and fish and their interaction with these tanks often leads to less accurate measurements and disruption of marine life at these sites. Eliminating humans from these extractions increases the range of measurements, reduces the impact on fauna and flora from human action, and reduces human error.

Another much-discussed topic, although the conclusions are far from unanimous, is the pollution caused by fish farms. This industry has an impact on the surrounding environment, i.e., the effluents that are deposited in the soil pollute it and impair its quality for other uses (such as some types of agriculture); as for the water, they do not have the proper treatment for its return to the sea/river, carrying with their antibiotics, contamination agents, nutrients and sometimes even residues of infestations and parasites. In addition, some populations in the world depend on local waters for consumption and production, so the contamination of these waters will reflect in diseases and impoverishment of the entire surrounding area. These concerns have led to legal measures for antibiotics and wastewater treatment. However, enforcing these directives must still be solid and assertive enough to ensure healthy and sustainable production.

Through close collaboration with the Fish2Fork initiative, which seeks to establish a collaborative network to facilitate information exchange and cooperation among stakeholders in the fish industry, the importance of analyzing and sharing data is understood. This project aims to implement innovative and disruptive technologies such as IoT, Blockchain, data analytics, and advanced logistics and operations management techniques for fisheries. These cutting-edge solutions will enable enhanced transparency, effective monitoring of illegal fishing activities, and implementing more efficient processes in both wild fish capture and aquaculture operations. Our contribution is the understanding of the influence of water quality parameters on fish mortality on fish farms.

The search was made with the PRISMA[4] methodology, and occurred in March 2022, in Web of Science[5] (WoSCC) and Scopus[6] repositories. The results concerned 15 scientific articles and 2 conference papers between 2000 and 2022, written in English and Portuguese.

2.2 Data extraction and results

A qualitative analysis based on PRISMA was carried out for data analysis. The published work relevant to the topics was rigorously searched in WoSCC and Scopus using the concepts “water quality”, “fish” and “mortality”, the population “fish farm” and “fish farm”, and within the context of “analysis”, “pH”, “nitrite”, “nitrate”, “salinity”, “carbon dioxide”, “alkalinity”, “redox”, “temperature” and “ammonium” – variables that we have in your data.

In each repository, the query utilized was ("water quality" AND "fish" AND "mortality") AND ("fish farm" OR "fishfarm") AND ("analysis" AND "ph" OR "nitrite" OR "nitrate" OR "salinity" OR "carbon dioxide" OR "alkalinity" OR "redox" OR "temperature" OR "ammonium").

The use of pre-defined keywords oriented to the specific problem allowed 17 articles to be returned - 10 in Scopus and 7 in WoSCC. The selection of terms used was rigorously thought out in order to obtain information targeted to our problem. For the concept, the parameters of the data

[4] http://www.prisma-statement.org

[5] clarivate.com/webofsciencegroup/solutions/web-of-science

[6] https://www.scopus.com

available to us and the category of the object studied, "fish", were chosen. This last filter allowed us to reduce the sample, discarding articles referring to other marine faunas and floras – the same reason was also used to exclude the term “aquaculture”. Regarding the methodology used, “CRISP-DM” was also considered for the keywords, but its applicability is very generic, although it has yet to be explored to analyze data on fish farms; In this sense, its use with “OR” returned articles was totally outside the scope of the study, and with “AND” eliminated any results. Concerning the population, the sample was intended to be limited to fish in fish farm environments, omitting all those in the sea, rivers, lakes. During the development of this work, it was observed that there was no consensus on how to write the concept fish farm. Therefore, both variations of the term – “fishfarm” and “fish farm”- were included to ensure comprehensiveness. In the context, the search was further refined by incorporating the terms related to the variables available for analysis in order to identify articles that refer to similar studies. The 17 articles that fit this query were all produced in English and Portuguese, and since the authors feel comfortable in both languages, they were all accepted for a first exploration.

From 17 papers retrieved, an in-depth analysis was done. 7 of the studies were rejected because they were deemed to be outside the analysis's purview: they concern to transportation of fish, prawns, pollution of nearby waters, lake fish, microbial and dissolved oxygen (DO)- a parameter we don’t have. As no duplicate documents were found, a more detailed exploration of the 10 selected articles was conducted.

In addition, Food and Agriculture Organization of the United Nations (FAO) was also accessed to provide some statics about the fish farms.

2.3 Related work

Based on the articles reviewed, a comprehensive state-of-the-art analysis emerges, shedding light on various aspects of water quality management in aquaculture. One study, Sanou et al., focuses on the impact of water temperature and salinity on fish growth and survival. The values of the measured parameters were compared with the standards recommended for the breeding of tilapia Oreochromis niloticus. The findings indicate that optimal temperature and salinity levels promote better growth performance and lower mortality rates in fish populations [16]. Similarly, another study by Pereira et al., monitored and analyzed biotic and abiotic parameters of water quality in production ponds from a temperate aquaculture. It is proved that maintaining adequate DO concentrations supports fish health and reduces stress-related mortalities [17].

The work of Sousa-Filho et al., employs multivariate statistical analysis to investigate the correlations between water quality parameters and fish mortality rates. The research findings emphasize the importance of monitoring parameters such as temperature, DO, ammonia, and pH to ensure optimal water quality conditions and minimize fish mortalities [18].

Comparing these studies, common themes emerge. The importance of temperature, salinity, dissolved oxygen, and water quality parameters such as ammonia and pH in influencing fish health and mortality rates is consistently highlighted. Additionally, several authors like Adeogun et al., Tedesco et al., and El-gohary et al., delve into the impact of water quality on disease prevalence and immune responses in fish populations [19], [20], [21]. El-gohary et al., found, also, that physicochemical water analysis revealed different ranges in relation to the fish farms and seasons [21]. The three works demonstrate that poor water quality conditions, such as low DO and high ammonia levels, can compromise the immune system, making fish more susceptible to diseases. Furthermore, the findings emphasize the importance of tailored water quality management strategies that align with specific aquaculture systems and species. For instance, studies focusing on Nile tilapia done by Teixeira et al., provide insights into this species' growth performance and mortality rates under different water quality conditions. The author proves that the cause of the fish mortality was attributed to opportunistic infection by imbalances in water quality [22]. This knowledge enables aquaculture practitioners to make informed decisions regarding temperature, DO, and ammonia levels to support the well-being of Nile tilapia populations. Researchers Ferreira et al., Johnsen et al., and Ali et al., explores the effectiveness of various water quality management strategies in mitigating fish mortalities. These studies propose biofiltration systems, aeration techniques, and proper waste management to maintain optimal water quality conditions and enhance fish survival [23], [24], [25]. Johnsen et al. and Ali and al. assess the impact and conditions that caused parasite blooms and real severe losses. Water quality influences the growth and development of these parasites, and there is an immediate need to control and identify root causes [24], [25].

In analyzing and comparing these articles with the core theme of this thesis, it becomes evident that water quality management plays a critical role in achieving the goals of sustainable aquaculture. The studies highlight the impact of various water quality parameters, such as temperature, salinity, DO, ammonia, and pH, on fish health and mortality rates. By understanding the relationships between these parameters and their effects on fish populations, researchers and industry practitioners can implement proactive measures to optimize water quality conditions.

The comparison of these articles reveals both commonalities and variations in their approaches and findings. While each study focuses on specific aspects of water quality management, they collectively contribute to a more comprehensive understanding of the topic. The identified correlations between water quality parameters and fish mortalities provide a solid foundation for developing tailored management practices that address the specific needs of different aquaculture systems and species. Overall, these studies underscore the crucial role of water quality management in achieving sustainable aquaculture practices. By incorporating the insights gained from these articles into the broader framework of the thesis, a comprehensive and informed approach to water quality management can be developed, ensuring the well-being, productivity, and longevity of fish populations in aquaculture systems.

Table 2.1 was created to develop state of the art, which cross-references the documents analyzed and some of the keywords each contains. It was found that no results were presented for the parameters CO_2 and redox, highlighting, instead, the analysis of temperature and pH as the most studied parameters.

Table 2.1 - Key words on the selected articles

Paper	1	4	6	8	9	11	12	13	14	16
Year	2022	2011	2022	2014	2022	2020	2019	2010	2020	2000
Fish specie	Typalia	Bream	*Gymnotus carapo*	Five different species	Trout & *Salmo Salar*	Typalia	Pacu	*Salmo salar*	Typalia	Trout
Local	France	Portugal	Brazil	Niger	Italy, Spain & Scotland	Egypt	Brazil	Norway	Egypt	France
Mortality		X	X	X	X	X	X	X	X	X
pH	X	X		X	X	X	X		X	
Redox										
Salinity		X					X	X		X
Temperature	X	X			X	X	X	X		X
Ammonium	X									
Nitrate	X			X		X			X	
Nitrite	X					X			X	
Alkalinity				X						
Carbon dioxide										

CHARTER 3

Methodology: CRISP-DM

The Methodology chapter of this research adopts the widely recognized CRISP-DM methodology. CRISP-DM provides a structured approach to guide the exploration, analysis, and interpretation of data systematically and comprehensively. To ensure transparency, repeatability, and rigor in data mining, this study aims to outline the various stages of CRISP-DM in this chapter, including business understanding, data understanding, data preparation, and modeling. Each stage is described in detail, highlighting the specific techniques and tools utilized for data collection, pre-processing, feature selection, modeling, and evaluation. The utilization of CRISP-DM framework ensures a methodical and rigorous approach to uncover valuable insights from the data, contributing to the overall objectives of this research.

3.1 Business Understanding

In the Business Understanding phase of the CRISP-DM methodology, the objective is to understand the business requirements and goals related to fish farming production and the influence of water quality parameters on fish mortality. By conducting an industry analysis and engaging with relevant stakeholders, the aim is to identify critical challenges and opportunities in fish farming. The primary goal of this phase is to define the project's scope and establish clear objectives. To gather insights into the specific concerns and priorities in fish farming production, collaboration is established with fish farm operators, aquaculture experts, and researchers. This phase involves understanding the current practices, identifying pain points, and determining the potential benefits of improving water quality conditions. Furthermore, key performance indicators are identified to evaluate the success of the analysis and any subsequent interventions or recommendations.

It is intended that the research is aligned with the practical needs and challenges faced by fish farm operators. This understanding provides valuable insights for developing practical solutions and making informed decisions throughout the project. Overall, the Business Understanding phase establishes the groundwork for the subsequent stages of the CRISP-DM methodology, enabling progress with a clear comprehension of the business context, objectives, and the potential impact of enhancing water quality parameters on fish mortality in fish farms.

Given the defined objectives, at this stage, it is crucial to thoroughly understand the data that has been provided. The data analyzed is from a Norwegian fish farm. In the period between January 1, 2018, and December 31, 2022, the company collected daily mortality values on 19 groups of fish and took water quality samples to assess 9 parameters. These groups of fish belong to the same species, *Salmo Salar*, and were in the juvenile stage (fry) department, and sometimes some groups share the same period in the department but in different tanks.

As for the study site, it has 21 tanks with water recirculation, 7 cubic meter volume, and 3 meters across in size. The water exchange rate is low initially and increases as the fish grow. Changing all water in 60 minutes is a reasonable estimate for each tank's average water exchange rate. The water samples are taken from the same place where they have their water quality sensors, and the sensors continuously measure pH, salinity, temperature and redox; ammonium, nitrate, nitrite, alkalinity and CO_2 have to be measured in the lab.

Mortality is determined by collecting and counting all dead fish in each tank; Once per day, sometimes twice if the mortality is higher than usual. Then they use a computer program to register all mortality per tank daily. Based on the input data, the program keeps track of how much fish are in each tank at all times. Only fish that are dead are registered as mortality. Usually, weak, sick, or deformed fish are also collected and euthanized with an overdose of a sedative/anesthesia. These fish are never registered as mortality (because they do not die naturally); This is called *culling*, and the register is made separately. As a result, the «total loss» is always bigger than what is registered as mortality.

3.2 Data Understanding

In the Data Understanding phase, the focus shifts towards analyzing the provided data to gain insights, identify patterns, and address potential errors. This phase serves as a foundation for further analysis and modelling.

The analysis is initiated by importing the required libraries and data files, enabling effective access and manipulation of the data. Moreover, data visualization plays a crucial role in understanding the data's characteristics, facilitating the description and classification of variables for enhanced comprehension. The data utilized in this study originates from a Norwegian fish farm, encompassing water quality information spanning from January 1, 2018, to December 31, 2022. The original dataset comprises 1641 rows and 38 columns. The provided variables include pH, redox, salinity, temperature, ammonium, nitrate, nitrite, alkalinity and CO_2. The pH, redox, salinity, and temperature measurements were recorded daily, while the other variables were measured weekly. Throughout this period, 19 distinct groups of fish were in production, with daily Mortality values serving as the target variable.

It is important to note that a conversion to a consistent time frame, such as weekly, is necessary due to variations in the timing of measurements. Additionally, observations reveal instances of missing values in the variables measured weekly. Addressing these missing values follows the principles of good practices, where suitable methods such as substitution or elimination will be employed.

The Data Understanding phase enables us to delve into the provided data, identify its structure, and comprehend the nature of the variables and their relationships. This exploration lays the groundwork for subsequent stages, aiding in developing meaningful insights and formulating effective analytical strategies.

Variable Analysis

In this study, 9 water quality parameters were analyzed. An explanation of each variable follows.

a) pH – measure of the acidity of the water, which impacts the stress, growth and reproduction on fish and aquatic animals [26].

b) Redox – chemical reaction in which the oxidation number of a molecule, atom, or ion change by acquiring or losing an electron is referred to as an oxidation-reduction reaction. Some of the fundamental processes of life, such as photosynthesis and respiration, depend on redox reactions. In our case, oxidation-reduction is used to remove impurities from water [27].

c) Salinity – total concentration of dissolved salts in water, witch directly impact fish body metabolism and osmoregulation [28].

d) Temperature – impacts fish reproduction and metabolic rate, which in turn has an impact on how they balance their energy, how they move around and the feeding behaviour [29].

e) Ammonium – ionized form of poisonous ammonia (NH3), NH4, a harmless salt. It is helpful to know that while NH4 is not harmful in the aquatic environment, it can immediately transform into NH3 in response to changes in pH or temperature [30].

f) Nitrate – the persistence of high levels above the recommendation increases the risk of sickness in fish and prevents them from reproducing; High nitrate concentrations are particularly detrimental to young fish and have a negative impact on their growth [31] [32].

g) Nitrite - intermediate in the oxidation of ammonium to nitrate that converts hemoglobin to methemoglobin, which does not carry oxygen - can result in anoxia in fish and other aquatic creatures [33].

h) Alkalinity – measure of water's capacity to neutralize acids or fend against changes that make it acidic while keeping its pH steady [28].

i) CO_2 – decreases a fish's blood's capacity to transport oxygen, so large quantities can be harmful. The physiological effects of high CO_2 in water are primarily ascribed to decreased body fluid pH [34].

3.3 Data Preparation

During the Data Preparation phase of the CRISP-DM methodology, the focus is on transforming and pre-processing the data to ensure its suitability for analysis and modeling. This phase is crucial in enhancing data quality, addressing missing values, and creating consistent and standardized datasets.

To begin, the time discrepancies in the measurements are addressed by converting all variables to a uniform time frame, which, in this case, is weekly. This conversion enables better alignment and comparison of data across different variables. Subsequently, the presence of missing values in the dataset is examined. Missing values can significantly impact the accuracy and reliability of our analysis. Various techniques, such as imputation or elimination, handle these missing values following best practices. The chosen method depends on the nature of the missing data and its potential impact on the overall analysis. In addition to addressing missing values, data cleaning and validation procedures were processed, which involve identifying and rectifying any inconsistencies, outliers, or errors in the dataset. By implementing suitable data cleansing techniques, the reliability and accuracy of the data used for analysis can be ensured.

Moreover, the relevance and significance of each variable in relation to the research objective are considered. Variables that do not contribute significantly to the analysis or exhibit high collinearity may be excluded to simplify the dataset and improve the accuracy of subsequent modeling processes.

Throughout the Data Preparation phase, the goal is to establish a well-structured and reliable dataset that is ready for analysis and modeling. This involves addressing time discrepancies, handling missing values, cleaning the data, and selecting relevant variables. These steps lay the foundation for generating meaningful insights and conducting effective data analysis.

Duplicated records

It was validated that the data provided contains only one measurement per variable and date, so duplicate deletions were unnecessary.

Erase variables that do not contribute for the study

The following columns were erased due to issues:

1. Closing Date – will be the linking column between the Mortality and Water Quality tables.
2. Time - unnecessary; all measurements were taken at 12 pm.
3. Spedevann - the variable only shows measurements from September 23, 2021.
4. CaCO3 - a variable with an insignificant number of measurements.
5. pH Avløp – Avløp means " return water, out from the tanks ". Only half of the variables show these measurements, so we chose to stay with the "Bio" measurements - "input water".

