UNIVERSITY
INSTITUTE
OF LISBON

# Data Analysis for Precision Agriculture

Bruno Filipe Santos Saraiva

Master's in Telecommunications and Computer Engineering

Supervisor:
PhD Octavian Adrian Postolache, Associate Professor with habilitation, ISCTE

Supervisor:
PhD Pedro Joaquim Amaro Sebastião, Assistant Professor, ISCTE

October, 2024

TECHNOLOGY
AND ARCHITECTURE

Department of Electrical and Computer Engineering

# Data Analysis for Precision Agriculture

Master's in Telecommunications and Computer Engineering

Bruno Filipe Santos Saraiva

Supervisor:
PhD Octavian Adrian Postolache, Associate Professor with habilitation, ISCTE

Supervisor:
PhD Pedro Joaquim Amaro Sebastião, Assistant Professor, ISCTE

# Acknowledgements

Firstly, I would like to express my sincere gratitude to ISTA ISCTE and Instituto de Telecomunicações, IT-IUL for providing me with the opportunity and resources to complete this thesis, I am truly thankful for the support from the Department of Electrical and Computer Engineering at ISTA and the ISCTE community, who have contributed greatly to my academic journey.

I want to thank my coordinators, Professor Doctor Octavian Adrian Postolache and Professor Doctor Pedro Joaquim Amaro Sebastião that gave me a great initial push, which I needed to start this work, and Professor Doctor Nuno Manuel Branco Souto for his deep comprehension and compassion.

I want to thank my parents that gave me the opportunity to write this document, my mother Rosa Saraiva for the comfort and love she provided and my father Luís Saraiva for the tenacity he gave me. To my sister Beatriz Saraiva, that supported me in my hardships. To partner Daniela Nunes, with her love and compassion, closely saw my ups and downs, and hardships while managing my workload with my degree. Also, her parents João Nunes and Célia Gomes Nunes, that gave me strength. I also want to thank all my friends that pushed me further into developing this work and my familiars for their moral support. This journey would have been a lot harder if it weren't for these people, and for all of that I want to express my deepest gratitude.

Sincerely, thank you so much, it means so much to me.

# Resumo

A agricultura de precisão utiliza uma análise de dados sofisticada para maximizar o rendimento agrícola e a eficiência dos recursos, permitindo métodos agrícolas personalizados que reagem a condições específicas do solo e do clima, melhorando assim a sustentabilidade e a produção. Este trabalho apresenta o desenvolvimento e implementação de um sistema de análise de dados para prever as necessidades hídricas no contexto agrícola, integrando assim o projeto "Soil IoT". O sistema utiliza sensores simulados para monitorizar as condições do solo, como humidade, temperatura e condutividade, transmitindo os dados em tempo real através do protocolo MQTT para a plataforma ThingsBoard. Foram realizados estudos com os dados, um utilizando dados reais dos sensores e outro com um dataset sintético gerado a partir dos dados originais onde foi possível analisar de forma detalhada os padrões de humidade do solo. Estas investigações permitiram uma análise mais aprofundada, especificamente nos padrões de humidade do solo, permitindo assim a previsão das necessidades de irrigação e a geração de alertas para os utilizadores da plataforma, verificando assim o valor da plataforma e de todo o sistema envolvente da análise dos dados, na orientação de decisões de irrigação. A plataforma desenvolvida oferece uma interface intuitiva que permite aos utilizadores monitorizar e tomar decisões baseadas nos dados analisados promovendo uma gestão hídrica mais eficiente e sustentável, onde dado um determinado alerta sobre os níveis de humidade do solo estarem em níveis críticos, o utilizador sabe que deverá tomar as devidas medidas.

**Palavras-chave**: Dados, MQTT, ThingsBoard, Previsão, agricultura, LSTM.

# Abstract

Precision agriculture uses sophisticated data analysis to maximize crop yields and resource efficiency, enabling customized farming methods that react to specific soil and climate conditions, thus improving sustainability and production. This work presents the development and implementation of a data analysis system to predict water needs in the agricultural context, thus integrating the “Soil IoT” project. The system uses simulated sensors to monitor soil conditions such as humidity, temperature, and conductivity, transmitting the data in real time via the MQTT protocol to the ThingsBoard platform. Studies were conducted with the data, one using real data from the sensors and the other with a synthetic dataset generated from the original data where it was possible to analyze the soil moisture patterns in detail. These investigations allowed for a more in-depth analysis, specifically into soil moisture patterns, thus allowing for the prediction of irrigation needs and the generation of alerts for platform users, consequently verifying the value of the platform and the entire system surrounding data analysis in guiding irrigation decisions. The developed platform offers an intuitive interface that allows users to monitor and make decisions based on the analyzed data, promoting more efficient and sustainable water management, where given a certain alert about soil moisture levels being at critical levels, the user knows to take the appropriate measures.

**Keywords**: Data, MQTT, ThingsBoard, Prediction, Agriculture, LSTM.

# Table of Contents

# List of Figures

# List of Tables

# Glossary

IoT – Internet of Things

WSNs - Wireless Sensor Networks

N - Nitrogen

P - Phosphorus

K – Potassium

PO4-P, ppm - Phosphate Phosphorus parts per million

Total N, % - Percentage of total nitrogen (N) in a substance or sample

Soluble K, % - Percentage of total soluble potassium (K) in a substance or sample

EC - Soil electrical conductivity

IPCC - The Intergovernmental Panel on Climate Change

kg/ha - 1 kilogram / hectare

Crop Evapotranspiration - Total amount of water that is lost from a crop's growing area due to evaporation from the soil and transpiration from the plants themselves.

Edaphoclimatic conditions – This determines the suitability of a region for specific crops and agricultural practices.

EDA – Exploratory Data Analysis

LSTM – Long Short-Term Memory

VWC – Volumetric Water Content

LoRa - Long Range modulation technique

LoRaWAN - Long Range Network protocol

WHIN - Wabash Heartland Innovation Network

IQR - Interquartile Range

# Chapter 1
# **Introduction**

Data analysis in agriculture is increasingly pivotal as it drives smarter decision-making and enhances sustainability in farming practices. This chapter introduces the "Soil IoT" project, an initiative designed to harness real-time soil data through a network of advanced sensors, including moisture, humidity, temperature, and conductivity sensors. The project aims to collect and transmit environmental data continuously, facilitating an integrated decision-support system for optimized water and nutrient management.

The Soil IoT project is built on the foundation of comprehensive data collection and analysis, leveraging technology to deepen our understanding of soil conditions and support informed decisions. The introduction begins by detailing the motivation behind this project, emphasizing the need for efficient agricultural management amid rising challenges such as climate change and resource scarcity. It then outlines the specific objectives of this research, focusing on building an IoT platform capable of predictive modeling, real-time monitoring, and supporting smart farming solutions. Further, the investigation questions that guide this study are explored, including inquiries into the effective integration of IoT technology in agriculture and the ways in which sensor data can empower users to make better-informed decisions. This introduction also describes the methodological approach taken throughout this research, grounded in the Scientific Method model, and includes testing and validating predictive models that were developed during the project.

The document outlines the state of the art in data analysis, IoT, and machine learning in precision agriculture. It then details the development of the Soil IoT platform, sensor setup, and data collection methods. The data analysis section covers visualization and predictive modeling for optimizing irrigation. The machine learning models, including Random Forests and LSTMs, are discussed to support predictions. The design of the user interface is reviewed for usability and effective metric visualization. Finally, the document concludes with key findings, the project's impact, and recommendations for future improvements.

## 1.1 Motivation

Since the beginning of times water has been the basis of life. Without this element, life would not exist [1]. Nowadays sustainability is a particularly important matter, growing day by day, therefore emissions must be reduced to reduce the carbon footprint and create new sustainable technologies that aim to preserve nature and save the most valuable resource. This project plays a significant role in this matter because it addresses sustainable farming.

In the Mediterranean region, where unfavorable climatic conditions contribute to water shortages, tend to worsen with climate change. In this region agriculture is seriously threatened also by the risks of decreasing water quality and land salinization, under conditions of water scarcity and high atmospheric evaporation pressure [2]. The Intergovernmental Panel on Climate Change (IPCC) warns in its latest report that the Mediterranean is one of the areas where climate change is advancing the most [3].

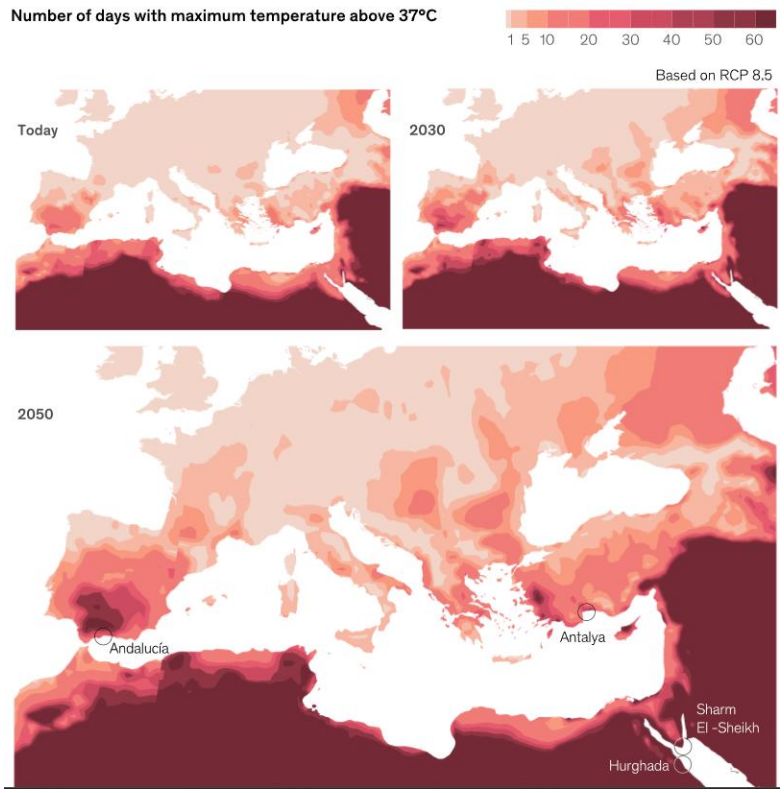



Figure 1 - Number of days above 37ºC in southern Spain, Turkey, and Egypt, that expected to double by 2050, from about 30 to 60, dataset by EURO-CORDEX RCM ensemble *[4]*

This figure shows that water will become scarce as draught scenarios will be more persistent. Therefore, it is essential to develop sustainable farming practices to tackle the

scarcity problem, considering that almost 75% of fresh water is used for agricultural purposes, a more resilient approach is needed [3].

It is intended to apply sensors on farming fields such as soil moisture sensors to measure the soil health and precisely irrigate the fields with water and nutrients, as necessary. The capacity to accelerate research and development towards smarter farming is made possible by the growing accessibility of big data and big data analysis techniques, that helps farmers meet the challenge of producing high yield crops on a larger scale and in a more sustainable manner [5].

Nitrogen (N), phosphorus (P), or potassium (K) concentrations, also known as NPK [6], soil moisture and soil conductivity are crucial for regulating the development of the plantations in the field therefore, a good soil composition is required for the plants to thrive. Soil moisture is important for the physical structural strength of a plant while temperature, humidity and light are required for the plant's photosynthesis process [7].

## 1.2 Objectives

Agriculture needs more sustainable methods in the face of climate change and extended drought. Precision farming is related to sustainability-focused solutions [4]. The objective evaluation of soil conditions and mapping using geographic information systems are of highest relevance in precision farming.

To optimize production efficiency by controlling soil moisture levels, it is essential to assess soil properties such as moisture content, as well as nutrient concentrations like nitrogen, phosphorus, and potassium (NPK), conductivity and even soil temperature. Sensor networks, utilizing IoT solutions with communication protocols like LoRa or Wi-Fi, play a vital role in monitoring these properties, enabling the development of predictive models for nutrient availability and water management optimization by controlling soil moisture levels.

This project aims to develop a comprehensive data analysis platform within an agricultural context, integrating Internet of Things (IoT) technologies to address the unique challenges of modern farming. By providing insightful data and actionable information, the platform seeks to empower farmers, agronomists, and agricultural researchers in making informed decisions about their fields. The IoT devices will relay the collected data through a chosen and optimized communication protocol, considering that the devices will be simulated with random data,

which is optimized to be like real data. Once sensor nodes have gathered the data from the simulated plantation environment, the platform will act as the centralized hub for data aggregation while providing visualization of metrics through an intuitive and user-friendly interface, users will have the ability to access the platform from any device with internet connectivity. The real-time nature of the data presentation will allow users to monitor the health and conditions of the plantation remotely, enabling them to make informed decisions promptly. This visualization layer will enable users to identify trends or anomalies that might have previously gone unnoticed.

## 1.3 Research Questions

- Why NPK concentration and Soil Moisture Assessment and Analysis?
- How will the data be received on the platform?
- How can be beneficial to use the data on the platform to empower the user acting in agriculture?
- How will the analyzed data be implemented to the platform?

## 1.4 Methodology

The type of methodology used to develop the dissertation is based on the Scientific Method Model which is a systematic approach on how to conduct proper research and gather knowledge. It involves the following general steps, which are illustrated in Figure 2 [8]:

- 1st Step – Observation/Question/Definition of Objectives: questions to be investigated under the project proposal (Chapter 1).

- 2nd Step – Research Topic Area: It defines the scope and context of the research. (Chapter 2).

- 3rd Phase - Hypothesis: A hypothesis is a testable conjecture that attempts to explain the observed behavior or answer the question. It is proposed that sensors can help with water management (Chapter 2).

- 4th Phase – Test with experiment: Consists of conducting experiments or suitable tests to compare the expected outcomes based on the hypothesis with the actual results obtained through experimentation (Chapter 3).

- 5th Phase – Analysis: Involves interpreting the data to draw meaning (Chapter 4 and Chapter 5).

- 6th Phase – Report Conclusions: It takes the analyzed data and draws conclusions, taken during the essay, while refereeing improvements in future work (Chapter 6).

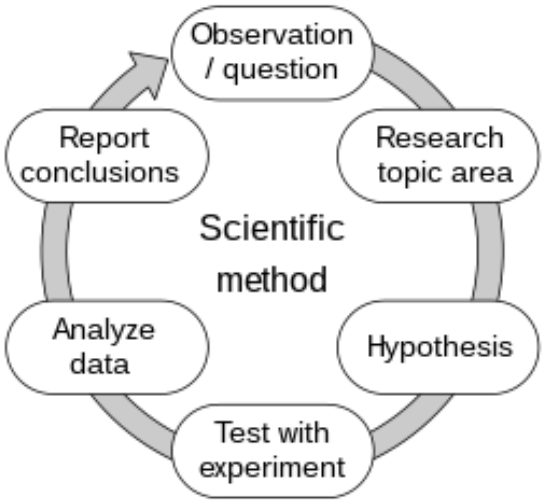



Figure 2 - Scientific Method Model [8]

## 1.5 Document Structure

The framework of this dissertation is outlined in the following manner:

• Chapter 2 - State of the Art: Literature review covering areas related to the dissertation.

•Chapter 3 - Platform Development: Presentation of the platform, as well as its development.

• Chapter 4 - Data Analysis: Introduction to formative and summative tests, including the analysis and validation of the work in question.

•Chapter 5 – Predictive Model: Study of the algorithm of choice to use in predictions.

•Chapter 6 - Conclusion and Future Work: Conclusions resulting from the work in question, including answers to the research questions.

# Chapter 2
# **State of The Art**

The Internet of Things (IoT) has revolutionized various industries by enabling the seamless connectivity of physical devices and the exchange of data between them [9] One such industry that has significantly benefited from IoT is precision agriculture. Precision agriculture refers to the use of advanced technologies and data-driven techniques to optimize agricultural practices, enhance crop yield, and reduce resource consumption [10]. The integration of IoT in agriculture has opened new ways for collecting and analyzing real-time data from diverse sources such as sensors, drones, and satellites [11].

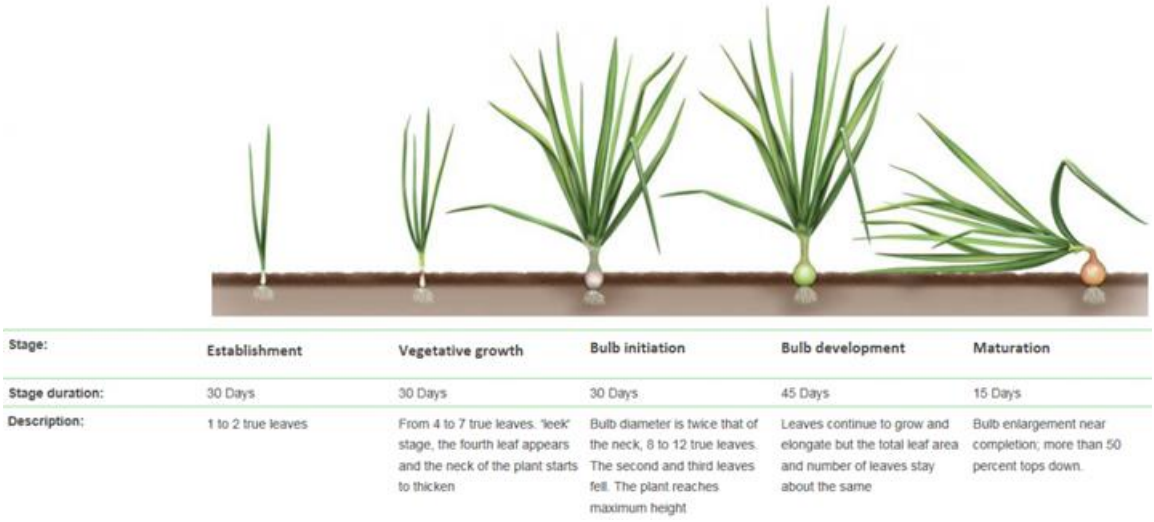
This literature review explores the fundamental role that precision agriculture and IoT technologies play in forwarding sustainable agriculture within the context of climate change and this project itself. To achieve the intended objective, an extensive and comprehensive literature review was conducted where the efforts involved inspecting a wide array of publications obtained from various sources, including conferences, journals, published documents and edited volumes, by systematically searching academic databases, mostly the EBSCOhost - Academic Search Complete, using targeted keywords such as "Internet of Things," "IoT in Irrigation" "Precision Agriculture," and "Weather Conditions," relevant literature was identified and extracted. These are the main keywords used to filter the articles, all peer reviewed, that sustain this State of the Art.

The investigation in question is fulfilled using the Boolean Methodology which relies upon Boolean searching, making use or operators like AND, OR, NOT, to search the pretended keywords in the title or in the text, in other words this methodology is a structured means of creating a search "equation" [12]. Following a strict review process, articles were thoughtfully selected from a total pool of 20 of 553 papers, however on top of the main keywords used in the Boolean research methodology, more filters were applied as a mean to section the information into the following categories: “Precision Agriculture and Its Significance”, “Climate Change and Agricultural Challenges”, “IoT Solutions in Precision Agriculture” and “Data Collection and Predictive Models”.

## 2.1 Precision Agriculture

Precision agriculture, fundamentally rooted in data-driven decision-making, redefines the way crops are cultivated and manage resources, in other words, precision agriculture seeks to optimize farming operations by adapting them to the specific needs of each plot of land. This section sets the stage by defining precision agriculture and highlighting its multifaceted benefits through key studies and examples, it showcases the tangible impact of precision agriculture on enhancing efficiency, conserving resources, and promoting environmental sustainability.

To put in perspective, onions should contemplate a level of moisture that doesn't drop below 75% [8], optimum pH is in the range of 6 to 7 where fertilizer requirements are normally 60 to 100 kg/ha N, 25 to 45 kg/ha P and 45 to 80 kg/ha K. [9]. However, it's not as straightforward as it may seem, as all plants go through various developmental stages. and each development stage will require a different nutrient management, as can be observed in the following Figure 2.



| Stage: | Establishment | Vegetative growth | Bulb initiation | Bulb development | Maturation |
|---|---|---|---|---|---|
| Stage duration: | 30 Days | 30 Days | 30 Days | 45 Days | 15 Days |
| Description: | 1 to 2 true leaves | From 4 to 7 true leaves. 'leek' stage, the fourth leaf appears and the neck of the plant starts to thicken | Bulb diameter is twice that of the neck, 8 to 12 true leaves. The second and third leaves fell. The plant reaches maximum height | Leaves continue to grow and elongate but the total leaf area and number of leaves stay about the same | Bulb enlargement near completion; more than 50 percent tops down. |

Figure 3 - Onion Growth Stages, figure by Haifa Group [13]

According to Haifa Group [9], as soil depth increases, rooting density decreases, therefore it is necessary to maintain soil moisture and nutrient levels in the shallow rooting area. To maximize growth and produce high yields, soil must be fertile and well-structured. When the tops fall off, bulbs are harvested. The plant should not flower to produce bulbs because flowering impacts yields. The length of the day, low temperatures (below 14–16°C) and low humidity are necessary for flowering.