6. Ammonium Avløp – Avløp means " return water, out from the tanks ". Only half of the variables show these measurements, so we chose to stay with the "Bio" measurements - "input water".
7. Nitrate Avløp – Avløp means " return water, out from the tanks ". Only half of the variables show these measurements, so we chose to stay with the "Bio" measurements - "input water".
8. Nitrite Avløp – Avløp means " return water, out from the tanks ". Only half of the variables show these measurements, so we chose to stay with the "Bio" measurements - "input water".
9. CO_2 Avløp - Avløp means " return water, out from the tanks ". Only half of the variables show these measurements, so we chose to stay with the "Bio" measurements - "input water".
10. Sm 19-1 - This group was culled due to a severe fungus outbreak.
11. Sm 20-3 - This group of fish shows measurements on only 17 days.
12. Sm 20-4 - This group of fish shows an absence of measurements in 1/3 and 1/2 of the data, depending on the variables.
13. Sm 21-4 – *rainbow trout*, instead of Atlantic Salmon, like the others.
14. Sm 22-5 - This group of fish shows measurements on only one day.

This results in 1215 rows for 24 columns. Our work evaluates 14 groups of fish for nine variables: pH, redox, salinity, temperature, ammonium, nitrate, nitrite, alkalinity, and CO_2.

Outliers

Regarding outliers, Table 3.1 presents the reference values for the variables analysed in this study and the range of values in our dataset [28], [35], [27], [36], [37], [38], [39].

Table 3.1 - Reference values in comparison with the range of values in our data

Variable	Reference values	Range of values in data
Alkalinity	60-300 mg/l	1.44 – 61.2 mg/l
pH	5.5 – 8.0	1.17 – 14
Redox	<320 mV	212 – 396 mV
Salinity	0-30 ppt	0.03 - 7.28 ppt
Nitrate	<150 mg/l	0.02 - 215 mg/l
Ammonium	*	0.1 - 11 mg/l
Nitrite	<0.75 mg/l	0.02 - 4.4 mg/l
Temperature	< 14ºC	8.3 – 17.3 ºC
CO_2	<15 ppm	1- 13 ppm

*Dangerous when pH values are above 10

The following conclusions were reached:

- **pH:** pH values outside the 5.5-8 range can cause stress on the fish, affecting their growth and reproduction. Based on these reference values and the visualization of this variable, values below 5 and above 10 are considered outliers and were eliminated. Three records were deleted because they were considered extreme values with fatal consequences for the fish: 1.17, 14, 14.
- **Redox:** an essential factor for determining the relative oxidation levels and reduction in aquaculture ponds. When the redox increases, the pH decreases, which is why it is a variable to be considered. There were 96 values found above the reference values, so excluding so much data would not be feasible or credible for the results. Considering that the pH values remained within the accepted measurement values, it was decided not to exclude any data for this variable.
- **Salinity:** Increases may impact how the osmoregulation process uses energy, affecting the growth rate and feed consumption. Considering that the measurements do not exceed the reference values, we decided not to exclude any outliers.
- **Temperature:** temperature only starts to create stress in the fish when it reaches values above 18º C and can be deadly at 21º C. Since the measurements obtained are within the reference values, we chose not to exclude any value.
- **Ammonium:** ammonium, or NH4, is not dangerous for fish. However, when pH rises above 7, NH4 turns into NH3 (ammonia), which is highly toxic; NH3 impacts the central nervous system and can result in convulsions and death -"acute ammonia poisoning". Since only two pH values greater than 10 were excluded as they were considered outliers, maintaining all values of this variable seems rational.
- **Nitrate:** nitrate levels below 80 mg/l are not dangerous for fish; however, when reached, nitrate poisoning can occur, characterized by lethargy, poor color, compromised immune system, and a decreasing reaction to feeding. Since there are several values above the referenced one, and we cannot disregard so many measurements, we consider outliers to be those greater than and including 150 mg/l – and remove them. This resulted in the exclusion of 9 values: 160mg/l, 170mg/l, 175mg/l, 175mg/l, 200mg/l, 200mg/l, 200mg/l, 210mg/l, and 215mg/l.
- **Nitrite:** nitrite causes fish stress at values above 0.75 mg/l and can be toxic above 5 mg/l. The hemoglobin in fish blood reacts with nitrite to generate methemoglobin, which is dangerous to fish – they may suffocate even when there is enough oxygen because hemoglobin transports oxygen through the body, but methemoglobin does not. Given that the collected measurements are all below the toxic value, it was decided not to remove any outliers.
- **Alkalinity:** directly related to pH, i.e., when one increases, the other follows, and vice versa. In this sense, alkalinity below 20 mg/l potentially harms fish farming. Although we find measures below recommended, the periods of exposure to these values would have to be longer to have drastic consequences. We choose to keep all the values collected for this variable.

- **CO_2:** High levels can be harmful because they make it harder for fish's blood to carry oxygen. Fish in water with high CO_2 concentrations can still suffocate even with high oxygen levels. Since our reference value are below 15 ppm, all values were maintained.

Under the assumptions made in identifying outliers, records that may bias the results were eliminated. In this sense, we removed 3 pH values and 9 nitrate values. This action reduces our sample to 1203 rows for 24 columns.

After this step, all data were converted to the same period - weekly - so they could be compared; In this sense, for the variables that presented daily values - Mortality, pH, redox, salinity and temperature - an average of these values was made. A Python function was created to achieve the desired result, averaging the values every 7 days. Considering that replacing 'Nan' with 0 could influence the average weekly value, when the function found an empty field, it ignored it. Instead, it did the average according to the number of actual values it found every 7 days, avoiding biasing the results. With this activity, the sample was reduced to 176 values for all variables.

Missing values

There are two potential solutions to address missing values in the dataset: imputation or deletion. However, before deciding on the appropriate approach, it is crucial to understand the nature of the missing values [40]. Three types of missing values can occur:

- Missing at Random (MAR): In this case, the missingness is unrelated to the studied subject. Although a relationship exists between the missing values and other variables, the missing data is random [41].
- Missing Completely at Random (MCAR): MCAR implies that the causes of the missing data are unrelated to the data itself. There is no discernible relationship between the missing values and other variables [41].
- Missing not at Random (MNAR): MNAR indicates that the missing values depend on the hypothetical value or the value of another variable. In this scenario, there is a relationship between the missing values and other variables; the missing data is not random [41].

Our study determined that the missing data occurred due to random human failures, such as forgetting to measure or record some values. Therefore, the missing values can be considered as MCAR. The best practice recommends that the missing data can be predicted based on the observed ones, as the aim is to avoid generating skewed estimates that could potentially result in erroneous outcomes.

By assuming MCAR and employing suitable imputation and deletion techniques, the missing values can be filled or removed based on observed patterns in the data. This approach aims to retain as much information as possible and minimize potential bias. Imputation is utilized when the quantity of missing values does not exceed 25-30%. Mean, median, or mode can be applied to these values: the mean provides the most accurate estimate for the actual dataset, the median is the optimal assessment when there are many outliers or extreme values in the dataset, and the mode is employed when the data is skewed. Deletion is applied when a variable has at least 30% missing values [42], which is the case in some fish groups

In 5 variables - ammonium, nitrate, nitrite, akalinity and CO2 - there are missing values; Table 3.2 exposes the percentage of missing values for each group of fish, highlighting in red the most critical situations and in yellow the borderline situations.

Table 3.2 - Missing values per group and variable

Fish Group	Data per variable	Ammonium	Nitrate	Nitrite	Alkalinity	CO2
Gr.1-18	18	5,26%	5,26%	5,26%	5,26%	31,58%
Gr.2-18	12	0%	0%	0%	41,67%	25%
Sm 19-2	14	28,57%	28,57%	28,57%	28,57%	35,71%
Sm 19-3	12	8,33%	8,33%	8,33%	33,33%	33,33%
Sm 19-4	8	0%	0%	0%	37,5%	25%
Sm 19-6	14	28,57%	28,57%	28,57%	28,57%	35,71%
Sm 20-1	13	15,38%	15,38%	15,38%	69,23%	23,08%
Sm 20-2	7	28,57%	14,28%	14,28%	85,71%	57,14%
Sm 21-1	12	25%	25%	25%	91,67%	25%
Sm 21-2	18	11,11%	11,11%	11,11%	77,78%	11,11%
Sm 21-3	12	25%	25%	25%	91,67%	25%
Sm 21-5	13	7,69%	7,69%	7,69%	23,08%	7,69%
Sm 22-1	15	20%	20%	20%	80%	60%
Sm 22-2	8	25%	37,5%	25%	50%	25%
Total	**176**	**16,48%**	**16,48%**	**15,9%**	**51,7%**	**28,41%**

In our specific case, the data provider recommended using the mean for imputation, specifically the average of the previous week's values with the later one. It was assured that this approach would yield results very close to reality.

Table 3.3 presents the percentage of deleted values and, among the retained values, the number that were imputed. It is worth noting that even though the total number of missing values for alkalinity exceeded 30%, was decided to retain those from the groups with a percentage below this threshold.

These variables' analysis will be presented by group and for the entire dataset. Given the small sample size, preserving the maximum number of values possible allows us to draw more robust and insightful conclusions.

Table 3.3 - Data deletion and imputation, per variable

Variable	Data deleted	Data after deletion	Data imputation	Data with imputation
Ammonium	0%	176	16,48%	29
Nitrate	4,54%	168	17,26%	29
Nitrite	0%	176	15,9%	28
CO_2	45,45%	96	20,83%	20
Alkalinity	61,93%	67	17,91%	12

3.4 Modeling

The data can be leveraged to gain insights through visualization techniques during the modeling phase. Visualizations such as histograms, heatmaps, and correlation maps enable us to identify the variables that have a significant impact on our target variable, which, in this case, is fish mortality.Once the data has been thoroughly understood, treated, and cleaned, it becomes crucial to examine its correlation. It is relevant to understand that two correlation coefficients can be applied according to the data under analysis: Pearson's and Spearman's. Pearson's coefficient is commonly used to measure the strength and direction of a linear relationship between continuous variables, assuming normality and linearity. On the other hand, Spearman's coefficient evaluates the monotonic relationship between variables, making it suitable for ordinal or non-linear data. Researchers can make informed decisions when exploring the relationships between variables by understanding the characteristics and appropriate usage of these correlation coefficients. In our analysis, we have chosen to use Pearson's coefficient due to the nature of our continuous variables and the exclusion made of outliers from the study.

The correlation coefficient is a metric to measure the extent of the relationship between two variables, with values ranging from -1.0 to 1.0. It is important to note that correlation coefficients cannot exceed 1.0 or fall below -1.0. A correlation of -1.0 indicates a perfect negative correlation, while a correlation of 1.0 represents a perfect positive correlation. A correlation coefficient above zero signifies a positive association, whereas a value below zero indicates a negative association. A value of zero suggests no correlation between the variables [43].

With knowledge of one variable's value, it’s possible to more accurately predict the value of another variable closely linked to it. The accuracy of these predictions improves as the strength of the association among variables increases. It is worth noting that while correlation provides valuable insights, it does not necessarily imply causality. Therefore, it is essential to exercise caution when inferring causal relationships based solely on correlation.

By analyzing the correlations among the variables, it is possible to uncover meaningful patterns and relationships that will inform subsequent modeling and analysis techniques. These insights will enable the drawing of reliable conclusions from the data.

The main goal of this thesis is to analyze 9 water parameters and their influence on the target variable - Mortality. Given the importance of the theme, it was decided to create a separate chapter where the entire study and conclusions to be drawn from the sample provided are presented in detail.

CHARTER 4

Insights

The Insights chapter presents detailed analyses of water quality parameters and their relationship with the target variable - fish mortality. These analyses were conducted for different fish groups and the entire dataset. The objective is to gain valuable insights into the impact of water quality on fish mortality and identify key variables that contribute to this outcome. Through comprehensive statistical analyses and visual representations, this chapter aims to provide a comprehensive understanding of the complex relationship between water quality parameters and fish mortality, shedding light on potential factors and patterns that influence the survival of fish in the aquaculture systems.

4.1 Analysis by group

Initially, for each of the 14 groups of fish analyzed, an individual analysis of water quality parameters and their relationship to mortality was performed. Table 4.1 shows the variables used in each group; it is visible that 9 water quality parameters contained sufficient data for analysis except for Alkalinity and CO_2, which can only be studied in 4 and 7 groups out of 14, respectively.

Table 4.1 - Variables used on each fish group

Fish group	1	2	3	4	5	6	7	8	9	10	11	12	13	14
pH	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Redox	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Salinity	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Temperature	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Ammonium	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Nitrate	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Nitrite	X	X	X	X	X	X	X	X	X	X	X	X	X	
Alkalinity	X		X			X						X		
CO_2		X			X		X		X	X	X	X		

Interpreting the size of a correlation coefficient is a fundamental aspect of data analysis. Correlation coefficients measure the strength and direction of the relationship between variables. Understanding the magnitude of the correlation coefficient is crucial for assessing the practical significance and meaningfulness of the relationship. Table 4.2 presents the interpretation rule used; It was chosen to be more cautious in the interpretation due to the small size of the analyzed samples.

Table 4.2 - Interpretation of correlation by size

Size	Interpretation
0 - |0.30|	Very Low
|0.30| - |0.50|	Low
|0.50| - |0.70|	Moderate
|0.70| - |0.90|	High
|0.90| - |1|	Very High

The 14 correlation matrices developed for each of the fish groups are presented in Figure 4.1 to Figure 4.14, followed by their analysis. On the left is shown the range of correlation by colors: red for positive correlations and blue for negative ones.

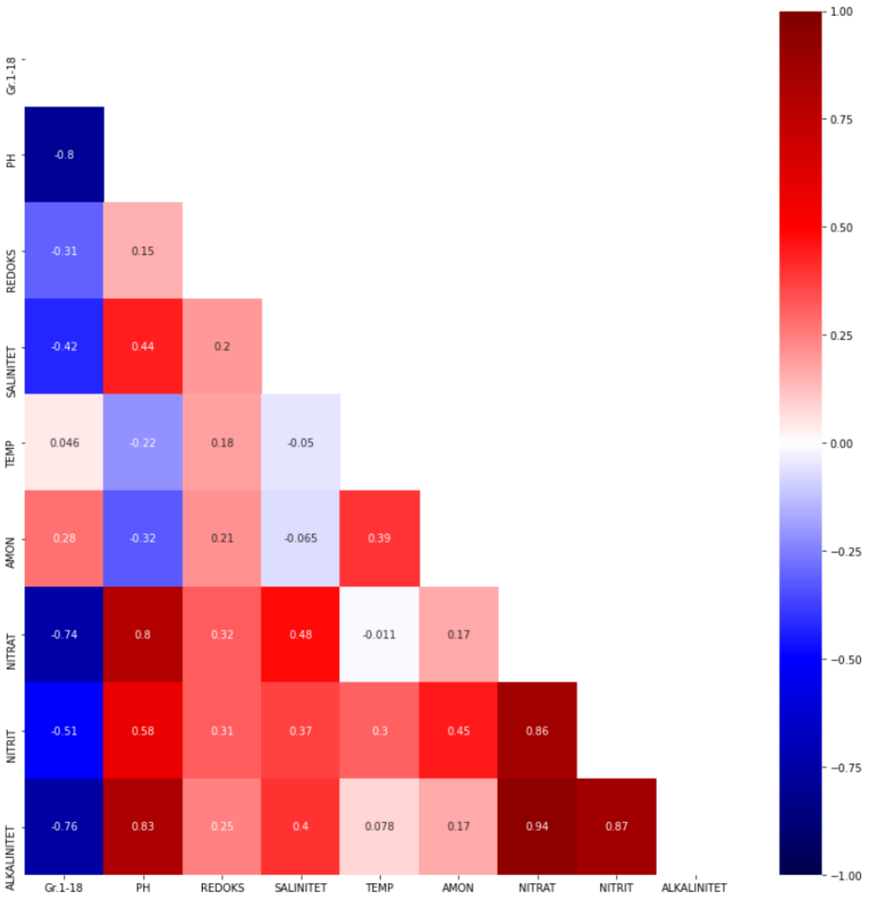

Figure 4.1 - Correlation matrix for the 1st group: Gr1-18

1st group (18 samples, with no values for CO2):

- Mortality shows strong negative correlations, i.e., when the parameter values increase, Mortality decreases, and vice versa, with pH, nitrate and alkalinity. The most robust inverse relationship appears in pH, with a correlation of -0.8. It is interesting to note that pH presents its highest correlations (and positive) with nitrate and alkalinity, with values of 0.8 and 0.83, respectively.