Table 1 - Tissue analysis guide for diagnosing crop nutrient status in onions, by Haifa Group [13]

| Sampling time and nutrient | Nutrient level | | |
|---|---|---|---|
| Deficient | Intermediate | Sufficient | |
| Early season (pre-bulbing) | | | |
| Total N, % | <3 | 3-4 | >4 |
| $PO_4$-P, ppm | <1000 | 1000-2000 | >2000 |
| Soluble K, % | <3 | 3-4 | >4 |
| Mid-season (bulbing) | | | |
| Total N, % | <2.5 | 2.5-3 | >3 |
| $PO_4$-P, ppm | <1000 | 1000-2000 | >2000 |
| Soluble K, % | <2.5 | 2.5-4 | >4 |
| Late seaon (post bulbing) | | | |
| Total N, % | <2 | 2-2.5 | >2.5 |
| $PO_4$-P, ppm | <1000 | 1000-2000 | >2000 |
| Soluble K, % | <2 | 2-3 | >3 |

Thus, labeled recommended values, according to Haifa Group should also be contemplated in sensor data, to ensure the quickest and healthiest way to produce, in this hypothetic example, onions while managing resources.

Regarding the soil conductivity, it is measured through the salt content in the soil, EC is an important marker of the health of the soil. It will also have an impact on crop output and quality, the availability of nutrients to plants, and soil microbial activity that is linked to important soil processes, including greenhouse gas emissions like nitrogen oxides, methane, and carbon dioxide [10].

As IoT technologies are evolving and becoming more sophisticated, have increasingly played a significant role in transforming Precision Agriculture practices, in a way that IoT's ability to connect physical objects to the internet has allowed for more precise and data-driven decision-making in agriculture, therefore turning it more productive, efficient, resource conservation, and environmental sustainability within the Precision Agriculture scope, keeping alive this evolutionary cycle [14].

## 2.2 Climate Change and Agricultural Challenges

The agricultural terrain is undergoing trough a multitude of transformations driven by climate change as prolonged droughts, erratic weather patterns, for example the Iberian Peninsula draught scenario, mentioned in section 1.1 and shifting growing seasons pose formidable challenges to farmers worldwide. This section delves into the current state of

climate change, presenting evidence of the growing threats faced by agriculture by also highlighting the urgency of innovative solutions that can adapt farming practices to these shifting environmental conditions.

The importance of energy efficiency in agriculture has risen to the top of the list of world issues, where the need for sustainable and effective agricultural practices has grown stronger as the world battles severe obstacles including climate change, resource shortages, and rising population. Thus, energy is essential to agricultural production since it powers a variety of procedures, including processing and transportation as well as machinery and irrigation systems so, understanding the developments in scientific research and cooperative initiatives relevant to energy efficiency in agriculture is crucial. One mention that is very relevant to underline in this article is that in the Netherlands, the United Kingdom, Portugal, and Belgium, methods that calculate Crop Evapotranspiration (ETc) or rely on incoming solar radiation are commonly used in 60-95% of crops. In Spain, Italy, the Netherlands, and Portugal, 10-15% of soilless crops utilize soil/substrate moisture sensors, including tensiometers, whereas three distinct irrigation control methods were evaluated for their impact on various agronomic parameters. The irrigation methods included the gravimetric method, amongst Volumetric Water Content (VWC) Control System and Radiation (AR) Control System, which assessed water needs by monitoring the weight of the growing medium, the accumulated radiation method, which relied on solar radiation data, and the Volumetric Water Content (VWC) sensor, measuring substrate water content. Results indicated that the gravimetric method was the most effective, yielding the highest commercial fruit production, especially under control conditions. These findings underscore the significance of the gravimetric method in optimizing irrigation, given its ability to precisely meet plant water requirements [15].

Also, in line with these concerns, the development of Smart Irrigation Equipment for Soilless Crops, is identified as a critical research area, in southeast Spain, where edaphoclimatic conditions are expected to worsen due to climate change, efficient water use in agriculture is imperative. The study suggests that innovative smart irrigation technologies can address this challenge effectively, currently, due to high costs, 60–80% of soilless crop irrigation relies on grower or advisor experience, whilst focusing on a designing a cost-effective control system to optimize soilless crop irrigation [16].

A comprehensive study has been conducted on the regional climate effects of irrigation under the warming of Central Asia by 2.0 °C. This research employed the Weather Research

and Forecasting (WRF) model to design three types of experiments: historical experiments, warming experiments using future driving fields, and warming experiments involving increasing surface energy where, two scenarios were considered, one with irrigation and one without. By comparing the results of these experiments with historical data, was revealed that a 2.0°C warming in Central Asia is demanding an increase in irrigation by 10-20%. It demonstrates variable impacts on precipitation dynamics depending on the type of warming experiment, highlighting the need for adaptive agricultural and water resource management strategies in response to climate change, therefore it must be considered irrigation in climate models to accurately assess the impact of climate change on water resources worldwide, that face similar challenges at the intersection of water resource management and climate change [17].

## 2.3 IoT Solutions in Precision Agriculture

This section provides details on several IoT technologies and communication protocols, such as LoRa, Wi-Fi, and NB-IoT, and describes the idea of IoT and its application in agriculture, demonstrating how IoT applications are changing agriculture through practical examples, focusing on soil monitoring, crop management, and resource optimization. The section also includes case studies that have been successful, highlighting the significant advantages of IoT solutions in contemporary agriculture.

Important technological advancements in agriculture, particularly in the domain of crop disease detection using Unmanned Aerial Vehicles (UAVs) and deep learning techniques, provide the ability to detect crop diseases. It prevents yield losses and increases agricultural benefits while integrating UAVs, remote sensing, and artificial intelligence which has shown promise in revolutionizing crop disease detection, offering efficient and cost-effective solutions. It is achieved by offering a detailed taxonomy and meta-analysis, in other words accesses the performance of machine learning models employed. It is highlighted a multitude of challenges, opportunities, and potential research directions in the field of drone-based remote sensing for crop disease estimation, that represents a significant step towards addressing the challenges faced by modern agriculture in a rapidly changing world [18].

With a primary focus on effective irrigation control, a crucial area given the growing challenges posed by climate change and the need for responsible water resource

management, in contrast to many existing IoT solutions that primarily stress sensor hardware development. The authors provide a software framework that reimagines the design of IoT systems, emphasizing software services in a client-server format made possible by REST interactions that are done in recognition of the crucial role that software plays in modern computer systems, while using irrigation in agriculture to demonstrate the viability and usefulness of their methodology, offering insightful information on how to create adaptable and effective IoT solutions for precision agriculture. The mentioned software framework has been put into practice in a real-world IoT irrigation use case, where it processes data, monitors field conditions, and quickly recognizes and reacts to warnings, while showing a significant performance, it also guarantees that the temporal overhead stays within bounds and is appropriate for the domain for which it is intended, with a response time of about 11 seconds even under difficult circumstances. This work offers a promising foundation for the development of software-centric IoT solutions in agriculture and beyond, addressing crucial issues like water conservation, crop health, and sustainable farming methods in an era where efficient resource management and real-time data-driven decision-making are dominant [19].

Thanks to LoRa's low bandwidths of 7.8 kHz or a maximum of 500 MHz, the range is currently about up to 10km to 11km, which is ideal for long-range transmissions [20] covering large fields, whereas Zigbee registers at around 2 MHz of bandwidth, balancing bandwidth and energy efficiency for medium-range communication from 10m-100m. It is ideal for applications like smart lighting or building automation but cannot match the long-range capabilities of LoRa [21]. With a bandwidth of around 1.4 MHz, LTE-M offers longer ranges than LoRa and supports mobility, particularly suitable for narrowband and long-range applications [22]. Given this information the only protocol that combines both those features and the agricultural field, LoRa is the ideal choice.

The impact of packet size on energy efficiency is examined, and it is suggested that, in contrast to the 11-bit packets used in standard commercially available equipment, smaller 6-bit packets are adequate for energy-efficient data collection in precision agriculture. This will help to mitigate the impact of climate change on agricultural production by achieving both long communication range and energy efficiency. The research also explores in an open-area IoT implementations by also carefully selecting parameters such as bandwidth, spreading factor, and error correction rate to achieve reliable communication with minimal energy

consumption, where the environment is high in interferences. Overall, it is intended to underline the critical role of LoRa-based wireless sensor networks in improving energy efficiency and environmental monitoring in precision agriculture, with a focus on optimizing key parameters to reduce energy consumption and extend the lifespan of sensor nodes [23].

## 2.4 Data Collection and Predictive Models

Integration of data from various sources, such as weather data, soil data, and satellite imagery, allows for a holistic view of the agricultural ecosystem, thus, combining data from different sensors and platforms provides farmers with a comprehensive understanding of their fields, leading to more effective and efficient agricultural practices. Integrated data can be used to generate accurate and localized recommendations for irrigation, fertilization, and pest control, considering the specific needs of each crop and field.

The Machine Learning (ML)-based weather forecasting model, which harnesses the power of the Social Spider Algorithm-Least Square-Support Vector Machine (SSA-LS-SVM) algorithm is employed to predict crucial weather and soil parameters, including atmospheric temperature, pressure, and soil humidity, for 24, 48, and 72 hours. These predictions are derived from a comprehensive dataset comprising 39 days of hourly data for Amman city. Notably, the predicted values exhibit low relative mean square errors when compared with both the actual values and the LS-SVM predictor, underscoring the model's accuracy and effectiveness in weather prediction [24].

Regarding the use of Unmanned Aerial Vehicles (UAVs) for crop disease estimation, which categorizes the methods Spectral-Texture (ST)-based, conventional Machine Learning (ML)-based, and Deep Learning (DL)-based approaches. It evaluates the impact of UAV platforms and sensors on crop disease estimation and compares the performance of ML and DL methods against traditional ST-based technique, where it is concluded that DL-based models prove to be the most successful due to its adaptability and reliability. By combining various remote sensing data modalities for enhanced crop disease detection and developing lightweight DL models for edge computing platforms like the Internet of Things (IoT), represent emerging chances of exploration. Addressing these challenges will contribute to greater reliability in DL models for this application[18].

A new concept of quantifying rice growth utilizes prediction models, applied to analyze the impact of environmental factors on agricultural production, with a particular focus on the use of neural networks. It is highlighted that the significance of rice growth prediction and the current research status in the field of rice prediction, carefully studies the features of rice growth recurring to neural networks and integrates the quantitative indicators of each growth stage as rice growth, thus providing a quantitative basis for rice growth prediction models. It was found that the Elman neural network is easy to fall into the local extreme value, which leads to a large deviation of individual points. After analyzing a variety of optimization methods, the genetic algorithm is finally accustomed to optimizing the weight and threshold of the Elman neural network efficiency, which not only ensures the diversity but also improves the search ability of the algorithm [26].

Despite the significant progress made in IoT, monitoring, and precision agriculture, several challenges persist, some of the key challenges include the interoperability of devices and systems, energy efficiency and data security, however, these challenges also present opportunities for research, innovation, and collaboration amongst stakeholders in the agriculture and technology sectors.

# Chapter 3
# Setup, Development, and Implementation of Soil IoT Platform

This chapter delves into the core of the "Soil IoT" platform development process to achieve its goals. This chapter's main goals are to clarify the approaches used, which were crucial to develop the project, and to examine the important facets of the development stage.

## 3.1 Development Methodology

The decision to adopt an Agile methodology for the development of the precision agriculture Internet of Things (IoT) ecosystem is a well-founded choice that aligns closely with the project's complexity, goals, and the evolving nature of modern farming practices.

The “Soil IoT” involves multiple intricate components, such as sensor mockup integration, and user interface design. Agile's iterative approach allows the project to be broken down into manageable iterations, ensuring that each component can be developed, tested, and refined incrementally. This iterative cycle facilitates constant feedback and adjustments, leading to a more refined product [27].

### 3.1.1 Flexibility and Adaptability:

The field of precision agriculture is fast developing due to recent technologies, altering farming practices, and growing stakeholder demands. Because of agile's flexibility, the project can be easily adjusted to these changes and can adjust to new features, changed priorities, and unexpected obstacles, all of which help to ensure that the final product stays in line with the needs.

### 3.1.2 Risk Mitigation:

Agricultural systems are inherently subject to uncertainties, such as variations in weather conditions and crop health. This methodology approach of delivering functional components incrementally allows for early identification of potential risks or issues, enabling these changes to be address challenges promptly, reducing the likelihood of major setbacks, during development stages.

Agile's iterative nature ensures that functional components are delivered in short cycles, allowing users to start benefiting from the ecosystem's capabilities sooner marking this as particularly beneficial for a project that aims to provide real-time insights to aid decision-making.

### 3.1.3 Continuous Improvement:

Regular retrospectives at the conclusion of each cycle are one way that agile approaches promote continual improvement. This approach promotes a culture of learning and creativity by enabling the project team to consider what worked, what could be improved, and how to improve procedures. In summary, the Agile methodology is a prudent choice for this precision agriculture IoT ecosystem project due to its capacity to accommodate the complexities of modern farming, adapt to future evolving requirements. This approach not only increases the chances of project success but also aligns with the dynamic and ever-changing nature of the agricultural industry.

### 3.1.4 Tasks Setup

Taking the definition of the Agile methodology a board with tasks was created to follow the very principles of this methodology. In the Table 2, can be observed the tasks created and planned to achieve this project objectives. The tasks became deliverables, and each one has an estimate that was attributed using story points, which in this methodology replaces the hours required, it’s an approximation to facilitate the development. Using the referred methodology 7 tasks were created regarding the aim of the project, having a total of 191 hours, and 95 story points estimated.

Table 2 - Repartition of Development Stages with Agile Methodology

| Deliverable | Hours Required | Story Points |
|---|---|---|
| Project Scope and Objectives | 10 | 8 |
| Communication Protocol Specification | 4 | 3 |
| Platform UI Design | 20 | 13 |
| Backend Development | 40 | 20 |
| Pilot Testing | 5 | 3 |
| Platform Deployment | 12 | 8 |
| Data Analysis | 120 | 50 |
| Total | 211 | 95 |

### 3.1.5 Project Scope and Objectives

On this task it is defined the criteria used to choose frameworks and the ideal platform to display the data, taking in consideration the possible escalation of the project.

- Define project scope and objectives.
- Identify available resources (sensors, communication technologies, development tools).
- Identify key data points to be collected (soil moisture, temperature, humidity, conductivity levels).

### 3.1.6 Communication Protocol Specification

This task plans the design and specifications of the chosen communication protocol. It focuses on the choice of the sensor's client communication protocol, which received data from simulated devices.

- Evaluates available communication protocols to communicate with the cloud platform.

### 3.1.7 UI Platform Design:

The UI design encompasses the user interface elements of the platform. This includes the layout, visual elements, and interactive features that stakeholders will engage with. The design ensures a user-friendly and intuitive experience, catering to users ranging from farmers to researchers.

- Design the user interface for the platform, including real-time data visualization.

### 3.1.8 Platform Backend Development

Regarding platform's backend development, this task involves building the core functionality that handles data aggregation, processing, and storage. This component ensures that data collected from sensors is processed, analyzed, and made available for visualization on the platform's frontend.

- Design the architecture to reach the platform, to include mocked real-time data visualization.

### 3.1.9 Data Analysis Platform

As soon as backend deployment is concluded, the platform will receive the data and the dashboard is assembled. This task also involves scaling up the infrastructure, ensuring stability, and making the platform accessible to all intended users.

- Develop the backend infrastructure for data aggregation, storage, and processing.
- Implement data visualization tools for data interpretation in the dashboard.
- Debug and resolve issues.

### 3.1.10 Pilot Testing

The pilot testing phase involves deploying the ecosystem to a subset of users for validation. Feedback collected during this phase helps refine the system, addressing any issues and improving functionality based on real-world usage.

- Employ test users on the platform.

### 3.1.11 Data Analysis

This task presents the findings of in-depth data analysis, highlighting trends, correlations generated from the collected data. It aids users in making informed decisions regarding nutrient management, water usage, and crop health.

- Analyze collected data to identify trends, correlations, and insights.
- Choice of the

## 3.2 Architecture

The heart of the chapter unfolds in the section dedicated to the actual development process from gathering the right technologies to the implementation of them. It provides insights into the design, implementation, and integration of the Soil IoT system. The subsection elucidates the technical aspects, such as the creation of the virtual moisture sensors, data transmission via MQTT, and the integration with the ThingsBoard platform, which collectively form the core of the project by showing the visualization of the received data.

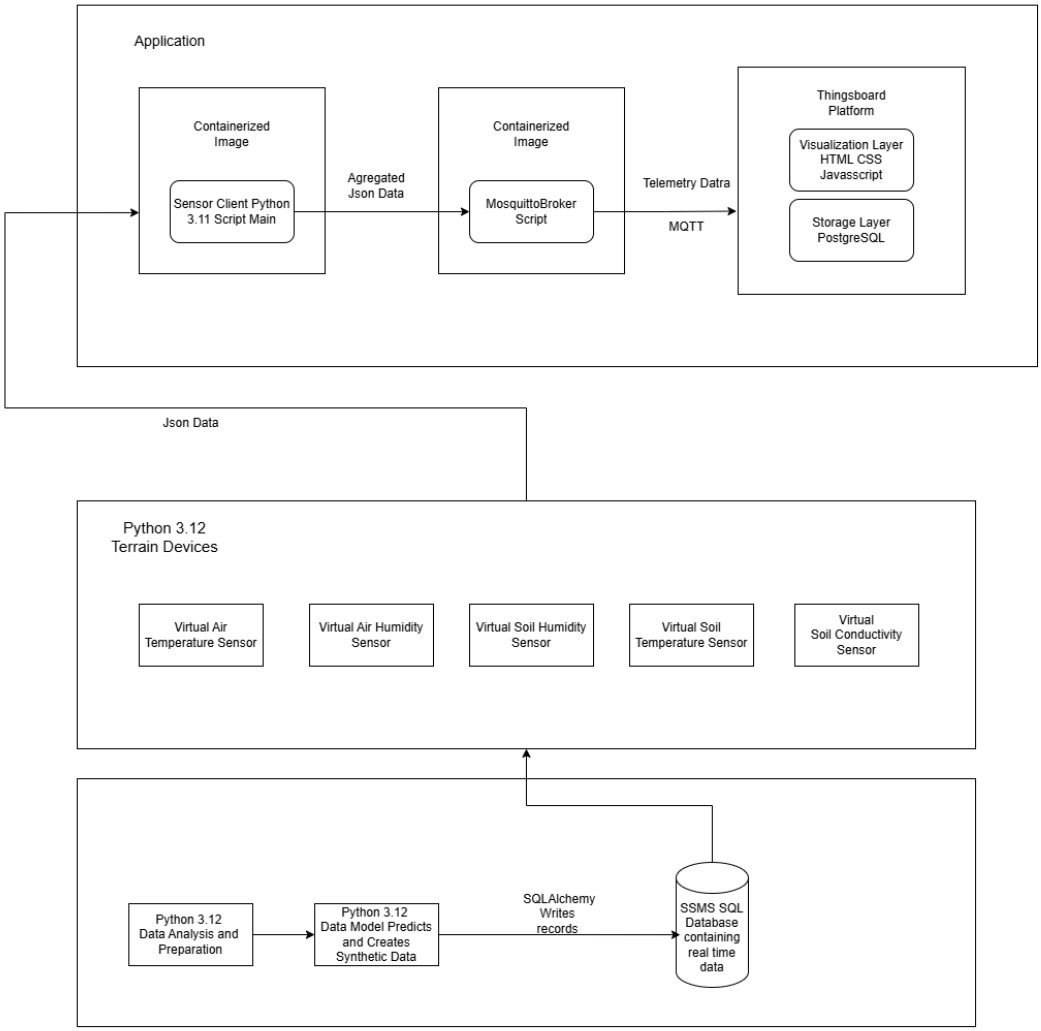



Figure 4 - Main Architecture Design

As it can be observed in Figure 4, the architecture has 5 different types of sensors (virtual sensors) whose goal is to send real-time synthetic data to mimic real live data. This simulation attempts to recreate the sensors, that are in a terrain communicating, using synthetic data created from a real dataset, as it will be thoroughly analyzed in chapter 4 and 5. This synthetic data that is aggregated in a database, posteriorly it is linked to the virtual sensors which are then linked to a client server that aggregates all the information received by these devices and sending it via MQTT to the Mosquitto broker.

The Mosquitto MQTT Broker serves as the MQTT broker that facilitates communication between the simulated sensors and the ThingsBoard IoT platform by providing a secure, lightweight, and efficient data transmission, by balancing the data stream load into, and publishing onto different topics, that correspond to different types of data [27].

Following the diagram in figure 4, the broker sends the data to the cloud platform, ThingsBoard is responsible for receiving, processing, and visualizing the data from the simulated sensors. It also hosts the rule chain for data processing and automation, enabling possible real-time decisions based on incoming data.

The architecture allows for scalable and efficient management of many devices while enabling real-time data processing and automation for agricultural monitoring and control.

### 3.2.1 Technologies Stack

When considering an IoT platform for the elaboration of an agricultural project, ThingsBoard stands out as a strong contender for several compelling reasons. ThingsBoard is an open-source IoT platform that enables rapid development, management, and scaling of IoT projects [28]. Primarily, is an open-source platform, which grants the flexibility to customize and adapt it to the specific needs of your agricultural IoT application with its rich and versatile toolset. The platform provides a comprehensive suite of features encompassing device management, data visualization, rule-based automation, and user interface creation, this is particularly valuable for agricultural monitoring and irrigation control, as it offers the tools necessary to manage the difficulties of data management and automation in an agricultural context. Regarding the platform's rule chain engine, it is possible to set up intricate automation scenarios based on the data coming in from IoT devices such as RPC calls the trigger exterior API's [28]. This capability is invaluable for making data-driven decisions, such as determining when to trigger irrigation based on sensor data and weather forecasts. Data visualization is another strength, by providing with ease, customizable dashboards and widgets that facilitate the creation of user-friendly interfaces for displaying sensor data and device statuses, simplifying data interpretation, and enhancing user experience. Integration with Python-based simulated devices and other IoT hardware is made relatively straightforward thanks to the platform's various integration options, including REST APIs, when the ThingsBoard installation is run as server. This connectivity flexibility ensures seamless communication between your devices and the platform, finally from a cost perspective, it offers a cost-effective solution, particularly suitable for small to medium-sized agricultural IoT projects, without incurring expensive licensing fees, making it an appealing choice for budget-conscious projects.