The lowest correlation is with temperature, which is justified because our data did not include values outside the range necessary for the species' survival. Although nitrite is not the parameter that most influences Mortality, it shows high correlation values with nitrate and alkalinity, suggesting an indirect influence on the target variable.

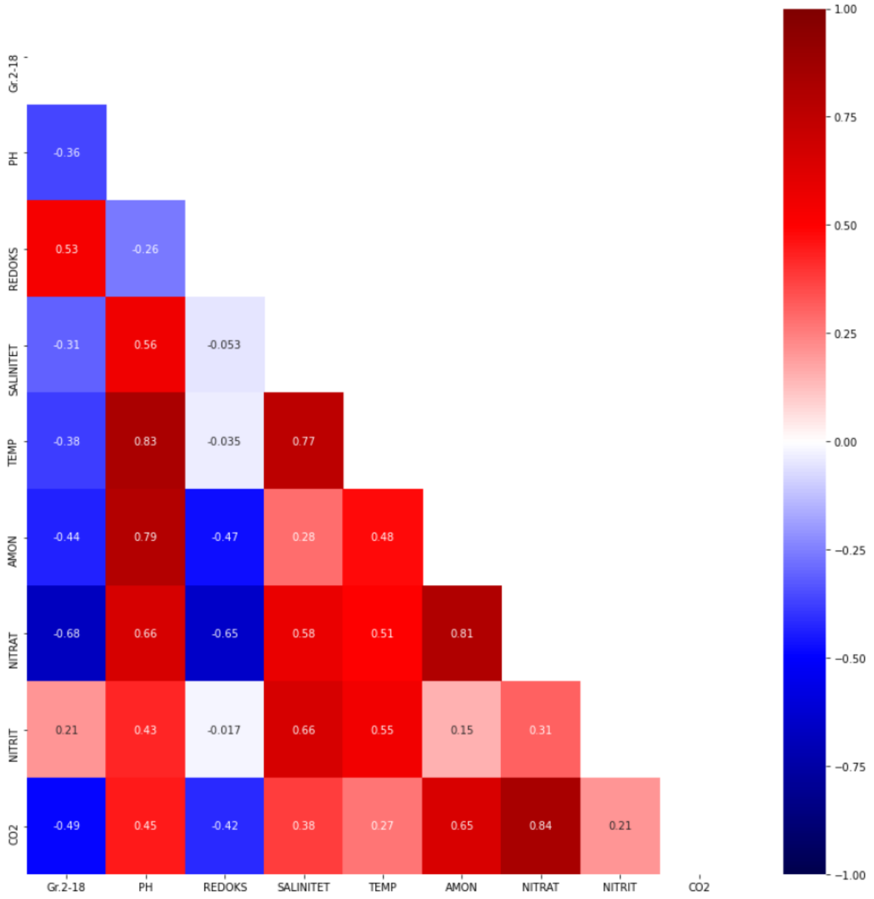

Figure 4.2 - Correlation matrix for the 2nd group: Gr2-18

2nd group (12 samples, with no values for alkalinity):

- In the second group of fish, Mortality shows the highest negative correlation with nitrate variable and the highest positive correlation with the redox variable, although it is a moderate correlation. However, interestingly, pH is the parameter that shows the highest correlation with all variables except redox, with which it shows a very low negative correlation of -0.26, contrasting with its influence on ammonium and temperature, with high positive correlations, close to 0.8. Thus, despite the weak direct correlation with Mortality, the pH significantly influences the other variables, showing positive correlations of some relevance. Concerning redox, its correlations are only moderate in nitrate and ammonium, all the others are lower. Additionally, the high positive correlations of nitrate with ammonium and CO2 stand out, putting this parameter in the spotlight in this analysis.

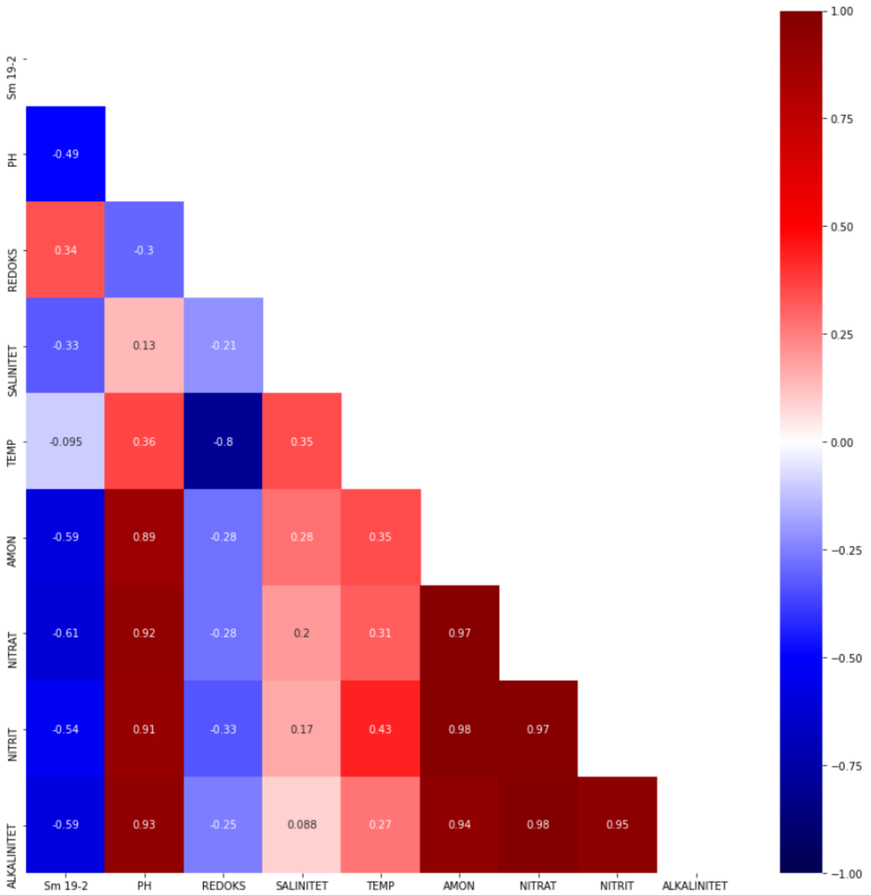

Figure 4.3 - Correlation matrix for the 3rd group: Sm19-2

3rd group (14 samples, with no values for CO2):

- The 3rd group shows mostly moderate negative correlations of Mortality with all variables except temperature, which is very low (-0.095). The very high positive correlations of ammonium, nitrate, nitrite and alkalinity, among themselves, close to 1, stand out. It is interesting to note that these variables are the ones that also stand out in their relationship with Mortality in this graph, although with only moderate and negative correlations. Again, we can see that although the pH does not have the highest correlation with the target variable and is negative, it has a significant influence on the other variables: with nitrate, nitrite and alkalinity, it has very high and positive correlations, and ammonium is very close to these values as well; this fact is relevant since these are the parameters that show the highest correlation with Mortality.

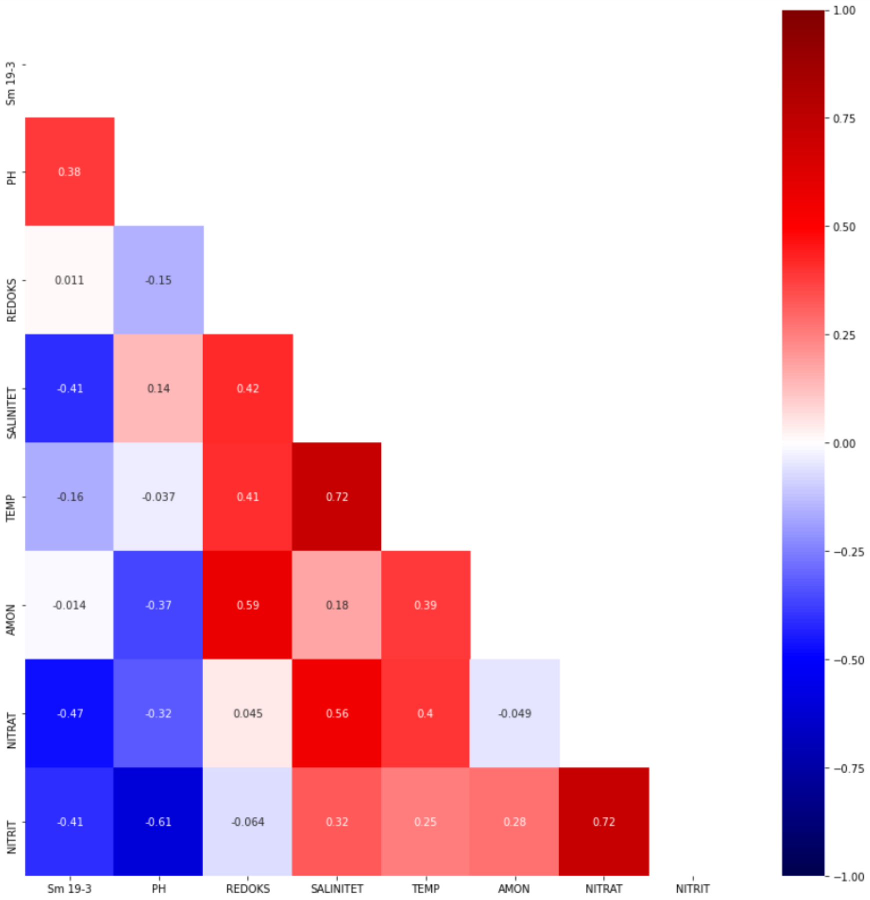

Figure 4.4 - Correlation matrix for the 4th group: Sm19-3

4th group (12 samples, with no values for CO2 and alkalinity):

- In the 4th group, the correlations are less impressive than in the previous groups, with Mortality showing the highest negative correlation, and only low, with nitrate (-0.47), followed by nitrite and salinity (-0.41), and a positive correlation with pH of 0.38. Divergent from the previous analyses is the fact that the redox and ammonium variables show very irrelevant correlations with the target variable. The connection of the nitrate and nitrite parameters is evidenced here, showing the highest positive correlation on the graph, 0.72 -a moderate one. It becomes clear that these two variables are dependent on each other, as well as being constantly relevant to Mortality.

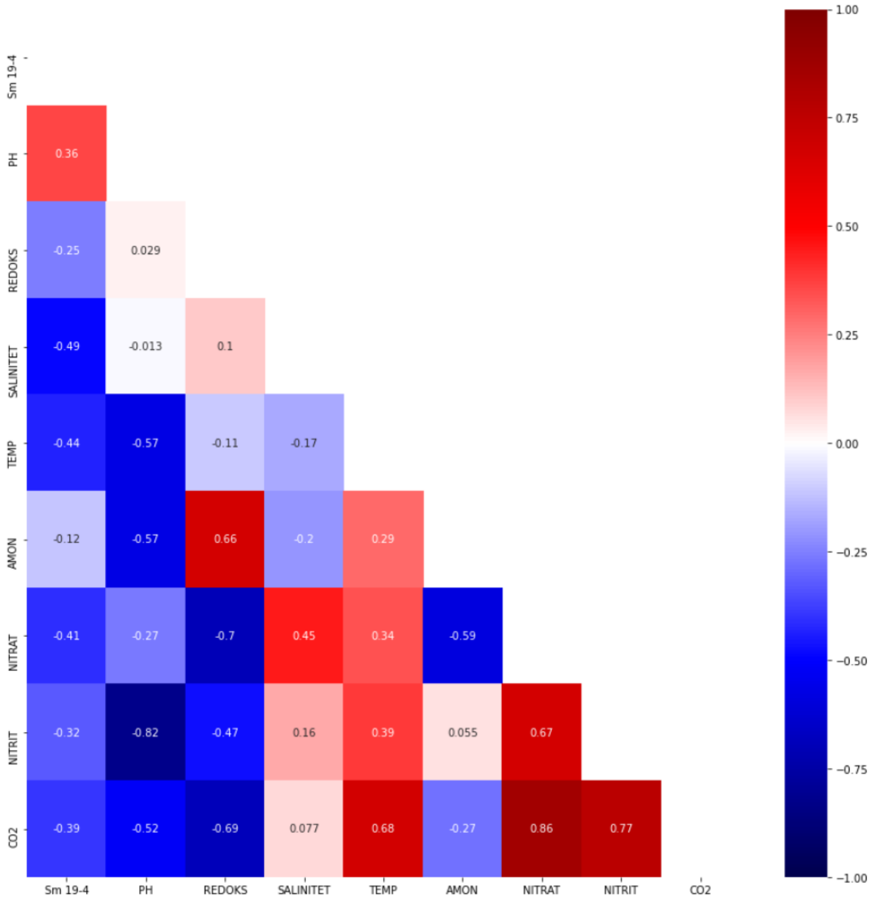

Figure 4.5 - Correlation matrix for the 5th group: Sm19-4

5th group (8 samples, with no values for alkalinity):

- For the 5th group, the variable target has the highest negative correlations with salinity, temperature and nitrate, and they are all low; Both ammonium and redox have a very low negative correlation. The only parameter positively correlated with Mortality is pH, with only 0.36. As for its relationship with other parameters, pH shows significant negative correlations: it has a high correlation with nitrate, and moderate correlations with temperature, ammonium, and CO2. Once more, we are inclined to consider pH as a relevant variable because of its indirect relations with the parameters that correlate most with the target variable.

  The graph also reveals that most correlations are negative, except for nitrate, which shows high positive correlation values with CO2. As far as salinity is concerned, it only shows some correlation with nitrate, a low correlation of 0.45.

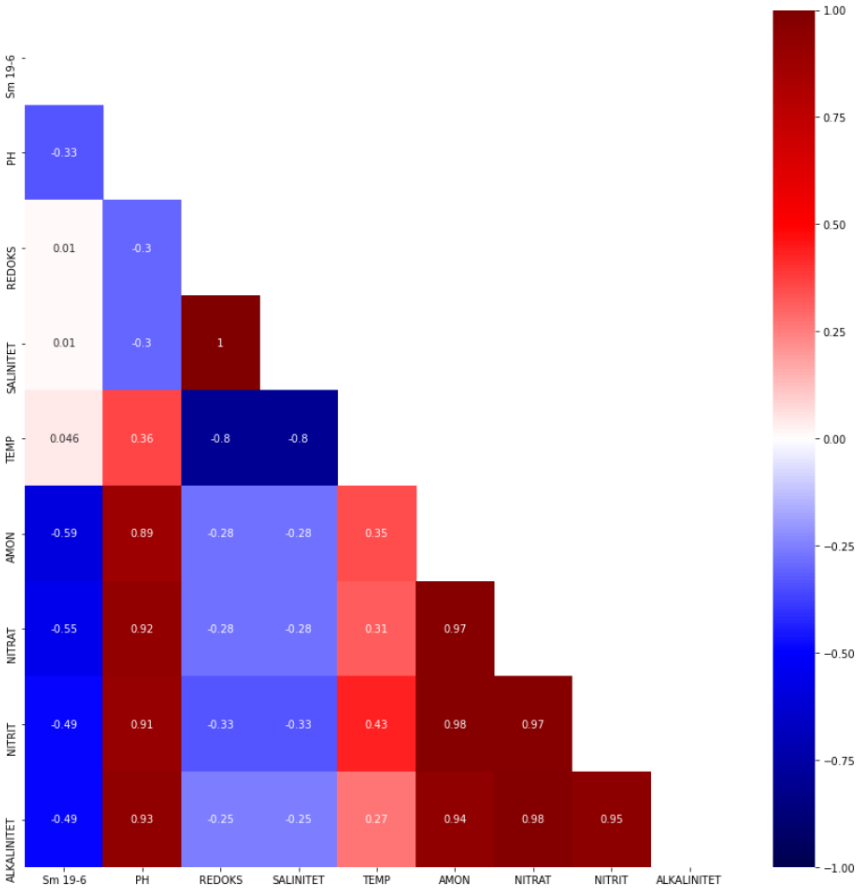

Figure 4.6 - Correlation matrix for the 6th group: Sm19-6

6th group (8 samples, with no values for CO2):

- This graph stands out by presenting stronger colours, i.e., it shows excellent correlations. However, regarding Mortality, there are three parameters with insignificant correlations -redox, salinity and temperature- and the strongest correlations are negative and only moderate. Again, ammonium and nitrate are the essential correlations of the target variable, although only moderate, followed very closely by nitrate and alkalinity. When analyzing the indirect relationships, we see that the variables with no relevance to Mortality have high negative correlations with each other. In the opposite direction, the variables that most impact the variable target show very high positive correlations with each other, close to 1. Although the pH shows a low negative correlation with Mortality, we again observe solid positive relationships of this parameter to the ones that most influence the target variable. It is important to remember that this group comprises only 8 values, which may bias the results.