Mosquitto, an MQTT broker, is an excellent choice for this project due to its lightweight and efficient design, making it well-suited for resource constrained IoT devices. It is cross-

platform, compatible with various operating systems and ensures flexibility in deployment. It has reliable message delivery through MQTT's Quality of Service levels, offering message assurance, given its ability to retain and persist data, makes it a plus for preventing data loss, especially in cases of intermittent network connectivity. Additionally, it integrates easily with various IoT platforms, including ThingsBoard, streamlining communication and data flow between the simulated sensors and the IoT platform. These qualities collectively make Mosquitto an excellent choice for managing data in monitoring and control systems [29].

Docker plays also plays an interesting role for several compelling reasons as it offers a layer of isolation, ensuring that the simulated sensor scripts and the Mosquitto MQTT broker run independently in their own containers, creating stability, and preventing conflicts between different components. It ensures that the IoT solution behaves predictably across various environments, from development and testing to production. Since one of Docker's key advantages is its ease of setup and reproducibility, integrates perfectly with image control and continue delivery, it can horizontally scale the IoT solution by adding more container instances. Deploying this project is made notably more straightforward with Docker launching, the solution becomes simple, greatly enhancing convenience. To conclude, Docker's infrastructure-agnostic nature is a final point of strength, enables the images to run on diverse infrastructure platforms, encompassing on-premises servers, cloud providers, and edge devices [30].

## 3.3 Implementation

In this section, the practical implementation of the "Soil IoT" platform is detailed. It outlines the steps taken to transform the conceptual framework into a functional system. This chapter serves as a comprehensive account of the project's execution and is intended to provide a practical guide to implement similar precision agriculture initiatives.

Starting with the database, there are two scripts that are responsible for outlier removal and data analysis. After this data undergoes a cleaning process, it is inserted in a database. This database will serve as a pivoting point, to run predictive models, as well as to create reliable synthetic data with the aid of predictive modeling. The data is subsequently transmitted through a MQTT broker, which arranges and disseminates information based on subjects that ThingsBoard is configured to monitor. Every device type or profile on

ThingsBoard is set up to listen to a specific MQTT topic and receive telemetry which is JSON-formatted data that is then saved in a database.

This data flow is further controlled and improved by a strong rule chain mechanism in ThingsBoard as it can be observed in figure 5, the rule chain serves as a collection of modular processing nodes that carry out tasks including filtering, transformation, and conditional routing. The platform can efficiently prioritize and manage incoming information, automate responses, and initiate its own alerts or actions based on predetermined criteria by using this methodical approach. The system's scalability is improved by this dynamic and flexible rule chain structure, which enables sophisticated decision-making processes and improved real-time data handling.

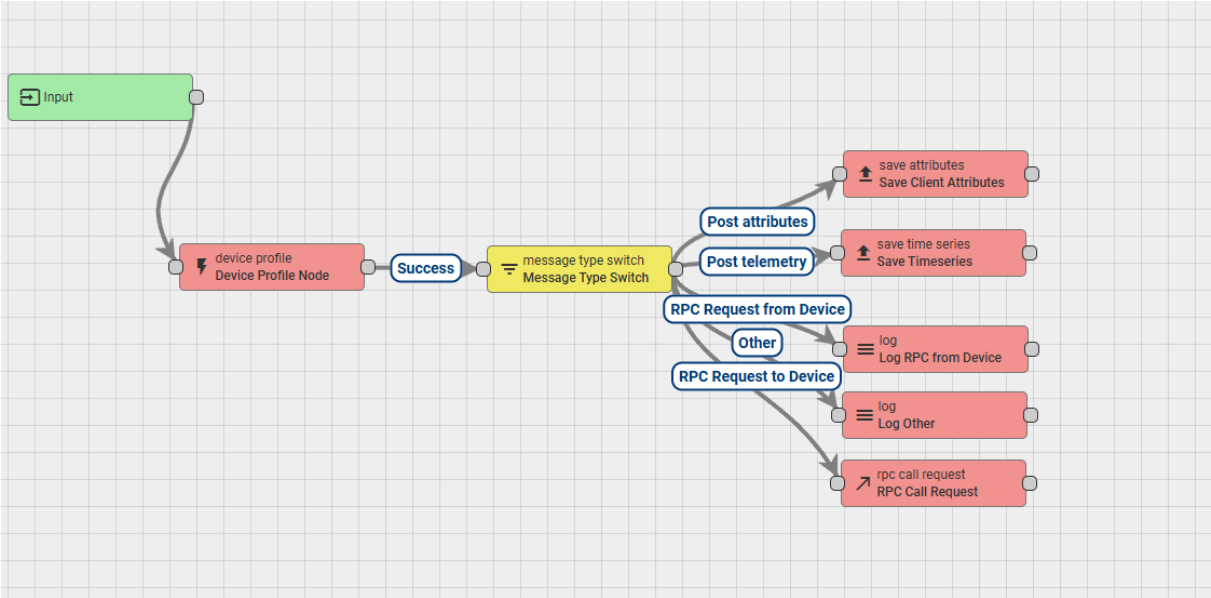



Figure 5 - Rule chain from Thingsboard platform

# Chapter 4
# **Data Analysis**

This chapter focuses on the process of data visualization and the insights drawn from the data before it was deployed to the dashboard and widgets. Emphasis is placed on the integral role that the data played in the Soil IoT project and the essential steps of pre-processing that were undertaken to ensure its accuracy and reliability. The processed data lays the groundwork for generating synthetic data and developing predictive models, both of which are essential in providing real-time alerts and notifications to the platform, aiding decision-making, and soil management.

## 4.1 Dataset

The dataset utilized in this project was sourced from Purdue University's Digital Agriculture Group through the Wabash Heartland Innovation Network (WHIN) initiative [32] where WHIN focuses on supporting data-driven agricultural innovation across Indiana and offers a vast range of high-quality agricultural data to support research and development efforts. The specific dataset used in this project contains sensor readings for several key variables, including temperature, humidity, soil moisture, soil temperature and soil conductivity. These variables provide essential insights into soil conditions and help in assessing soil health and its direct impact on crop yields.

Each of these variables are truly important in understanding the environmental and physical factors that influence optimal farming practices both soybean and corn fields were represented in the dataset. There was a need to merge both type of crops provided due to the small size of the dataset, this merging of soybean and corn datasets allowed for a more comprehensive analysis of soil dynamics across various seasons, enabling an understanding of how factors like water retention, conductivity values and temperature variability are influenced.

Additionally, this merged dataset was essential for creating flexible, accurate predictive models allowing for the simulation of a range of soil conditions, enabling the platform to generate timely alerts and provide actionable insights to users. This is especially crucial in

precision agriculture, where resource management and decision-making heavily rely on real-time data.

| | time | battery_voltage | temperature | humidity | soil_sensor1_temperature | soil_sensor2_temperature | soil_sensor1_moisture | soil_sensor2_moisture | soil_sensor1_conductivity | soil_sensor2_conductivity | crop_type |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 190 | 2021-07-27T10:33:29 | 6,125501953 | 33,67 | 73,81 | 22,8 | 22,3 | 0,369 | 0,317 | 0,521 | 0,312 | soy |
| 191 | 2021-07-27T11:02:46 | 6,125501953 | 34,78 | 73,4 | 22,8 | 22,4 | 0,369 | 0,317 | 0,52 | 0,311 | soy |
| 192 | 2021-07-27T11:32:04 | 6,119753906 | 35,12 | 72,8 | 22,8 | 22,6 | 0,369 | 0,316 | 0,519 | 0,307 | soy |
| 193 | 2021-07-27T12:01:22 | 6,119753906 | 35,96 | 70,7 | 17,9 | 23,1 | 0,368 | -0,001 | 0,518 | 0,821 | soy |
| 194 | 2021-07-27T12:30:39 | 6,117837891 | 36,34 | 70,34 | 22,7 | 23,1 | 0,368 | 0,315 | 0,509 | 0,307 | soy |
| 195 | 2021-07-27T12:59:57 | 6,117837891 | 36,77 | 69,14 | 22,7 | 23,4 | 0,367 | 0,315 | 0,516 | 0,309 | soy |
| 196 | 2021-07-27T13:29:14 | 6,106341797 | 36,13 | 69,7 | 22,7 | 23,8 | 0,367 | 0,314 | 0,514 | 0,31 | soy |
| 197 | 2021-07-27T13:58:32 | 6,106341797 | 35,61 | 69,84 | 22,7 | 24,2 | 0,367 | 0,313 | 0,514 | 0,306 | soy |
| 198 | 2021-07-27T14:27:50 | 6,106341797 | 37,57 | 67,17 | 21,5 | 24,6 | 0,366 | 0,313 | 0,514 | 0,308 | soy |
| 199 | 2021-07-27T15:26:25 | 6,127417969 | 36,74 | 70,65 | 22,9 | 25,3 | 0,366 | 0,312 | 0,512 | 0,3 | soy |
| 200 | 2021-07-27T15:35:15 | 2,939167969 | -240,2 | 629,9 | 815,5 | 2839,5 | 1,389 | 1,379 | 25,707 | 43,4 | soy |
| 201 | 2021-07-27T15:55:42 | 6,114005859 | 32,26 | 75,23 | 4919,3 | 25,6 | 0,365 | 0,311 | 0,513 | 0,308 | soy |
| 202 | 2021-07-27T16:25:00 | 6,114005859 | 34,36 | 72,26 | 23 | 25,8 | 0,365 | 0,311 | 0,513 | 0,309 | soy |
| 203 | 2021-07-27T16:54:18 | 6,112089844 | 34,47 | 72,58 | 23,1 | 25,9 | 0,365 | 0,31 | 0,509 | 0,303 | soy |
| 204 | 2021-07-27T17:23:35 | 6,112089844 | 33,93 | 73,36 | 23,2 | 26 | 0,365 | 0,31 | 0,512 | 0,309 | soy |
| 205 | 2021-07-27T17:52:53 | 6,115921875 | 31,64 | 75,63 | 23,3 | 26,1 | 0,365 | 0,31 | 0,512 | 0,309 | soy |

Figure 6 - Sample from merged MergedCornSoyData in database

The acquired data are expressed by essential agricultural parameters such as:

- Time: The timestamp when each reading was taken.
- Battery Voltage: Reflecting the battery status of the sensors.
- Temperature (°C): Soil surface temperature.
- Humidity (%): Air humidity near the sensor.
- Soil Sensor 1 and 2 Temperature (°C): Measuring soil temperature at different points.
- Soil Moisture Sensor 1 and 2 (VWC/%): Measuring volumetric water content to understand soil moisture levels.
- Soil Conductivity Sensor 1 and 2 (µS/cm): Measuring soil conductivity to provide insights into nutrient availability and soil health.
- Crop Type (soy or corn): Identifying the crop growing during the data collection period.