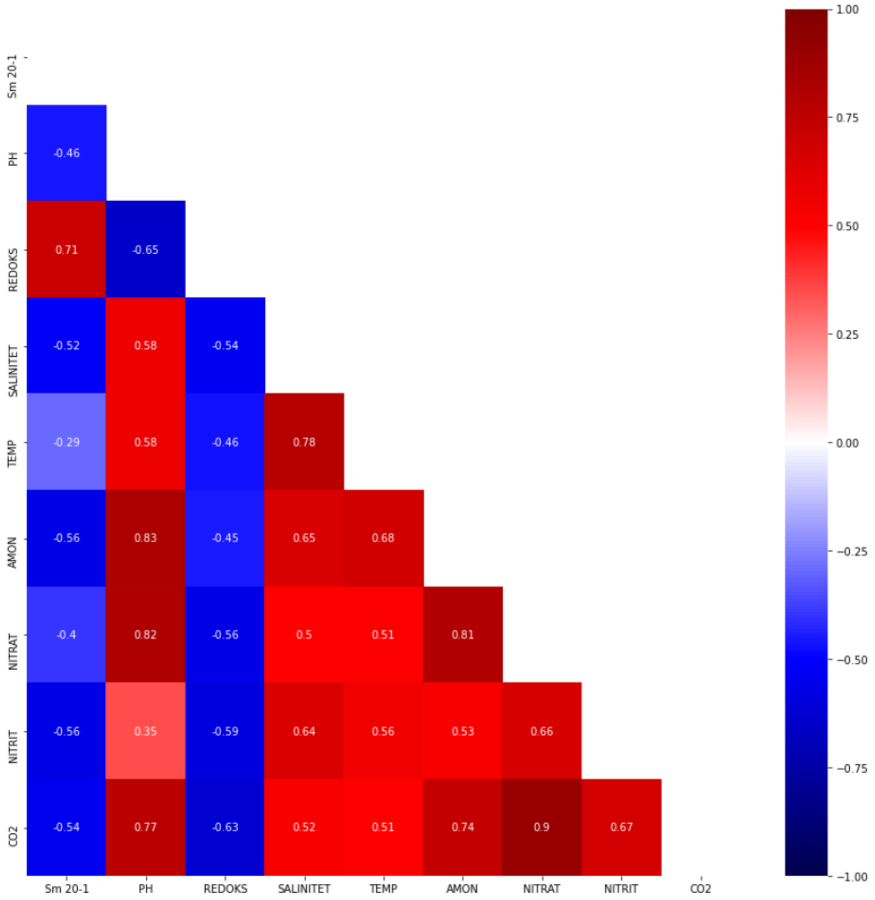

Figure 4.7 - Correlation matrix for the 7th group: Sm20-1

7th group (13 samples, with no values for alkalinity):

- The 7th group differs from the previous ones by presenting a panorama of mostly positive correlations and by the fact that all variables show strong correlations with each other. The lowest correlation in the graph is between Mortality and temperature, with a negative correlation of -0.29. The strongest negative correlations with Mortality are ammonium and nitrate, followed closely by pH and CO2. The highest positive correlation of the target variable is 0.71 with redox; this variable shows moderate negative correlations with all the others, mostly moderate ones. pH, in this analysis, stands out again as the parameter with the highest (and positive) correlations with all other variables except redox. CO2 also presents strong positive correlations with other variables, with particular emphasis on the very high correlations with nitrate.

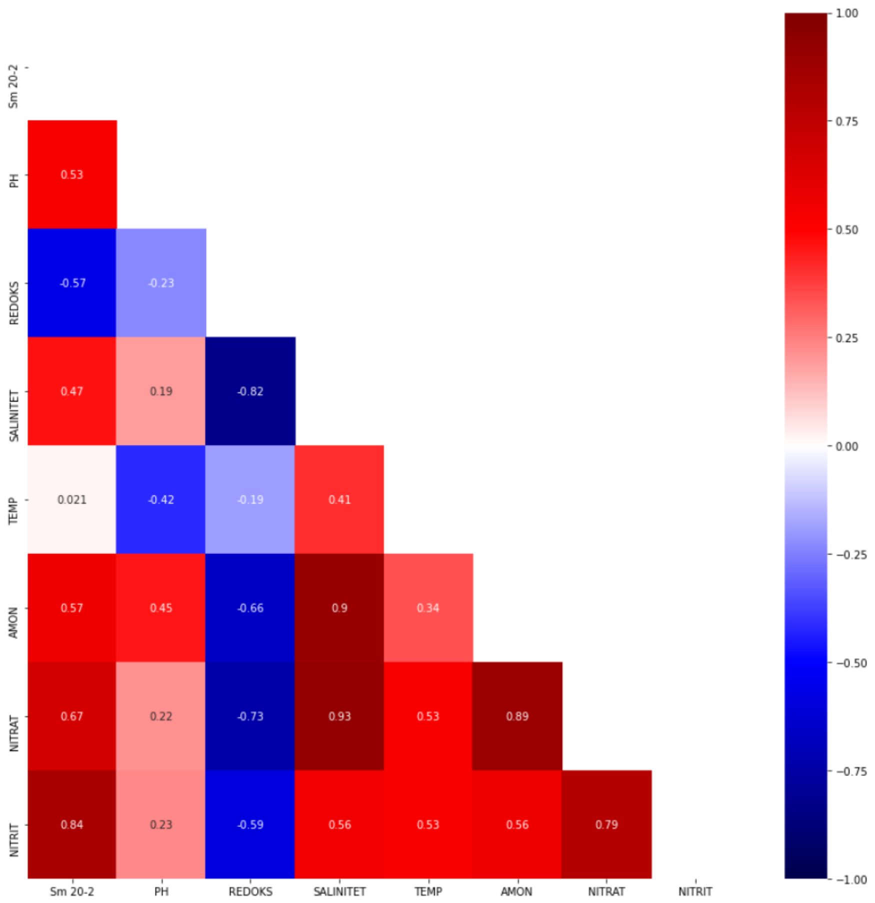

Figure 4.8 - Correlation matrix for the 8th group: Sm20-2

8th group (7 samples, with no values for CO2 and alkalinity):

- In the 8th group, temperature again shows almost zero correlation with the target variable. The highest Mortality correlation with other parameters is positive and high, with nitrite (0.84). Redox is the only negative correlation for mortality, and a low one; this variable also shows negative correlations with all the others, with salinity standing out at -0.82. Once again there is an emphasis on the positive correlations of Nitrate with other variables, especially with salinity (0.93), ammonium (0.89) and nitrite (0.79). It also shows a moderate negative correlation with the redox (-0.73). pH and temperature show the most insignificant correlations in all directions, against nitrate and nitrite, that collect the highest values. This group presents the smallest number of analyzed values, a fact that our interpretations will consider.

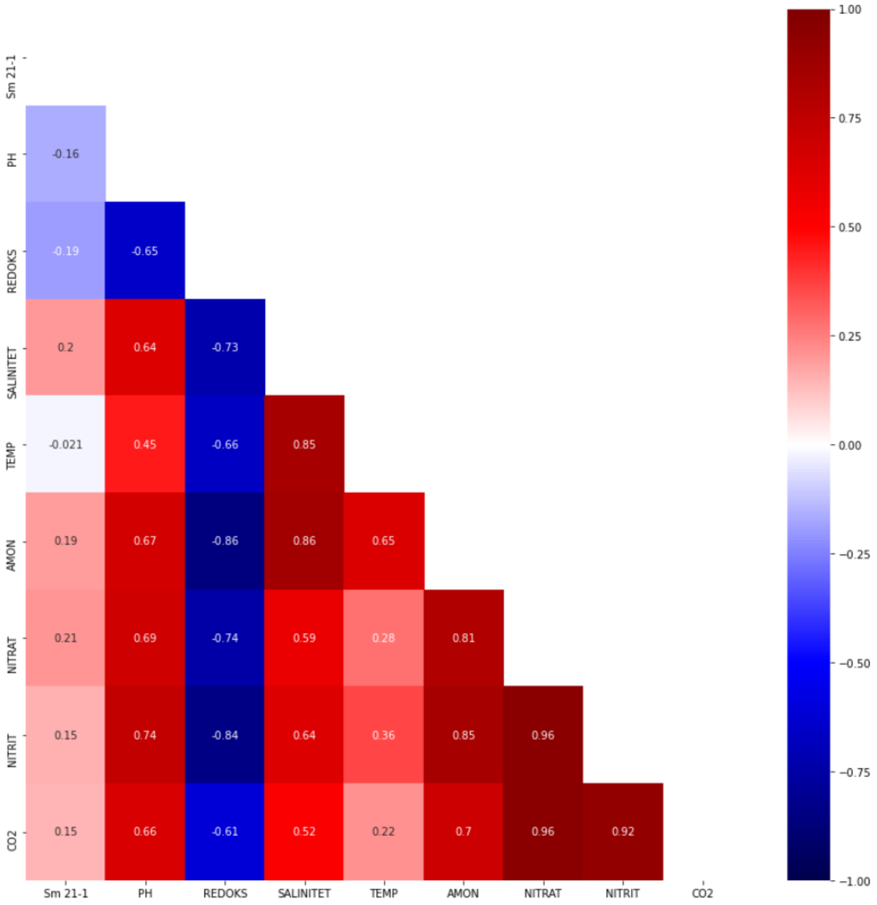

Figure 4.9 - Correlation matrix for the 9th group: Sm21-1

9th group (12 samples, with no values for alkalinity):

- The 9th group shows the most insignificant correlations between the target variable and all the parameters, compared to the other groups. No parameter presents correlations higher than |0.21|, which are very low. It should be noted, however, that the highest correlations of the variable target, in this chart are with nitrate, ammonium and salinity. Regarding the correlations of the other parameters, is visible that redox has the only negatives correlations in the graph, all of them moderate or high. The biggest positive correlations are positive, between nitrate, CO_2 and nitrite - all very high-, and we also have high ones between these variables and ammonium. The high positive correlation of salinity with temperature and ammonium is also noted.

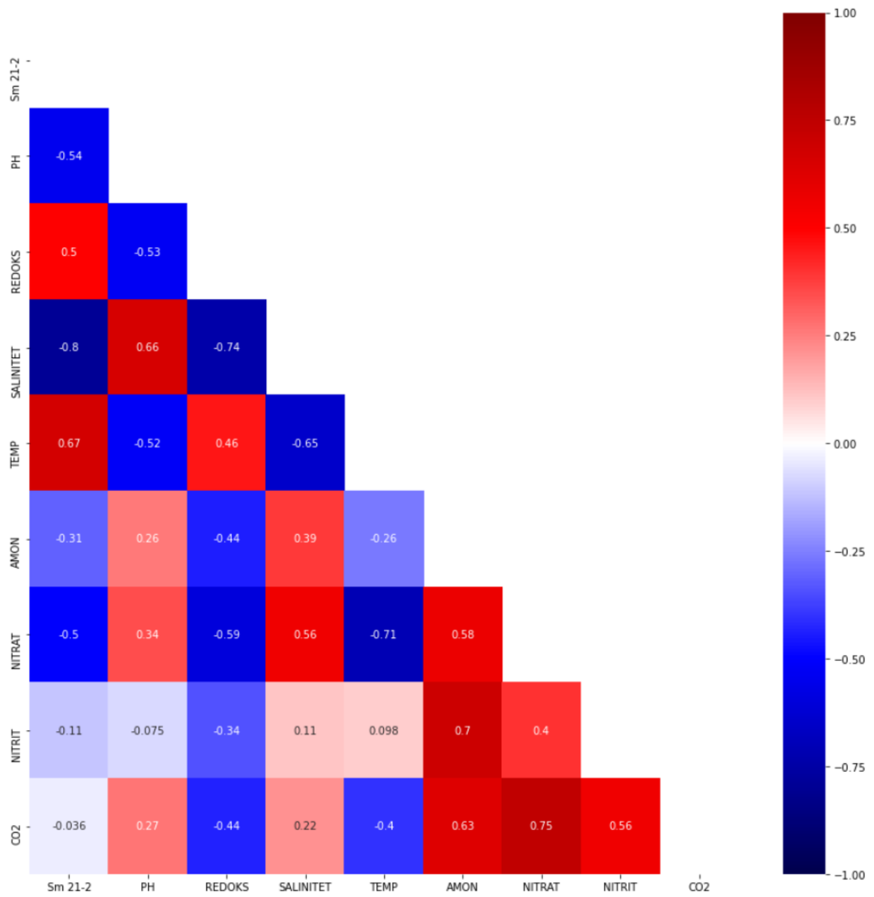

Figure 4.10 - Correlation matrix for the 10th group: Sm21-2

10th group (18 samples, with no values for alkalinity):

- The 10th group scenario contrasts a little with the previous ones, showing the weak correlations of nitrite with all parameters. Regarding the target variable, the most significant relationship is with salinity, with high negative correlation of -0.8, and, on the opposite side, the largest positive correlation is only moderate, 0.67, with temperature. In the big picture, salinity and redox are the parameters with the highest correlations, contrasting with the previous higher results for nitrate, which in this graph shows only moderate correlations. Nevertheless, ammonium, nitrate, nitrite and CO2 continue to show positive and moderate correlations with each other.

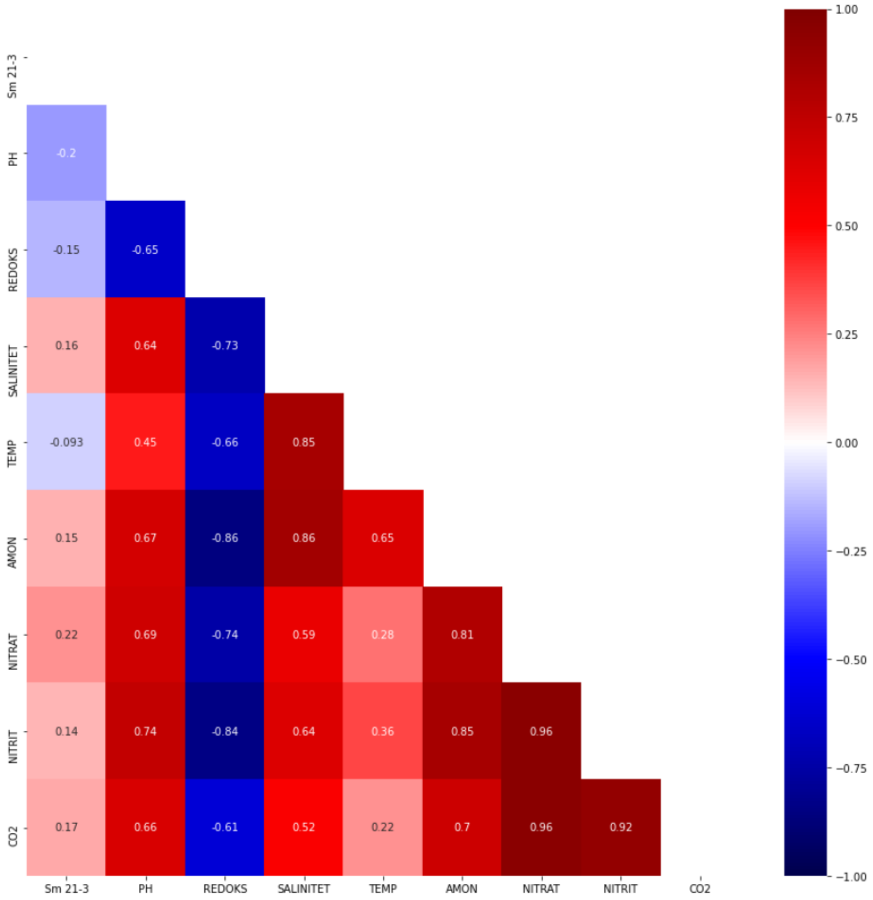

Figure 4.11 - Correlation matrix for the 11th group: Sm21-3

11th group (12 samples, with no values for alkalinity):

- The 11th group again exhibits weak correlations between Mortality and all the other parameters, being the one with the highest correlation nitrate, which shows a low positive of 0.22. It is noteworthy that redox is the only variable to show negative and significant correlations with all the others. There are patterns that start to be felt in the analyses, as is the case of temperature, which again presents the most insignificant correlation with the target variable in the graph. Overall, the highest positive correlations are between nitrate, nitrite and CO2, all above 0.92. Also, ammonium shows a high positive correlation with nitrate, nitrite and CO2, and a high negative one with redox. We point out, moreover, the high positive correlation of salinity with temperature and ammonium, something already verified in the analysis of the 9th group.