The MergedCornSoyData table consisted of 14,334 rows which is the totality of the Purdue University's Digital Agriculture Group 2021 dataset, considering that all rows were merged and that each row represents a unique set of readings for the specific environmental variables at the time of data capture. This raw data set was directly loaded into the database from CSV files obtained from Purdue University's Digital Agriculture Group, following WHIN (Wabash Heartland Innovation Network) guidelines.

## 4.2 Exploratory Data Analysis (EDA) and Data Pre-Processing

Exploratory Data Analysis (EDA) is essential as it involves investigating the characteristics, relationships, and underlying patterns within a dataset before performing advanced analyses such as predictive modeling. In the context of the Soil IoT project, EDA provided valuable insights into the environmental data collected through sensors over several months. This data, consisting of variables like temperature, humidity, soil moisture, and conductivity, served as the foundation for building predictive models and driving actionable insights.

### 4.2.1 Purpose of EDA

The purpose of conducting EDA in this project was to understand the characteristics of the sensor data in a comprehensive manner before proceeding with advanced analytical methods like predictive modeling. By leveraging a combination of statistical techniques and data visualization tools, this analysis allows the identification of important patterns and trends, such as variations in soil moisture across different periods like the cyclical nature of temperature changes throughout the day. Additionally, EDA helped to explore potential relationships between environmental factors, including the effect of temperature on soil moisture retention. This investigation also aims to detect anomalies or outliers in the dataset, such as erroneous sensor readings, which could negatively affect the accuracy of the analysis.

### 4.2.2 Data Preparation for EDA

The data collected from sensors available in the dataset was initially raw and required several pre-processing steps to ensure its integrity and usability for analysis, so one of the first challenges encountered was the handling of missing data, which occurred due to big discrepancy in measured values most likely due to intermittent sensor failures or network issues. To address this, missing values were inserted using the median of the respective variable, ensuring that the dataset remained robust while avoiding any bias introduced by extreme values. This method of median imputation was chosen as it is less sensitive to outliers compared to mean imputation, preserving the central tendency of the data without distorting the results [33].

Another key aspect of pre-processing involved the identification and removal of outliers, since that extreme outliers were found to be present in various sensor readings, particularly in temperature and soil moisture values, so the Interquartile Range (IQR) method was employed to detect and eliminate these outliers [34]. By testing two different thresholds k=1.5

and k=1.0 the approach was refined to settle on k=1.5, which provided a balanced approach that removed erroneous data points while preserving the natural variability inherent in agricultural sensor readings because k = 1.0 eliminated many reliable data points.

### 4.2.3 Outlier Detection and Treatment

One of the most critical findings during the EDA process was the identification and treatment of outliers. The dataset contained significant outliers due to inconsistent data, especially in the soil temperature and soil moisture readings. As referenced in the anterior section, these outliers were detected using the Interquartile Range (IQR) method, with k=1.5 standard, selected as the optimal multiplier normalizing the dataset, by removing outliers such as temperatures below -100°C and above 100°C, which were physically impossible given the provided data. After processing the removal of outliers, the remaining data provided a realistic range of temperatures (10°C - 40°C), as shown in the figures 7 and 8, where the same comparisons can be observed in the other attributes.

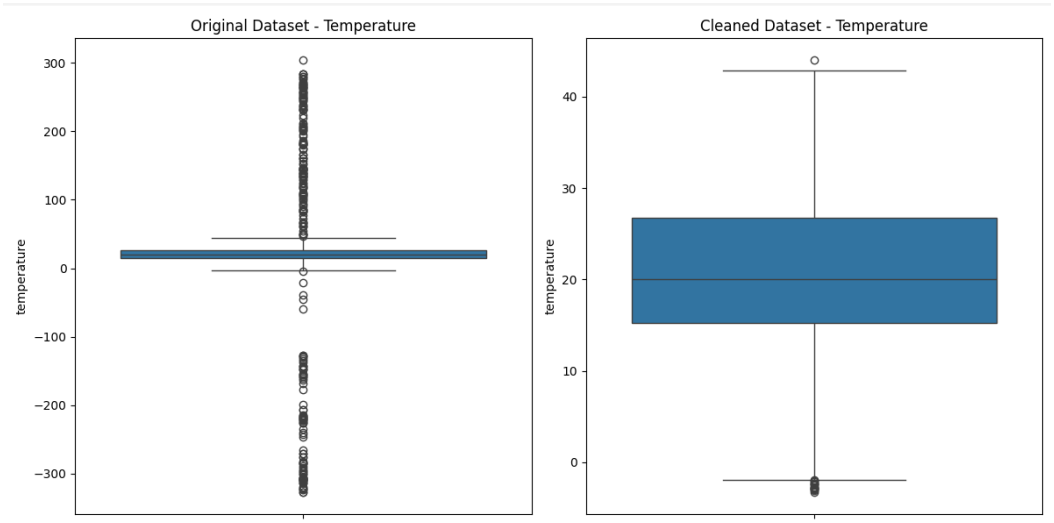



Figure 7 – Whisker plot comparison between datasets MergedCornSoyData and PreparedData regarding temperature

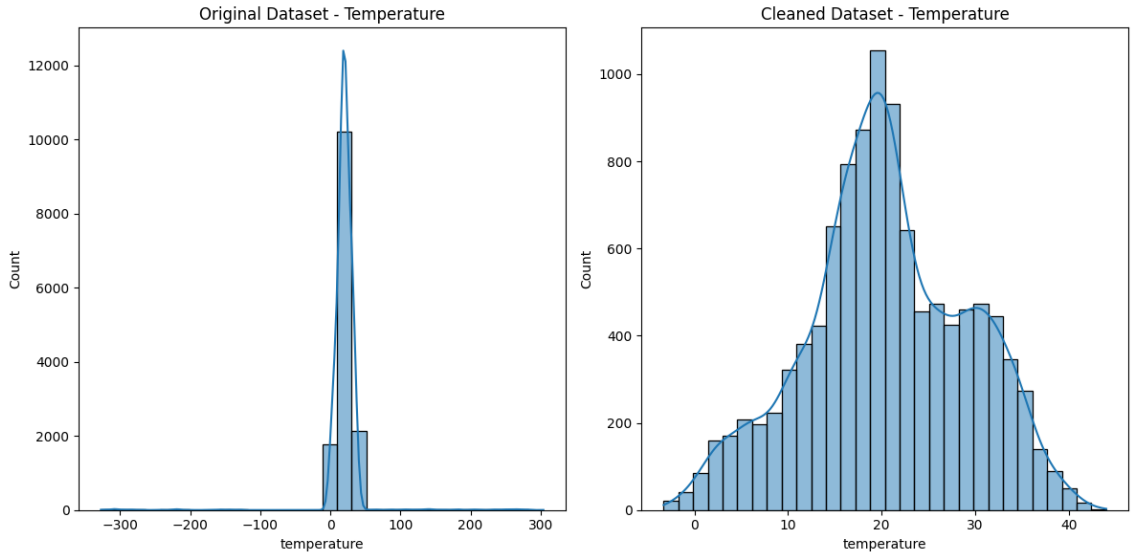



Figure 8 – Histogram comparison between datasets MergedCornSoyData and PreparedData regarding temperature

The same behavior can be observed on the temperature histogram, figure 8, where its observable a greater distribution in different values, the same goes for the whisker plot where abnormal outlier values were removed.

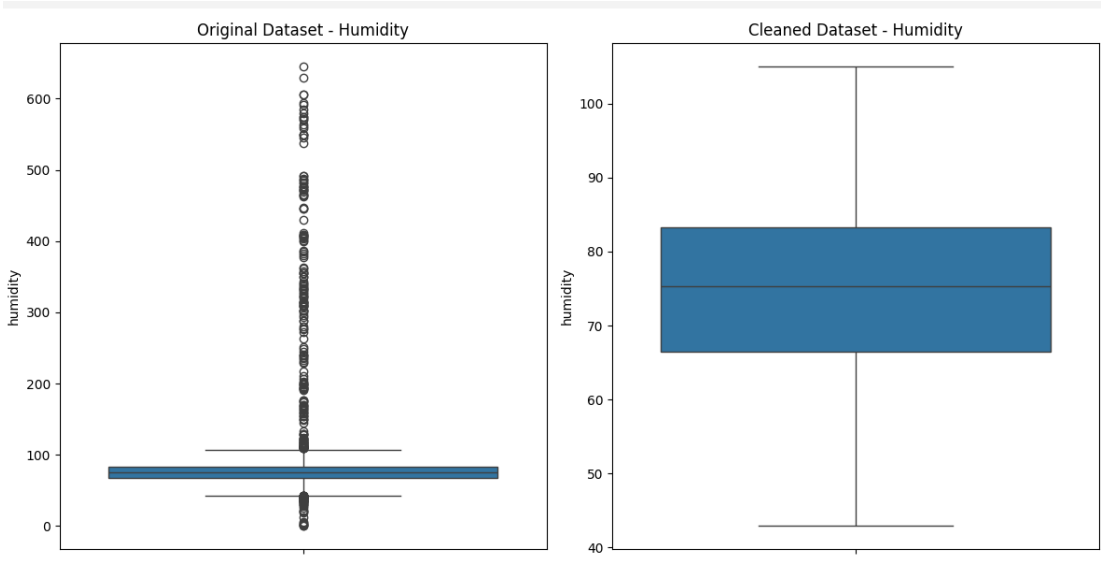



Figure 9- Whisker plot comparison between datasets MergedCornSoyData and PreparedData regarding humidity