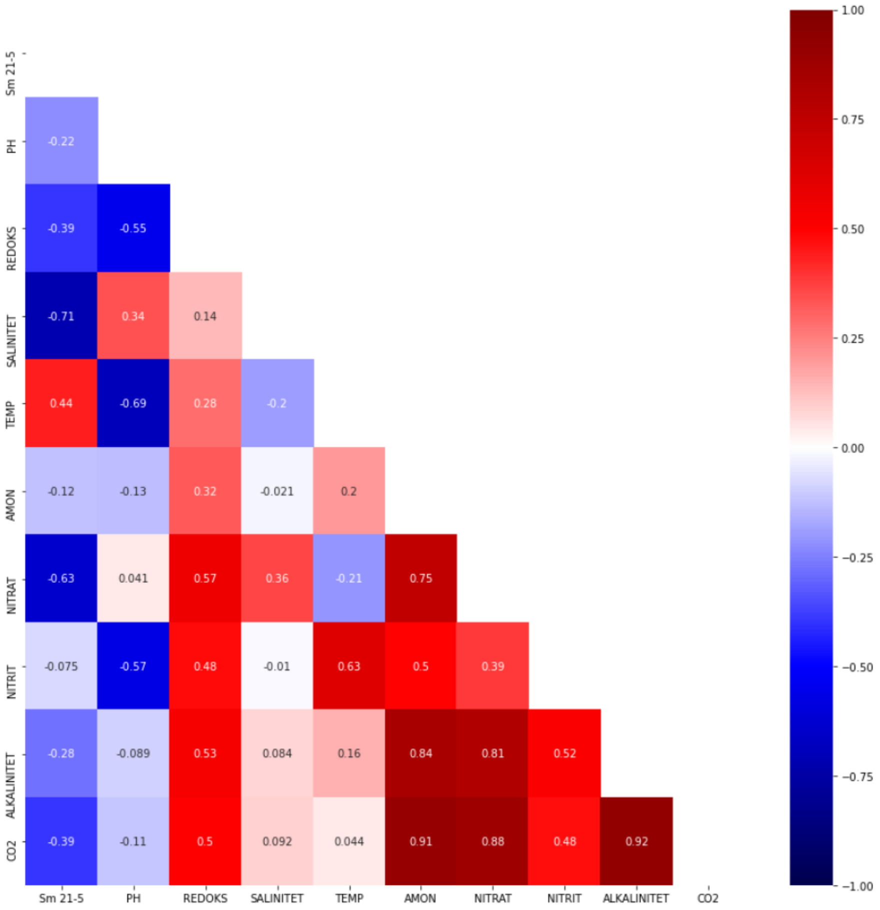

Figure 4.12 - Correlation matrix for the 12th group: Sm21-5

12th group (13 samples):

- In the analysis of this group, we see that Mortality has the highest correlation with salinity, a negative one of 0.71, and its only positive correlation is with temperature, with 0.44. Nitrite stands out in this graph for presenting the most insignificant correlation with the target variable, which did not happen in the previous analyses. Regarding salinity, apart from Mortality, it shows only weak correlations, being the best the moderates ones with nitrate and alkalinity. Overall, it stands out the positive correlations between alkalinity and CO2, and between these parameters and ammonium and nitrate - all above 0.75.

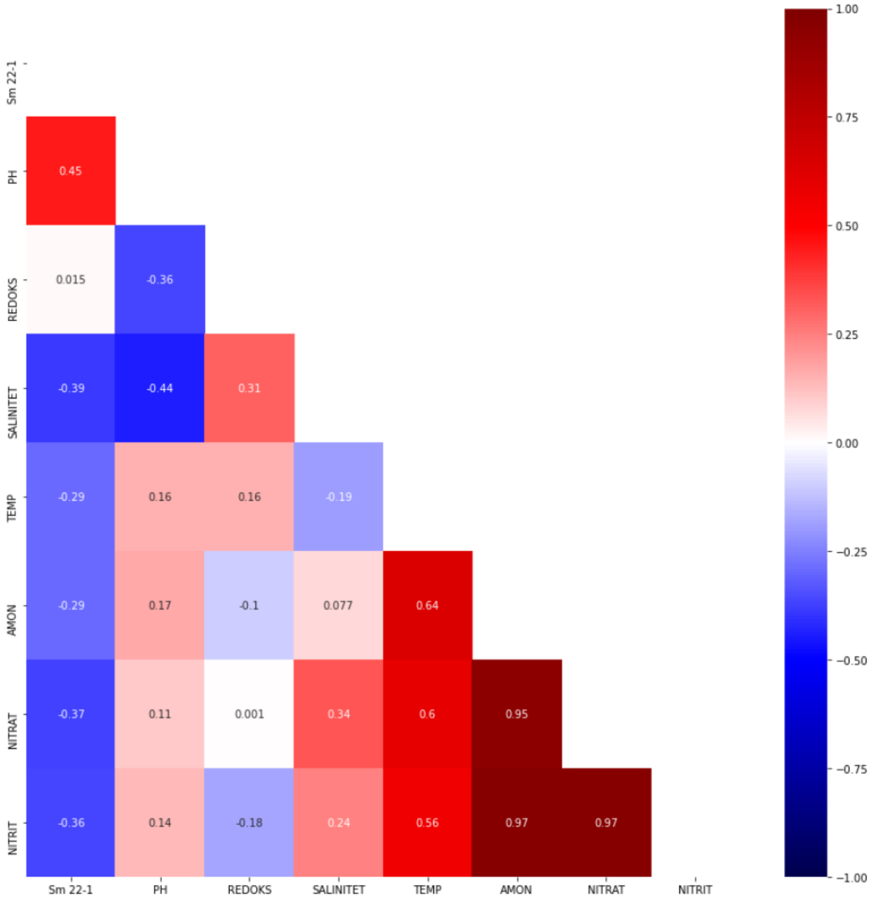

Figure 4.13 - Correlation matrix for the 13th group: Sm22-1

13th group (15 samples, with no values for CO2 and alkalinity):

- This group is the only one where it was impossible to use the parameters CO2 and alkalinity because they needed more values for analysis. The variable target presents its highest correlation with pH, with a value of 0.44. Salinity, nitrate and nitrite stand out in the negative correlations, although they show only moderate correlations. The general panorama presents very pastel colours, showing the weak correlations obtained here. Even so, temperature appears as a variable with moderate positive correlations with ammonium, nitrate and nitrite; and once again is reinforced the dependence of these three parameters, which present the highest correlations of the analysis, all close to 1.

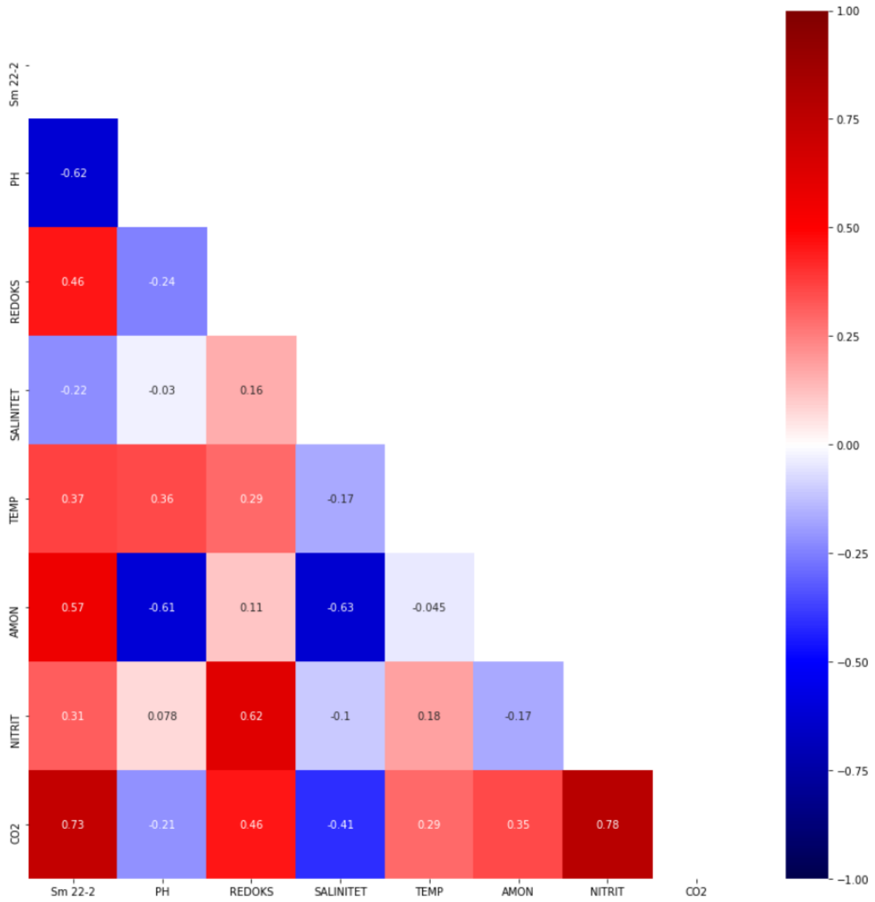

Figure 4.14 - Correlation matrix for the 14th group: Sm22-2

14th group (8 samples, with no values for nitrate and CO2):

- The last group analyzed presents a more pastel colour correlation matrix, that is, with not very strong correlations, which is not surprising, since it is one of the ones with the fewest samples. For the variable target, the biggest trend are the positive correlations, showing a high one with CO2 (0.73). In the opposite field, the highest negative correlation is with pH, with a value of -0.62. CO2 reveals itself as the variable with the highest and best correlations in this graph, with its relationship with nitrite (0.78) again standing out. It also highlights the moderate negative correlations of ammonium with salinity and pH.

In this first interpretation of results, it is important to emphasize that not all groups contain the same variables, which justifies some discrepancies found. The following deductions can be taken into account:

- nitrate is the variable that presents the highest correlations with Mortality;
- temperature is the variable that presents the smallest correlations with Mortality;

- nitrate and nitrite are the variables with the highest correlation with each other, and it is verified that ammonium also shows a high affinity with these;
- CO2 and alkalinity, when they have enough values to be considered, show significant correlations with the variable target, nitrate and nitrite;
- redox and pH have an inverse tendency to nitrate and nitrite concerning their correlation with Mortality.

4.2 Analysis by variable

To gain insights into the impact of each parameter on the target variable, the analysis was made individually.

Figures 4.15-4.22 present the average of all target variables and their distribution among the 14 groups studied.

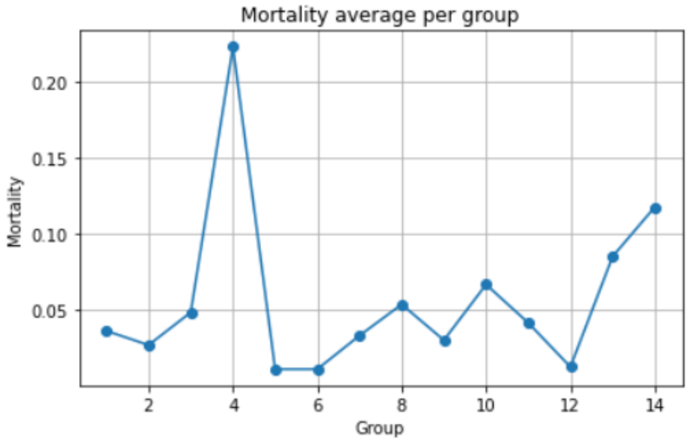

Figure 4.15 -Average mortality per group

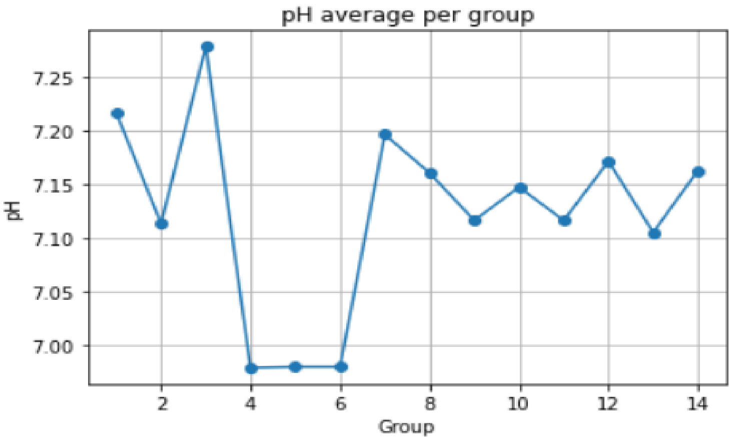

Figure 4.16 - Average pH per group

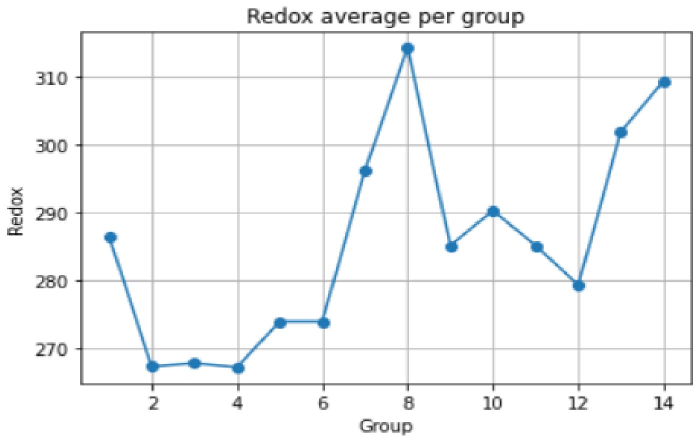

Figure 4.17 - Average redox per group

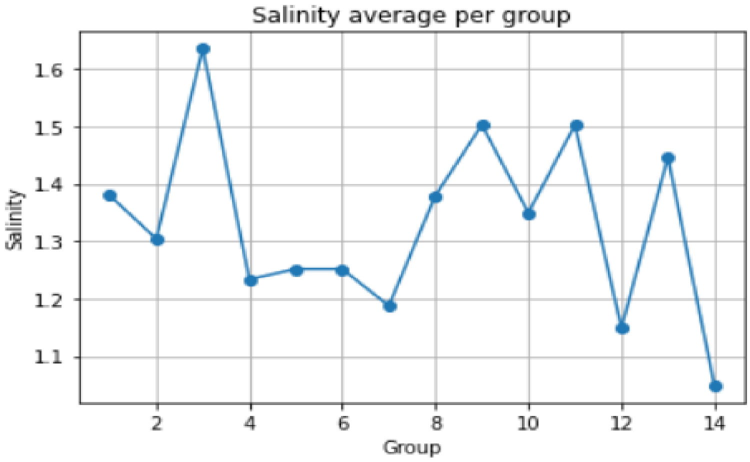

Figure 4.18 - Average salinity per group

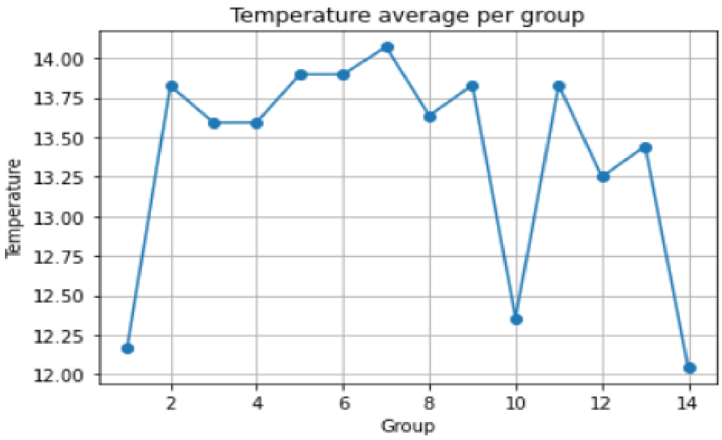

Figure 4.19 - Average temperature per group

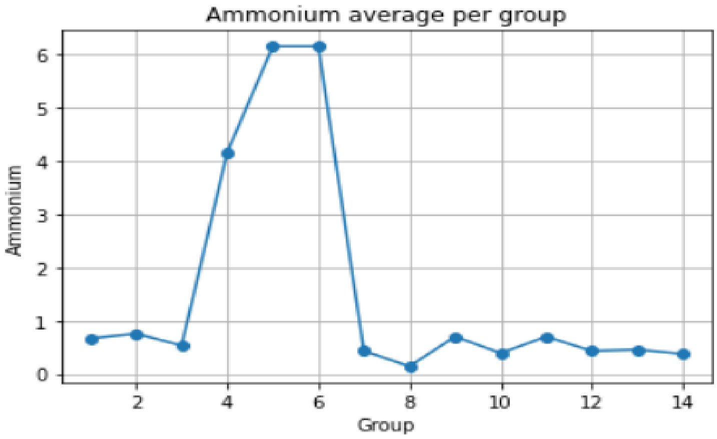

Figure 4.20 - Average ammonium per group

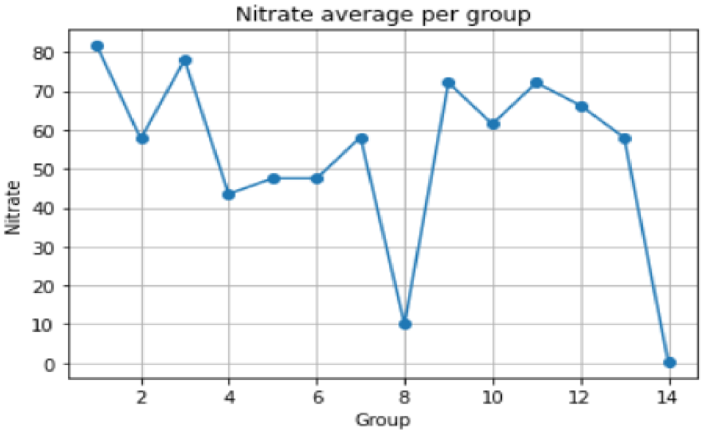

Figure 4.21 - Average nitrate per group

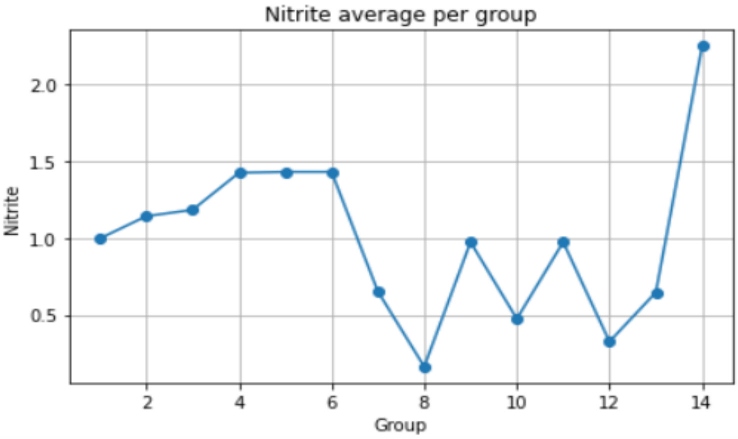

Figure 4.22 - Average nitrite per group

The analysis of the Mortality revealed interesting insights; Among the variables examined, redox (Figure 4.17) displayed a similar positive trend across groups 6 to 14, indicating that higher redox values correspond to increased Mortality. Additionally, pH (Figure 4.16) demonstrated a similar positive association within groups 8 and 11, while temperature (Figure 4.19) and salinity (Figure 4.18) exhibited a negative relationship within the same groups.