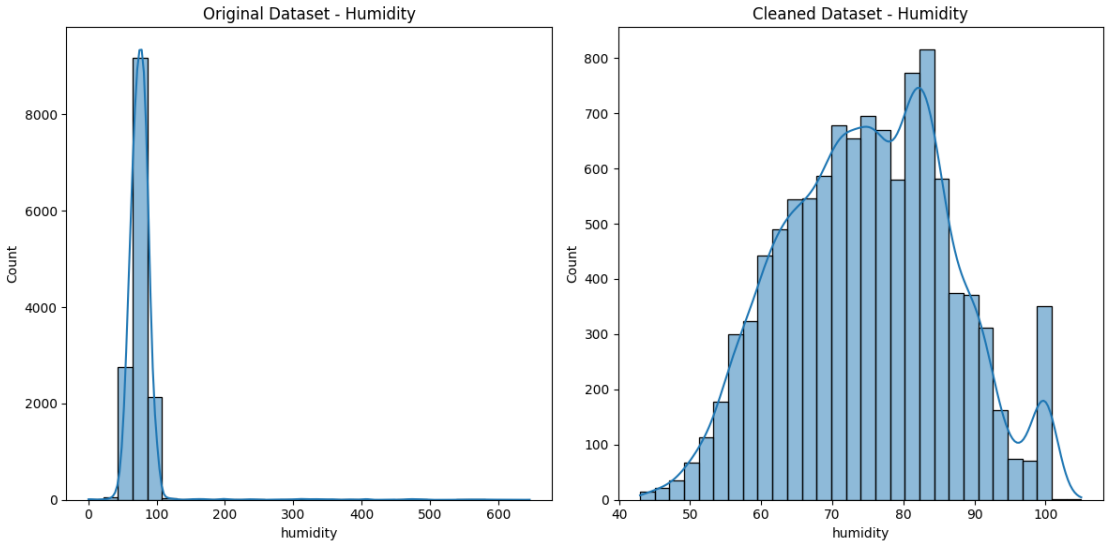



Figure 10 - Histogram comparison between datasets MergedCornSoyData and PreparedData regarding humidity

The comparison of soil moisture, specifically the sensor parameter soil_sensor1_moisture, for the two datasets, the original, and the cleaned, shows the extent of the cleaning procedures to eliminate noise and outliers from the data. In the whisker plots in Figure 11, it can be observed that for the original dataset, the ratio of average to substance outlier points is also greater, which indicates inconsistency of values in terms of soil moisture amounts. Such outliers add noise to the data and compromise the integrity of any analysis or conclusion based on it. Once cleaning has occurred, there are no more extreme outliers on the range of whisker plot (right), which means that the data cleaning process has improved the quality of soil moisture values such as whippers. In the histograms in Figure 12, the difference is clearly noticeable where the original dataset (left) has a very narrow distribution range and an over peaked area on a particular value.

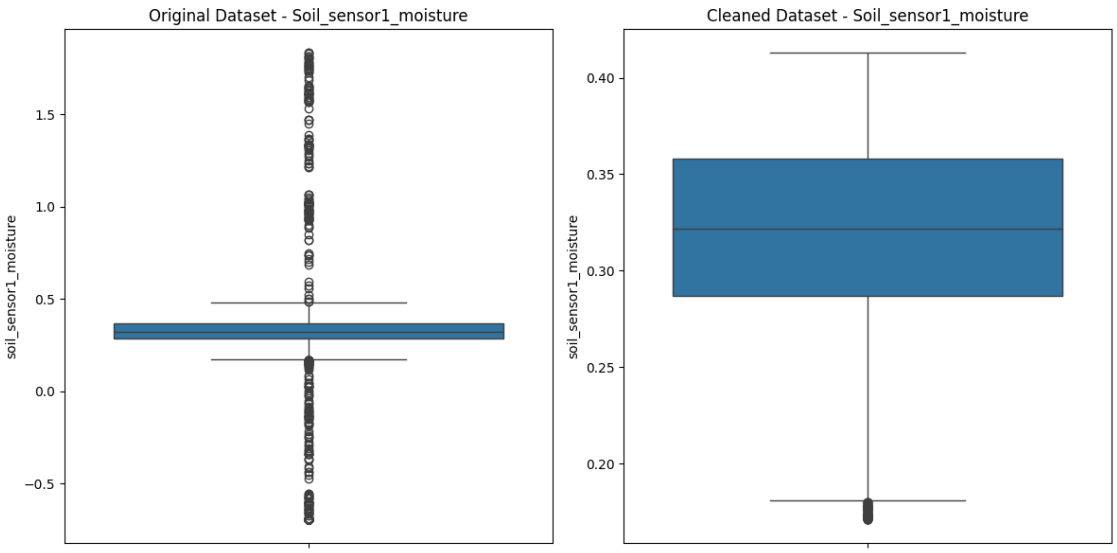



Figure 11 – Whisker plot comparison between datasets MergedCornSoyData and PreparedData regarding soil_sensor1_moisture

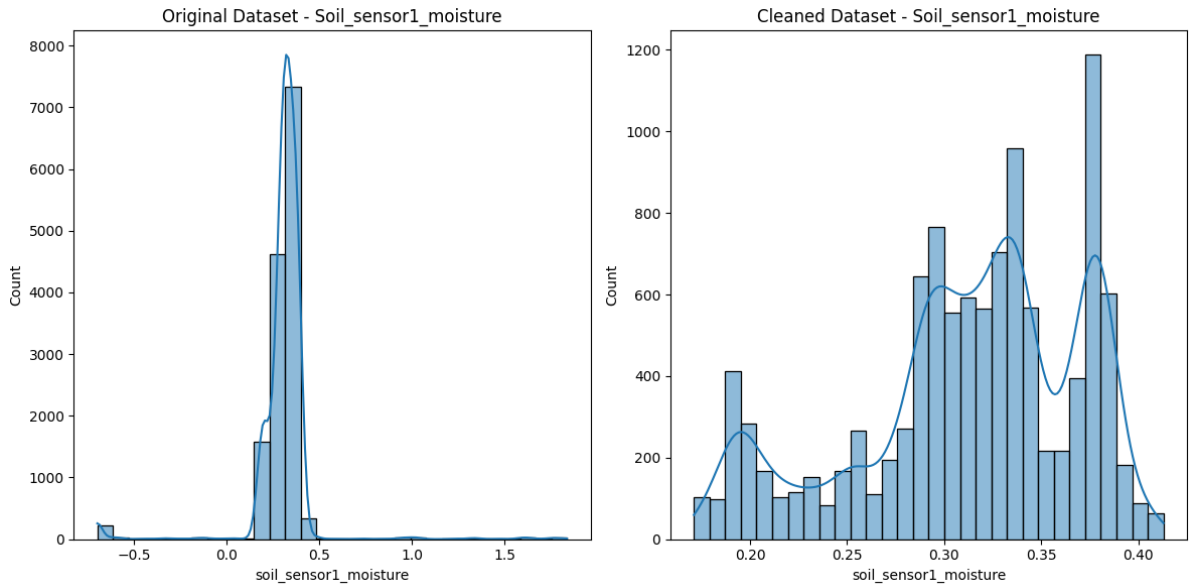



Figure 12 – Histogram comparison between datasets MergedCornSoyData and PreparedData regarding soil_sensor1_moisture

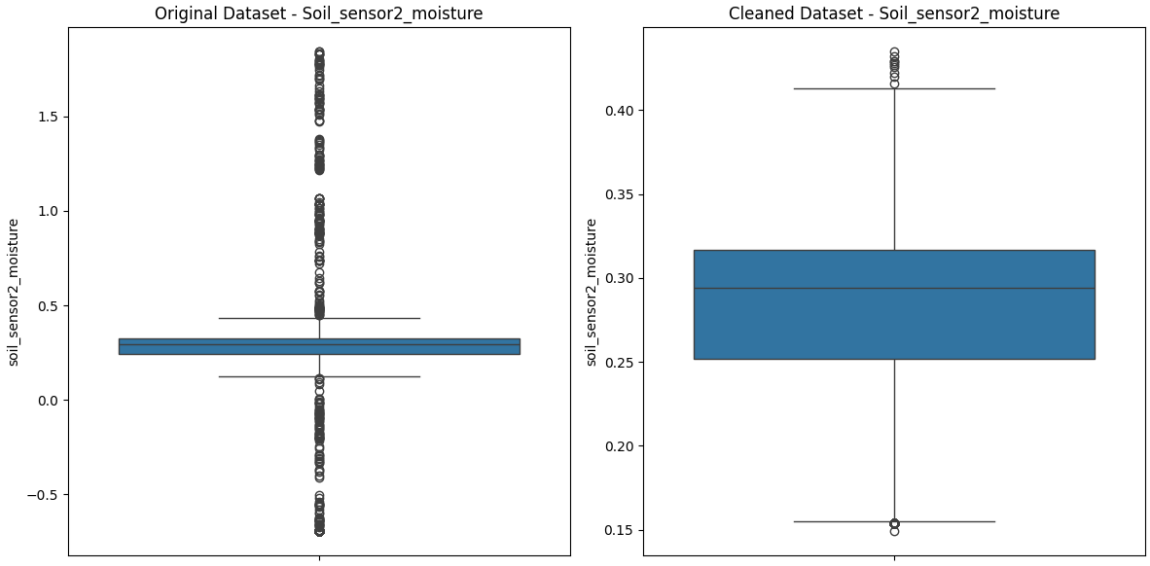



Figure 13- Comparison between datasets MergedCornSoyData and PreparedData regarding soil_sensor2_moisture

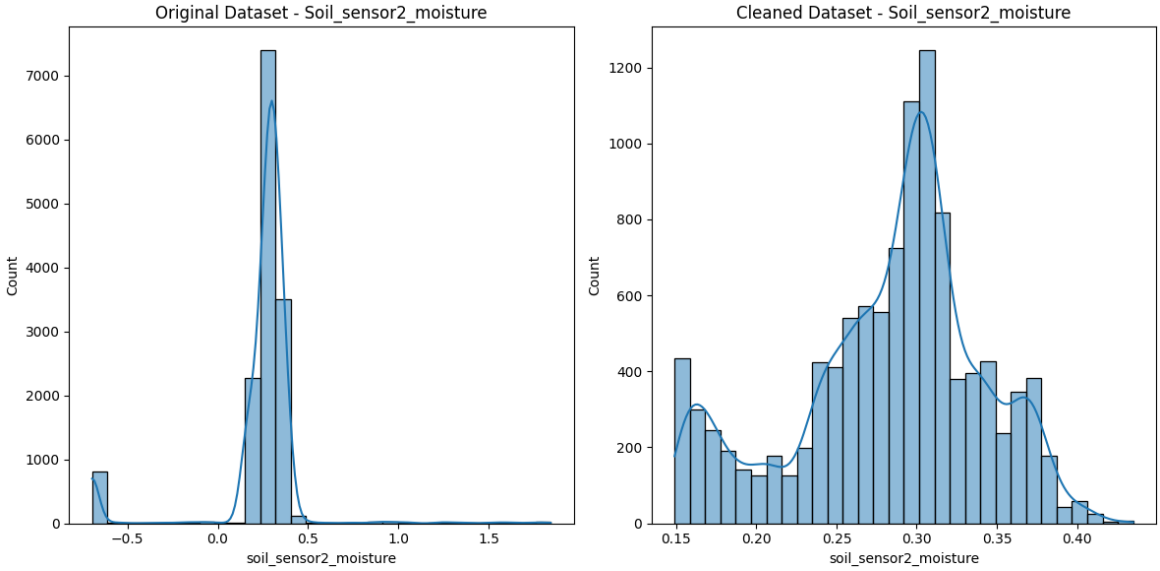
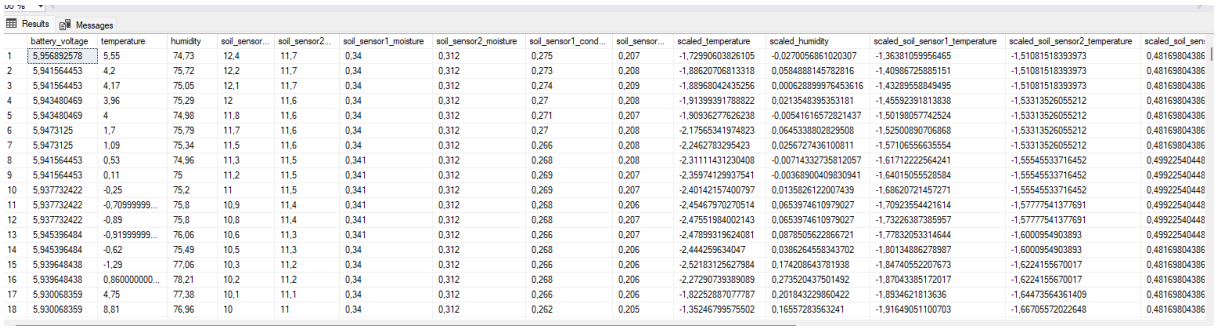