The investigation unveiled notable patterns regarding nitrate and nitrite (Figure 4.21 and 4.22). These two variables showcased a similar negative association across groups 8 to 14 (excluding group 12), suggesting that higher nitrate and nitrite levels coincide with lower Mortality rates.

Moreover, the relationship between pH (Figure 4.16) and ammonium (Figure 4.20) displayed an inverse correlation. As pH levels increased, ammonium levels decreased, and vice versa. This finding highlights a correlation, and the interdependence between pH and ammonium concentrations.

Another observation was observed within groups 8 to 12, where temperature, salinity, nitrate, and nitrite demonstrated a synchronized behaviour. These variables exhibited similar trends, with one variable increasing or decreasing alongside the others. Conversely, pH displayed an inverse relationship with these variables, indicating that the other variables decreased as pH increased.

Overall, these findings provide valuable insights into the relationships among the variables and their impact on mortality rates. Studying the variables individually provides validation for some of the conclusions mentioned in the analysis by group, reinforcing the following points:

- nitrate and nitrite are the variables with the highest correlation with each other, and it is verified that ammonium also shows a high affinity with these;
- redox and pH have an inverse tendency to nitrate and nitrite concerning their correlation with Mortality.

However, it is essential to note that further analysis is required to establish definitive conclusions, as the interpretations are based on a limited dataset. Nevertheless, these findings contribute to a deeper understanding of the factors influencing mortality rates within the studied context.

To attempt to validate these inferences, the correlations of each variable with the average Mortality, by group, are shown in Figures 4.23-4.29.

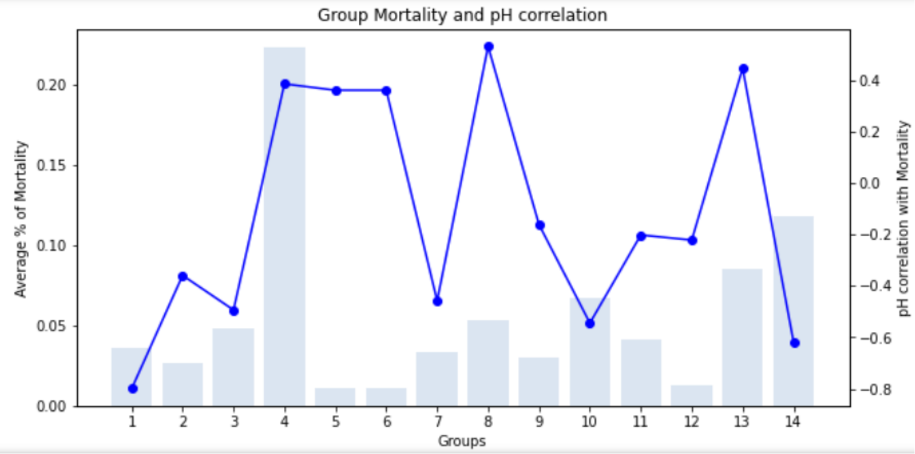

Figure 4.23 - Average percentage of Mortality vs pH correlation with Mortality, per group

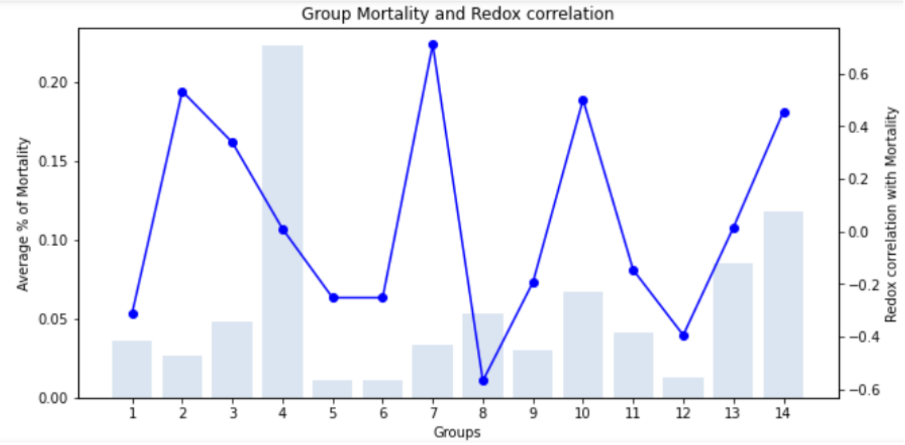

Figure 4.24 - Average percentage of Mortality vs redox correlation with Mortality, per group

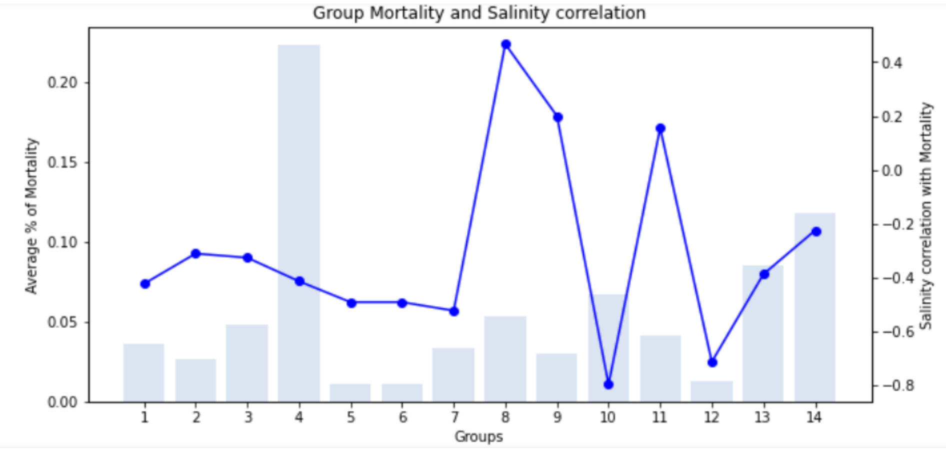

Figure 4.25 - Average percentage of Mortality vs salinity correlation with Mortality, per group

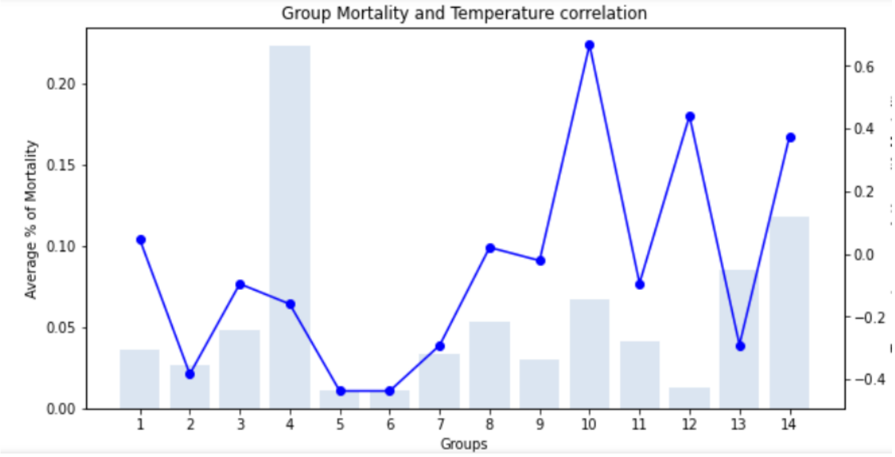

Figure 4.26 - Average percentage of Mortality vs temperature correlation with Mortality, per group

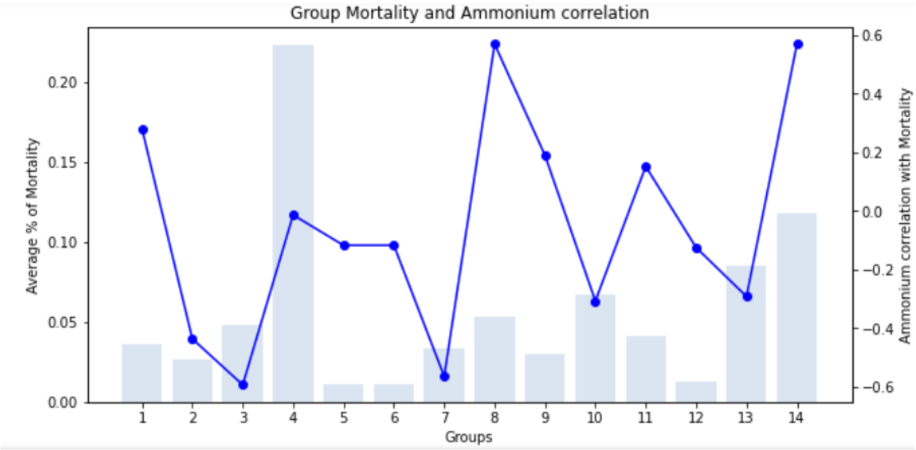

Figure 4.27 - Average percentage of Mortality vs ammonium correlation with Mortality, per group

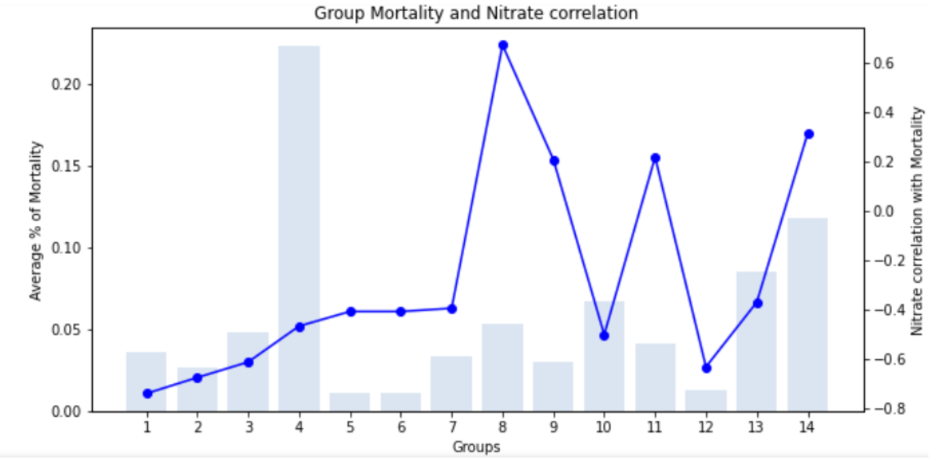

Figure 4.28 - Average percentage of Mortality vs nitrate correlation with Mortality, per group.

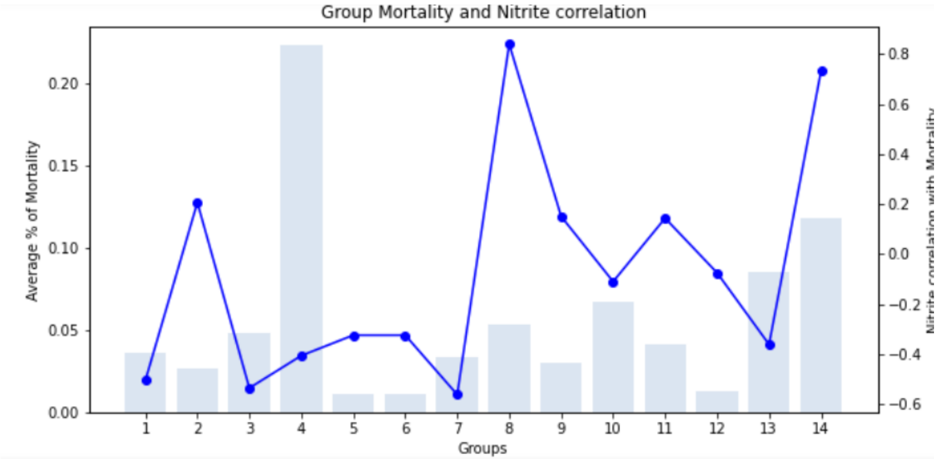

Figure 4.29 - Average percentage of Mortality vs nitrite correlation with Mortality, per group

The thorough interpretation of the 7 correlation graphs depicting parameter relationships with mortality in each group proves challenging and complex. The limited size of the analyzed dataset hinders the establishment of logical and consistently validated conclusions. In this regard, it is observed that the two groups with the highest mortality rates (groups 4 and 14) exhibit inverse correlations in 4 out of the 7 parameters: while some parameters show positive correlations in group 4, they also display strong negative correlations in group 14.

It is interesting that groups 5 and 6, which have the same quantity of data and identical mortality rates, also demonstrate similar correlations among the parameters. Notably, negative correlations stand out for salinity (Figure 4.25), temperature (Figure 4.26), and nitrate (Figure 4.28), while pH (Figure 4.23), exhibits a positive correlation. It is essential to mention that despite this consistency, these correlations are relatively low, around 0.4.

Regarding indirect relationships between the parameters, noticeable correlations between nitrate (Figure 4.28), nitrite (Figure 4.29), and ammonium (Figure 4.27) are consistently observed in the same direction. These correlations indicate a consistent interdependency between these parameters.

Through this analysis, 3 conclusions should be considered:

- High similarity between the behaviour of nitrate and salinity;
- High similarity between the behaviour of nitrite and ammonium;
- Redox shows opposite behaviour regarding nitrate between the groups 2 until 11.

It is important to emphasize that due to the limited size of the dataset, caution should be exercised in drawing definitive conclusions, and the results obtained in the various matrices do not reveal a perfect picture of any outputs.

4.3 Global Analysis

To obtain more substantiated results, a global analysis of the data is conducted. At this stage, the values for CO2 and alkalinity are excluded as they only appear in some groups of fish, and even the available values have undergone several imputations. This analysis consists of 170 data points with weekly values for 8 variables: the target variable - Mortality -, pH, redox, salinity, temperature, ammonium, nitrate, and nitrite. In Figure 4.30, a correlation matrix is created to analyze the overall correlation between the parameters.

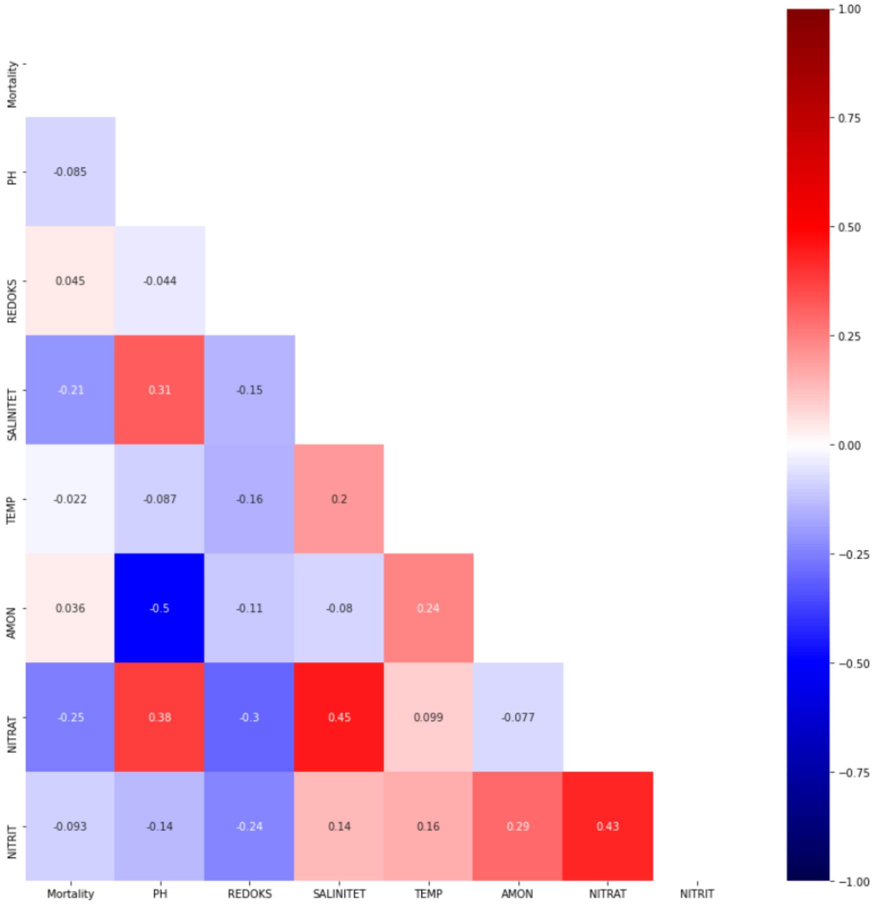

Figure 4.30 – Global correlation matrix

The group analysis highlighted the fact that the number of samples differed from each other. In this sense, we consider balancing the data for the global analysis. Data balancing is crucial for accurate and unbiased analysis. Imbalanced data, where one class significantly outweighs the others, can lead to biased models and misleading results. By balancing the data, it is ensured that each class is represented fairly. Proper data balancing helps mitigate the risk of skewed outcomes and enables more robust insights and decision-making based on a comprehensive understanding of the data distribution. It is crucial to ponder the implications of balancing the data set's characteristics and the analysis results. The choice between undersampling and oversampling depends on the problem's specific context. In this sense, the size of the data set was used as a criterion. If an extensive data set is available and the difference between classes is not significant, oversampling may be considered as a viable option to increase the representation of the minority class.Inversely, when the data set is already relatively small, subsampling may be preferable to avoid excessive loss of information. In Figure 4.31 and 4.32 the correlations obtained through undersampling and oversampling are presented for comparison.

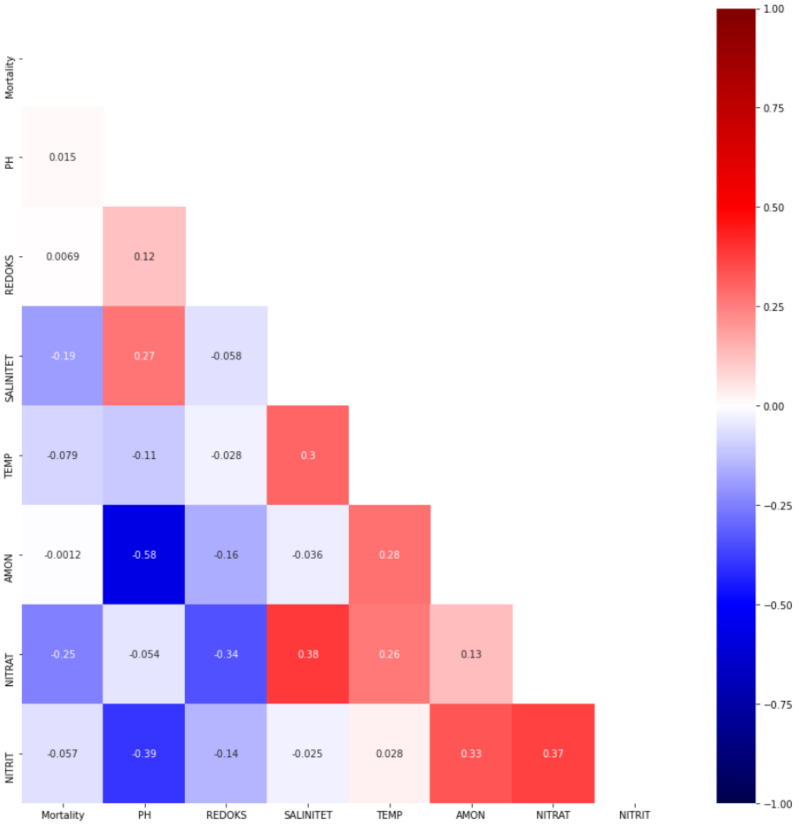

Figure 4.31 – Global correlation matrix with undersampling

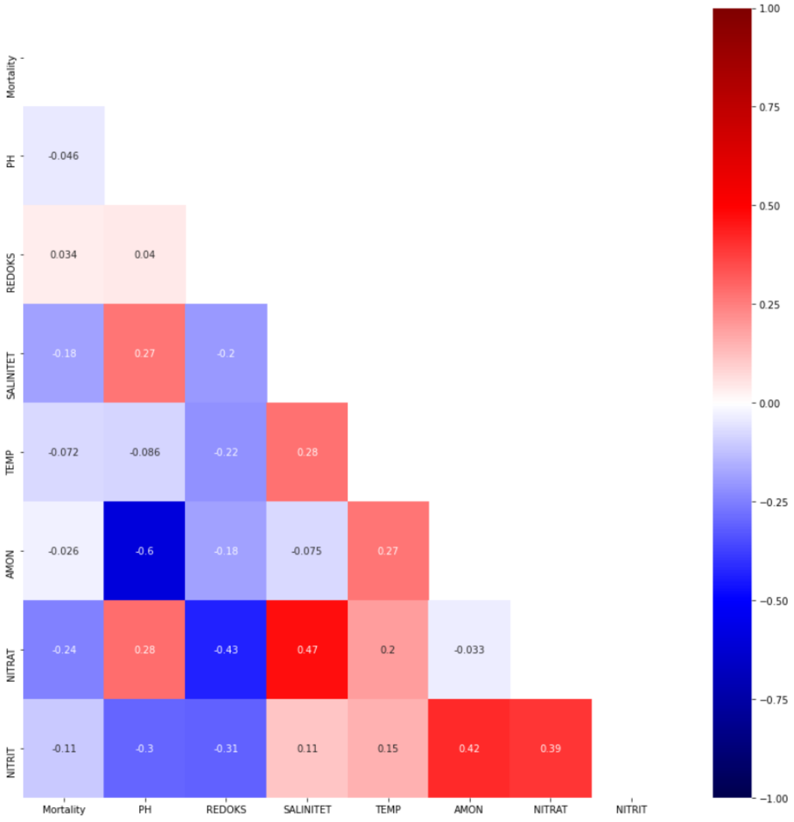

Figure 4.32 – Global correlation matrix with oversampling

For undersampling (Figure 4.31), based on the group with the lowest number of values, all groups were reduced to 7 values. On the other hand, for oversampling (Figure 4.32), all groups were equalised, aligning them with the group with the highest number of values, resulting in all groups having 18 values.

From the obtained results, there are no significant differences between the application of undersampling and oversampling. The limited sample size makes interpreting the results and drawing more definitive conclusions challenging. When comparing the original correlation graph with those developed using undersampling and oversampling techniques, it is observed that the colours and values undergo subtle changes.

The variable with the highest correlation to the target variable is nitrate, albeit with a very low negative correlation. This result is consistently observed in all matrices, both with and without data balancing. Following nitrate, salinity demonstrates a similarly low negative correlation, corresponding to the graphs' highest correlations. The least significant correlations are observed in ammonium, temperature, pH, and redox, all below |0.1|.

Regarding the relationships between the parameters, the correlation between pH and ammonium is notable, approximately -0.5, indicating a moderate correlation. Additionally, positive correlations above 0.41 are observed in all graphs between nitrate, salinity, and nitrite.

The outputs achieved in the global analysis align with the results of all the analyses, reinforcing the following conclusions:

- nitrate is the variable that presents the highest correlations with Mortality;
- temperature is the variable that presents the smallest correlations with Mortality;
- nitrate and nitrite are the variables with the highest correlation with each other, and they present similar behaviour with salinity and ammonium, respectively;
- redox and pH have an inverse tendency to nitrate and nitrite in what concerns their correlation with Mortality.

To highlight the proximity between nitrate and nitrite, and between pH and redox, and that they run in opposites ways, Figure 4.33 was created. Blue and green were chosen to present the pH and redox variables, and red and orange for nitrate and nitrite. The graphics present the average values of pH, redox, nitrate and nitrite values by group. It shows that between groups 4 and 6 there is no variation in values, and all variables show stable values. The most significant contrast in the behaviour of these parameters is evident between groups 7 and 13, where nitrate and nitrite follow and behave contrary to pH and redox.

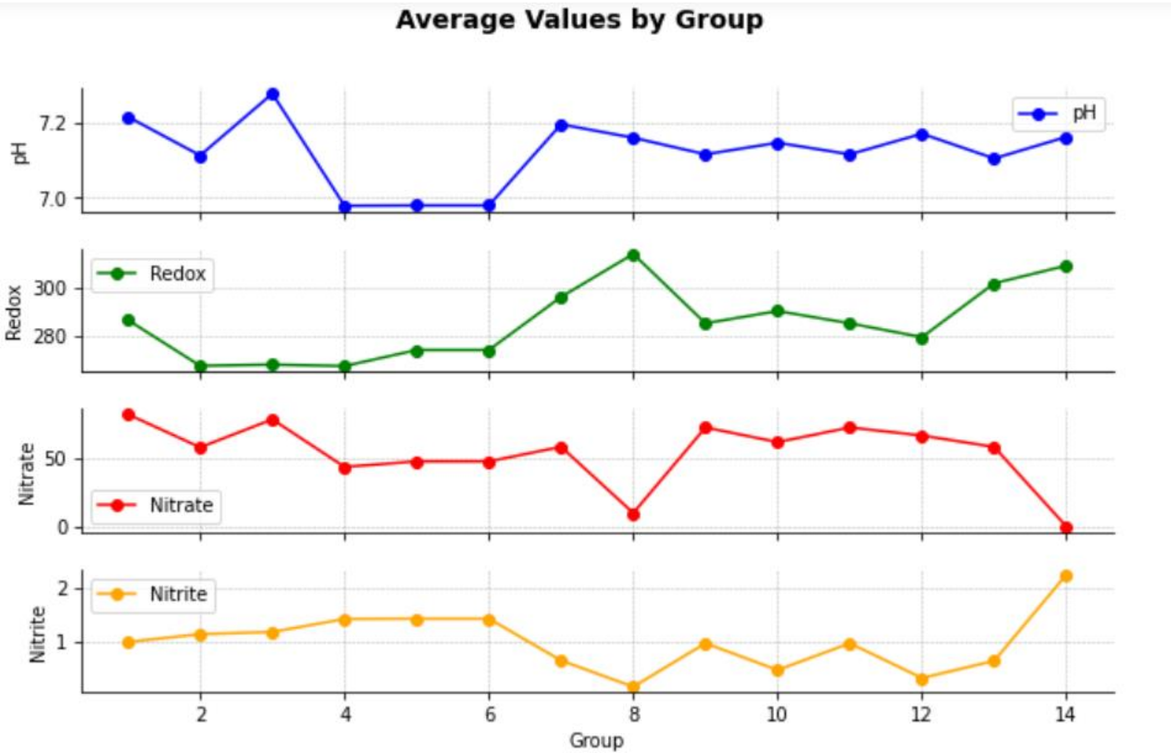

Figure 4.33 - Average values of pH, redox, nitrate and nitrite across all groups

It is important to acknowledge that although statistically significant, the observed correlations are generally low. Other factors and variables not considered in this analysis may contribute to the overall relationship. It is crucial to consider the limitations of these correlations and the need for further investigation to understand the complex dynamics within the studied context comprehensively.

CHARTER 5

Conclusion

The concluding chapter of this thesis presents a comprehensive summary of the findings and implications of the study on water quality management in aquaculture. Through an in-depth analysis of relevant literature and empirical research, this study has shed light on the crucial role of water quality parameters in influencing fish health and mortality rates. By critically examining the results and discussing their practical implications, this chapter aims to provide a concise overview of the contributions made by this research and its relevance to the field of study. Moreover, it outlines potential avenues for future research and highlights the significance of continued efforts to improve water quality management practices in achieving sustainable and profitable aquaculture production.

5.1 Discussion

The results obtained from the analysis largely aligned with the research objectives. These findings contribute to a deeper understanding of the 9 parameters of water quality and how they influence Mortality, within the context of fish farms. The results obtained add knowledge and pave the way for further exploration into the influence of water quality on fish mortality.

The cleaning and transformation of the data allowed us to understand the reality of the sample provided: the absence of data and its regularity forced data compression for a weekly period, the substitution of values and, in some groups, even disregarding some variables. However, with a clean dataset, it was possible to validate results consistently. This work also provides practical insights on the analyzed water parameters, as shown on the Insights charter: it was found that nitrate plays a very influential role in Mortality, with an inverse correlation to this variable; the pH and redox reveal a behaviour in the same direction, with a positive correlation tendency with Mortality, and the opposite occurs with nitrate and nitrite; temperature is a rather uninteresting factor in our results, which is understandable because fish farms develop their production in a controlled environment, so the temperature will have great control and will not expose significant variations; it is proven that nitrate and nitrite are closely related, growing and reducing similarly. Additionally, the variables CO_2 and alkalinity might also play a relevant role in Mortality, albeit at an indirect level. However, to validate this deduction, it is necessary to obtain more data

Identifying patterns and associations among the variables was achieved through rigorous data preparation and analysis. The findings provide valuable insights into the complex dynamics within the system and contribute to a deeper understanding of the factors influencing mortality rates. This research contributes to the field by pioneering the analysis of water quality parameters in a Norwegian fish farm, shedding light on their impact on fish health and mortality rates. While the study encountered data volume and consistency limitations, it serves as a foundation for future analyses that can delve deeper into the subject.

Despite the efforts made in this study, it is essential to acknowledge and discuss the limitations encountered during the research process. One significant limitation of this study is the limited volume and inconsistency of the data provided. The data available was sparse and lacked consistency, posing challenges in obtaining robust results. It is crucial to emphasize the need for investment in financial resources and time to ensure the accurate and error-free recording of all possible data, as sensor's calibration. The quality of the data directly influences the quality of the results obtained. By addressing the limitation of data quality, future analyses can be more focused and yield more insightful results. Recognizing the limitations encountered in terms of data volume and variability creates an opportunity to expand the scope of analysis. By collecting a larger volume of data and incorporating more variables, researchers can build upon this pioneering work and develop more robust models and frameworks for water quality management in aquaculture. Building a consistent and reliable database will enable the application of advanced algorithms and artificial intelligence techniques for forecasting and predictive modelling. This, in turn, can provide valuable insights into water quality management in aquaculture, allowing for more informed decision-making and proactive measures to ensure sustainable and profitable production. It is essential to recognize that achieving such a comprehensive database and implementing advanced analytical techniques will require a medium-term investment from the industry. This investment is crucial to maintain sustainable and profitable production in the long run by keeping pace with technological advancements and allocating resources to establish a robust data collection system. Therefore, collecting a maximum number of variables at regular intervals is recommended, establishing a coherent and comprehensive database that provides reliable information for subsequent analysis. With a more comprehensive dataset, future analyses can provide more accurate and detailed insights into the relationship between water quality parameters and fish health.

The study's findings highlight the importance of embracing technological advancements in the aquaculture industry. As aquaculture evolves, incorporating advanced analytics and artificial intelligence techniques becomes crucial. Applying algorithms and predictive modelling can enhance decision-making processes and enable forecast capabilities.

Comparing and analyzing the articles from Related Work with the core theme of this thesis, it is evident that water quality management is crucial for achieving sustainable aquaculture practices. The studies highlight the impact of various water quality parameters, such as temperature, salinity, DO, ammonia, and pH, on fish health and mortality rates. By understanding these relationships, researchers and industry practitioners can implement proactive measures to optimize water quality conditions. It proves no easy task to match the insights of these authors with our study since the focus of their analysis was on variables that we did not have access to, such as ammonia and DO; our study opens the way to research other variables that show a strong correlation with Mortality, such as nitrate. It is prudent not to disregard any analysis but rather to develop an examination with as many water parameters as possible.

In summary, this research holds practical implications for the aquaculture industry and related contexts. The findings can inform aquaculture practitioners and industry stakeholders about the significance of water quality management in maintaining fish health and mitigating mortalities. Understanding the relationships between water quality parameters and fish populations allows for developing tailored management strategies. The study's findings emphasize the need for more comprehensive analyses, including a broader range of variables. Industry can achieve sustainable and profitable production by embracing technological advancements and investing in medium-term initiatives. Our contribution to the field lies in paving the way for future research to explore a larger volume of data and variables, ultimately enhancing our understanding of water quality's impact on fish mortality rates and facilitating the development of tailored management practices. Ultimately, this research contributes to the advancement of knowledge in water quality management and supports the sustainable development of aquaculture as a vital source of food production.

5.2 Future Work

Building upon the findings and insights gained from this thesis, there are several potential opportunities for future research in the field of water quality management in aquaculture. The fish farm that provided the data allows its sharing, so the data used in this study may be requested and acquired. These future endeavors can further expand the knowledge and contribute to developing more effective and sustainable practices.

One promising direction for future research involves conducting studies with a larger volume of data and a wider range of variables. As mentioned previously, the limitations of this study in terms of data volume and variability highlight the need for a more comprehensive dataset. Researchers can enhance the accuracy and reliability of their analyses by collecting data from a more significant number of fish farms and incorporating additional variables, such as DO, which is referred to in several articles as a relevant parameter in water quality for maintaining marine life. This will enable a more robust examination of the relationships between water quality parameters and fish health and mortality rates.

In addition to expanding the dataset, future studies could focus on specific areas of interest within the field. For example, investigating the long-term effects of different water quality management strategies on fish populations would provide valuable insights into their efficacy and sustainability. The process could involve monitoring and analyzing the performance and health of fish populations over extended periods, considering factors such as growth rates, disease prevalence, and overall survival rates. The study of water recirculation is also relevant because it says much of the water quality the fish swam in, mainly when the difference between input and output is significant.

Another area for future exploration is the integration of advanced technologies and techniques into water quality management practices. As technology advances, there is a growing opportunity to leverage tools such as artificial intelligence, machine learning, and predictive modelling in aquaculture. By harnessing these technologies, researchers and industry practitioners can develop sophisticated algorithms and decision-support systems that optimize water quality conditions and improve fish health outcomes. Such advancements can facilitate real-time monitoring, early detection of anomalies, and timely intervention to mitigate potential risks.

Furthermore, it would be beneficial to investigate the economic implications of implementing improved water quality management practices. Assessing the cost-effectiveness and feasibility of different strategies can guide decision-making processes within the aquaculture industry. It is relevant to consider the initial investment required for technology adoption, operational costs, and potential returns regarding increased productivity, reduced mortalities, and improved market competitiveness.

Lastly, collaboration and knowledge-sharing among researchers, industry stakeholders, and policymakers are critical for advancing the field of water quality management in aquaculture. Establishing multidisciplinary research networks and platforms for exchanging information, best practices, and lessons learned can accelerate progress and facilitate the adoption of innovative approaches across different regions.

There is great potential for future research in the field of water quality management in aquaculture. By expanding the dataset, investigating specific areas of interest, integrating advanced technologies, evaluating economic implications, and fostering collaboration, researchers can enhance the understanding of the subject and develop practical and sustainable solutions to ensure the well-being and productivity of fish populations in aquaculture systems.

This thesis highlights the paramount importance of water quality management in aquaculture. It uncovers the intricate relationship between water quality parameters and fish health, emphasizing the need for comprehensive analyses and data-driven decision-making. Investing in technological advancements and fostering collaboration among researchers, industry stakeholders, and policymakers can pave the way for sustainable and profitable fish farming practices. With a deeper understanding of the factors influencing mortality rates, practical solutions that optimize water quality conditions and ensure the well-being and productivity of fish populations can be develop. Our work serves as a compass, guiding us towards a future where aquaculture thrives, providing a vital source of nourishment while safeguarding the delicate balance of our aquatic ecosystems.

REFERENCES

[1] «FAO publications catalogue 2022». https://www.fao.org/documents/card/en/c/cc2323en (accessed on 10th february 2023).

[2] M. Føre, K. Frank, T. Norton, E. Svendsen, J. A. Alfredsen, T. Dempster, H. Eguiraun, W. Watson, A. Stahl, L. M. Sunde, C. Schellewald, K. R. Skøien, M. O. Alver, and D. Berckmans, «Precision fish farming: A new framework to improve production in aquaculture», *Biosystems Engineering*, vol. 173, pp. 176–193, 2018, doi: 10.1016/j.biosystemseng.2017.10.014.

[3] J. P. Fry, N. A. Mailloux, D. C. Love, M. C. Milli, and L. Cao, «Corrigendum: Feed conversion efficiency in aquaculture: do we measure it correctly? (2018 *Environ. Res. Lett.* **13** 024017)», *Environ. Res. Lett.*, vol. 13, n.º 7, p. 079502, jul. 2018, doi: 10.1088/1748-9326/aad007.

[4] «Environmental Problems of Aquaculture | Earth Journalism Network». https://earthjournalism.net/resources/environmental-problems-of-aquaculture (accessed on 10th february).

[5] S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li, and R. Zhao, «Application of machine learning in intelligent fish aquaculture: A review», *Aquaculture*, vol. 540, p. 736724, jul. 2021, doi: 10.1016/j.aquaculture.2021.736724.

[6] U. F. Mustapha, A. Alhassan, D. Jiang, and G. Li, «Sustainable aquaculture development: a review on the roles of cloud computing, internet of things and artificial intelligence (CIA)», *Reviews in Aquaculture*, vol. 13, n.º 4, pp. 2076–2091, sept. 2021, doi: 10.1111/raq.12559.

[7] L. Parra, S. Sendra, L. García, and J. Lloret, «Design and deployment of low-cost sensors for monitoring the water quality and fish behaviour in aquaculture tanks during the feeding process», *Sensors (Switzerland)*, vol. 18, n.º 3, 2018, doi: 10.3390/s18030750.

[8] T. Kassem, I. Shahrour, J. El Khatabi, and A. Raslan, «Smart and sustainable aquaculture farms», *Sustainability (Switzerland)*, vol. 13, n.º 19, 2021, doi: 10.3390/su131910685.

[9] «How to become data-driven in aquaculture», 2021. https://thefishsite.com/articles/how-to-become-data-driven-in-aquaculture (accessed on 15th february).

[10] « Reducing disease risk in Aquaculture». (accessed on 16th february). Available on: https://documents1.worldbank.org/curated/en/110681468054563438/pdf/882570REPLACEM00NAME0Reantaso0Melba.pdf

[11] J. Taylor, M. Graham, A. Louw, A. Lepheana, B. Madikizela, C. Dickens, D.V. Champan, and S. Warner, «Social change innovations, citizen science, miniSASS and the SDGs», *Water Policy*, vol. 24, n.º 5, pp. 708–717, may 2022, doi: 10.2166/wp.2021.264.

[12] M. S. Ahmed, T. T. Aurpa, and Md. A. K. Azad, «Fish Disease Detection Using Image Based Machine Learning Technique in Aquaculture», *Journal of King Saud University - Computer and Information Sciences*, vol. 34, n.º 8, pp. 5170–5182, sept. 2022, doi: 10.1016/j.jksuci.2021.05.003.

[13] «How can the Internet of Things (IoT) enhance fish health and welfare? - FutureEUAqua». https://futureeuaqua.eu/index.php/2021/02/24/how-can-the-internet-of-things-iot-enhance-fish-health-and-welfare/ (accessed on 20th february).

[14] J. Saltz, I. Shamshurin, and C. Connors, «Predicting data science sociotechnical execution challenges by categorizing data science projects», *Journal of the Association for Information Science and Technology*, vol. 68, n.º 12, pp. 2720–2728, dec. 2017, doi: 10.1002/asi.23873.

[15] B. Paul, S. Agnihotri, Kavya B., P. Tripathi, and Narendra Babu C., «Sustainable Smart Aquaponics Farming Using IoT and Data Analytics»:, *Journal of Information Technology Research*, vol. 15, n.º 1, pp. 1–27, jul. 2022, doi: 10.4018/JITR.299914.

[16] A. Sanou, S. Coulibaly, A.M.L. Guéi, M. Baro, E.F.T. Tanon, N. Méité, and B.C. Atsé., «Assessment of some physico-chemical parameters of the fish farm water in Abengourou, Côte d'Ivoire», *Egyptian Journal of Aquatic Biology and Fisheries*, vol. 26, n.º 5, pp. 319–343, 2022, doi: 10.21608/EJABF.2022.261442.

[17] J. Pereira, I. Abrantes, I. Martins, J. Barata, P. Frias, and I. Pereira, «Ecological and morphological features of Amyloodinium ocellatum occurrences in cultivated gilthead seabream Sparus aurata L.; A case study», *AQUACULTURE*, vol. 310, n.º 3–4, pp. 289–297, jan. 2011, doi: 10.1016/j.aquaculture.2010.11.011.

[18] I. P. Sousa-Filho, R.S. Moares, K.C. Saturnino, M. Tavares-Dias, Í.A. Braga, H.M. Ziemniczak, C.N. Souto, and D.G.S. Ramos, «First record of Trichodina heterodentata (Ciliophora: Trichodinidae) in banded knifefish Gymnotus carapo (Gymnotidae) cultured in Brazil», *Brazilian Journal of Biology*, vol. 82, 2022, doi: 10.1590/1519-6984.240840.

[19] O. A. Adeogun, G.A. Oladosu, M.M.A. Akinwale, O.A. Okunade, I.A. Akintayo, N. Idika, A.A. Adeiga, S.M.C. Ezeugwu, E.E. Afocha, O.S. Peters, and A.F. Odusanya, «Identification, distribution and prevalence of ecto-parasites associated with cultured fish in Ogun state, Nigeria», *Journal of Fisheries and Aquatic Science*, vol. 9, n.º 5, pp. 413–418, 2014, doi: 10.3923/jfas.2014.413.418.

[20] P. Tedesco, M. Saraiva, J.V. Sandoval-Sierra, M.T. Alves, R. Galuppi, J. Dieguez-Uribeondo, P. van West, A. Cook, P. Posen, B. Oidtmann, and M. Fioravanti, «Impact of abiotic factors and husbandry on saprolegniosis in salmonid farms», *Aquaculture*, vol. 561, 2022, doi: 10.1016/j.aquaculture.2022.738679.

[21] F. A. El-gohary, E. Zahran, E.A. Abd El-Gawad, A.H. El-gohary, F.M. Abdelhamid, A. El-mleeh, E.K. Elmahallawy, and M.M. Elsayed, «Investigation of the prevalence, virulence genes, and antibiogram of motile aeromonads isolated from nile tilapia fish farms in egypt and assessment of their water quality», *Animals*, vol. 10, n.º 8, pp. 1–19, 2020, doi: 10.3390/ani10081432.

[22] R. Teixeira, R. de Souza, A. Sena, B. Baldisserotto, B. Heinzmann, and C. Copatti, «ESSENTIAL OIL OF Aloysia triphylla IS EFFECTIVE IN NILE TILAPIA TRANSPORT», *BOLETIM DO INSTITUTO DE PESCA*, vol. 44, n.º 1, pp. 17–24, 2018, doi: 10.20950/1678-2305.2018.263.

[23] M. W. Ferreira, R.M. Santos, A.R. Da Silva, E.D. Marino, S.S. Makimoto, G.R.C. Barbosa, and G.B. De Andrade, «Mortality in pacus (Piaractus mesopotamicus) caused by pantoea agglomerans and pseudomonas aeruginosa in excavated tank», *Acta Scientiae Veterinariae*, vol. 47, 2019, doi: 10.22456/1679-9216.90826.

[24] T. Johnsen, W. Eikrem, C. Olseng, K. Tollefsen, and V. Bjerknes, «PRYMNESIUM PARVUM: THE NORWEGIAN EXPERIENCE», *JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION*, vol. 46, n.º 1, pp. 6–13, feb. 2010, doi: 10.1111/j.1752-1688.2009.00386.x.

[25] G. I. E. Ali, H. A. M. Abd El-Hady, and M. A. M. Abou Zeid, «Rapid Detection of Streptococci in Cultured Tilapia Fish Using PCR and Chemical Analysis», *World's Veterinary Journal*, vol. 10, n.º 3, pp. 286–296, 2020, doi: 10.36380/scil.2020.wvj37.

[26] A. Mustapha, «IMPORTANCE OF pH CONTROL IN AQUACULTURE», sept. 2019.

[27] X. Li, B. Jean-paul, Y. Liu, S. Triplet, e L. Michaud, «Effect of oxidation–reduction potential on performance of European sea bass (Dicentrarchus labrax) in recirculating aquaculture systems», *Aquaculture International*, vol. 22, pp. 1263–1282, aug. 2014, doi: 10.1007/s10499-013-9745-3.

[28] M. T. Dinis e R. M. Rocha, «INTRODUÇÃO À AQUACULTURA».

[29] «Water temperature in aquaculture - Responsible Seafood Advocate», *Global Seafood Alliance*, , 2018. https://www.globalseafood.org/advocate/water-temperature-in-aquaculture/ (accessed on 21th april 2023).

[30] «Ammonia/Ammonium | Hach». https://www.hach.com/parameters/ammonia (accessed on 21th april 2023).

[31] aqua-admin, «Nitrite & Nitrate in RAS Aquaculture - Reducing Risk, Improving Profits», *Aquamonitrix® - Nitrate and Nitrite Analyser*, 2021. https://aquamonitrix.com/nitrate-and-nitrite-aquaculture/ (accessed on 25th april 2023).

[32] X. Yang, X. Song, L. Peng, E. Hallerman, and Z. Huang, «Effects of nitrate on aquaculture production, blood and histological markers and liver transcriptome of Oplegnathus punctatus», *Aquaculture*, vol. 501, pp. 387–396, feb. 2019, doi: 10.1016/j.aquaculture.2018.11.048.

[33] A. Ciji and M. S. Akhtar, «Nitrite implications and its management strategies in aquaculture: a review», *Reviews in Aquaculture*, vol. 12, jun. 2019, doi: 10.1111/raq.12354.

[34] J. Kita, A. Ishimatsu, T. Kikkawa, and M. Hayashi, «- Effects of CO2 on Marine Fish», em *Greenhouse Gas Control Technologies - 6th International Conference*, J. Gale and Y. Kaya, Eds., Oxford: Pergamon, 2003, pp. 1695–1698. doi: 10.1016/B978-008044276-1/50278-6.

[35] «The inevitable pH fluctuations of aquaculture pond water - Responsible Seafood Advocate», *Global Seafood Alliance*, jan. 2017. https://www.globalseafood.org/advocate/the-inevitable-ph-fluctuations-of-aquaculture-pond-water/ (accessed on 25th april 2023).

[36] «nitrate and nitrite». https://www.lenntech.com/hazardous-substances/nitrate-and-nitrite.htm (accessed on 25th april 2023).

[37] «RSPCA welfare standards for Farmed Atlantic salmon». (accessed on 27th april 2023). Available on: https://science.rspca.org.uk/documents/1494935/9042554/RSPCA+welfare+standards+for+farmed+Atlantic+salmon+%28PDF%29.pdf/60ae55ee-7e92-78f9-ab71-ffb08c846caa?t=1618493958793

[38] «Fisheries :: Home». http://www.agritech.tnau.ac.in/fishery/fish_water.html (accessed on 27th april 2023).

[39] «A fish farmer's guide to understanding water quality - Part 2», nov. 2007. https://thefishsite.com/articles/a-fish-farmers-guide-to-understanding-water-quality-part-2 (accessed on 27th april 2023).

[40] G. Papageorgiou, S. W. Grant, J. J. M. Takkenberg, and M. M. Mokhles, «Statistical primer: how to deal with missing data in scientific research?», *Interact Cardiovasc Thorac Surg*, vol. 27, n.º 2, pp. 153–158, aug. 2018, doi: 10.1093/icvts/ivy102.

[41] R. W. Nahhas, *9.2 MCAR, MAR, MNAR | Introduction to Regression Methods for Public Health Using R*. (accessed on 28th april 2023). Available on: https://www.bookdown.org/rwnahhas/RMPH/mi-mechanisms.html

[42] «How to Handle Missing Data. "The idea of imputation is both… | by Alvira Swalin | Towards Data Science». https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4 (accessed on 28th april 2023).

[43] M. Mukaka, «A guide to appropriate use of Correlation coefficient in medical research», *Malawi Med J*, vol. 24, n.º 3, pp. 69–71, sept. 2012.