Figure 14 - Data count between datasets MergedCornSoyData and PreparedData regarding soil_sensor2_moisture

Some outliers in humidity were also removed which ensured that the data reflected the true soil conditions more accurately, as can be better observed in all whisker plots (Figures 7,9,11 and 13, after processing the data is distributed in real values. Additionally, unit conversion was necessary for consistency across the dataset, particularly for temperature readings. Sensor data initially recorded temperature in Fahrenheit, therefore, was converted to Celsius to align with the standard units used for analysis in other environmental factors.



Results Messages

| | battery_voltage | temperature | humidity | soil_sensor... | soil_sensor2... | soil_sensor1_moisture | soil_sensor2_moisture | soil_sensor1_cond... | soil_sensor... | scaled_temperature | scaled_humidity | scaled_soil_sensor1_temperature | scaled_soil_sensor2_temperature | scaled_soil_sen |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 5,956892578 | 5,55 | 74,73 | 12,4 | 11,7 | 0,34 | 0,312 | 0,275 | 0,207 | -1,72990603826105 | -0,027005686102030 | -1,36381059956465 | -1,51081518393973 | 0,48169804386 |
| 2 | 5,941564453 | 4,2 | 75,72 | 12,2 | 11,7 | 0,34 | 0,312 | 0,273 | 0,208 | -1,88620706813318 | 0,0584888145782816 | -1,40986725885151 | -1,51081518393973 | 0,48169804386 |
| 3 | 5,941564453 | 4,17 | 75,05 | 12,1 | 11,7 | 0,34 | 0,312 | 0,274 | 0,209 | -1,88968042435256 | 0,000628899976453616 | -1,43289558849495 | -1,51081518393973 | 0,48169804386 |
| 4 | 5,943480469 | 3,96 | 75,29 | 12 | 11,6 | 0,34 | 0,312 | 0,27 | 0,208 | -1,91399391788822 | 0,0213548395353181 | -1,45592391813838 | -1,53313526055212 | 0,48169804386 |
| 5 | 5,943480469 | 4 | 74,98 | 11,8 | 11,6 | 0,34 | 0,312 | 0,271 | 0,207 | -1,90936277626238 | -0,00541616572821437 | -1,50198057742524 | -1,53313526055212 | 0,48169804386 |
| 6 | 5,9473125 | 1,7 | 75,79 | 11,7 | 11,6 | 0,34 | 0,312 | 0,27 | 0,208 | -2,17565341974823 | 0,0645338802829508 | -1,52500890706868 | -1,53313526055212 | 0,48169804386 |
| 7 | 5,9473125 | 1,09 | 75,34 | 11,5 | 11,6 | 0,34 | 0,312 | 0,266 | 0,208 | -2,2462783295423 | 0,0256727436100811 | -1,57106556635554 | -1,53313526055212 | 0,48169804386 |
| 8 | 5,941564453 | 0,53 | 74,96 | 11,3 | 11,5 | 0,341 | 0,312 | 0,268 | 0,208 | -2,31111431230408 | -0,00714332735812057 | -1,61712222564241 | -1,55545533716452 | 0,49922540448 |
| 9 | 5,941564453 | 0,11 | 75 | 11,2 | 11,5 | 0,341 | 0,312 | 0,269 | 0,207 | -2,35974129937541 | -0,00368900409830941 | -1,64015055528584 | -1,55545533716452 | 0,49922540448 |
| 10 | 5,937732422 | -0,25 | 75,2 | 11 | 11,5 | 0,341 | 0,312 | 0,269 | 0,207 | -2,40142157400797 | 0,0135826122007439 | -1,68620721457271 | -1,55545533716452 | 0,49922540448 |
| 11 | 5,937732422 | -0,70999999... | 75,8 | 10,9 | 11,4 | 0,341 | 0,312 | 0,268 | 0,206 | -2,45467970270514 | 0,0653974610979027 | -1,70923554421614 | -1,57777541377691 | 0,49922540448 |
| 12 | 5,937732422 | -0,89 | 75,8 | 10,8 | 11,4 | 0,341 | 0,312 | 0,268 | 0,207 | -2,47551984002143 | 0,0653974610979027 | -1,73226387385957 | -1,57777541377691 | 0,49922540448 |
| 13 | 5,945396484 | -0,91999999... | 76,06 | 10,6 | 11,3 | 0,341 | 0,312 | 0,266 | 0,207 | -2,47899319624081 | 0,0878505622866721 | -1,77832053314644 | -1,6000954903893 | 0,49922540448 |
| 14 | 5,945396484 | -0,62 | 75,49 | 10,5 | 11,3 | 0,34 | 0,312 | 0,268 | 0,206 | -2,44425963404 7 | 0,0386245583437 02 | -1,80134886278987 | -1,6000954903893 | 0,48169804386 |
| 15 | 5,939648438 | -1,29 | 77,06 | 10,3 | 11,2 | 0,34 | 0,312 | 0,266 | 0,206 | -2,52183125627984 | 0,174208643781938 | -1,84740552207673 | -1,62241556670017 | 0,48169804386 |
| 16 | 5,939648438 | 0,860000000... | 78,21 | 10,2 | 11,2 | 0,34 | 0,312 | 0,268 | 0,206 | -2,27290739389089 | 0,273520437501492 | -1,87043385172017 | -1,62241556670017 | 0,48169804386 |
| 17 | 5,930068359 | 4,75 | 77,38 | 10,1 | 11,1 | 0,34 | 0,312 | 0,266 | 0,206 | -1,82252887077787 | 0,201843229860422 | -1,8934621813636 | -1,64473564361409 | 0,48169804386 |
| 18 | 5,930068359 | 8,81 | 76,96 | 10 | 11 | 0,34 | 0,312 | 0,262 | 0,205 | -1,35246799575502 | 0,16557283563241 | -1,91649051100703 | -1,66705572022648 | 0,48169804386 |

Figure 15 - Sample from PreparedData after arranging data

The PreparedData table is the cleaned and transformed version of the original MergedCornSoyData. The data underwent several preprocessing steps, resulting in 10,826 rows after removing outliers, handling missing data, and standardizing relevant variables.

## 4.3 Univariate Analysis

Univariate is a term commonly used in statistics to describe a type of data which consists of observations on only a single characteristic or attribute. A simple example of univariate data would be the salaries of workers in industry [31].

### 4.3.1 Distribution of Temperature, Humidity, and Soil Moisture

The temperature readings, after removing outliers, followed a normal distribution, centered around 20°C to 30°C, which is typical for agricultural fields during the growing season, the plot also include 24-point rolling means (represented by the red lines). These rolling averages are meant to help smoothing out fluctuations to reveal underlying trends in the data. As can be observed in the original dataset, the rolling mean line appears mostly stable, but it does not accurately reflect the extreme spikes in temperature, indicating that the outliers had a minimal effect on the mean due to their sporadic nature.

In contrast, in the cleaned dataset, the rolling mean follows the general seasonal trend, capturing the gradual increase and decrease in temperature that aligns with seasonal changes, where regarding the cleaned dataset it exhibits a clear cyclic pattern, with temperatures gradually increasing from spring to summer, reaching their peak, and then starting to decrease.

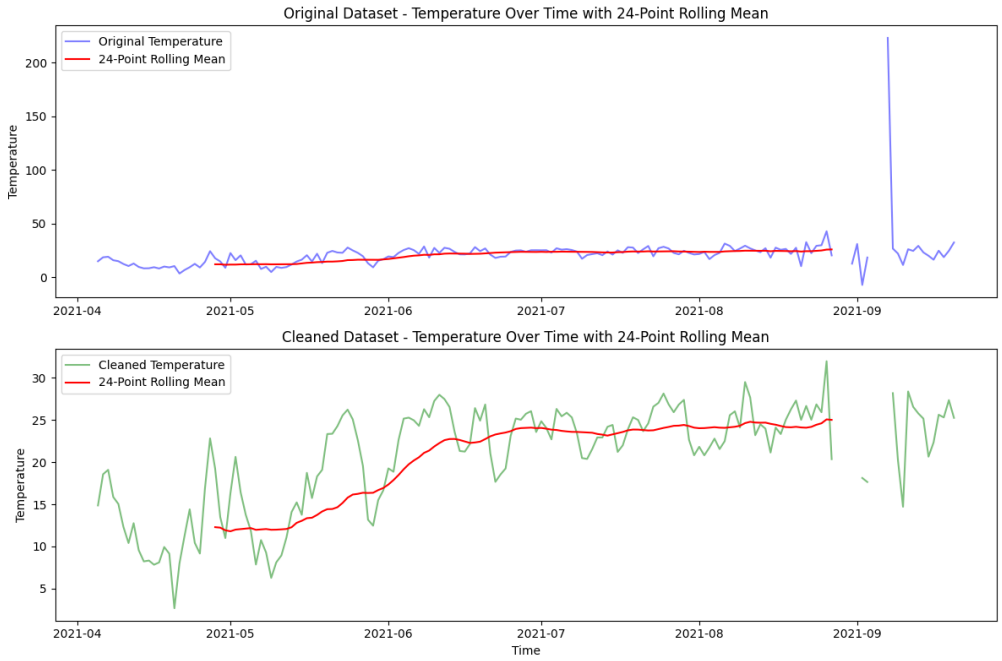



Figure 16 - Comparison between datasets using 24 Point Rolling Mean regarding temperature [36]

After outlier removal, temperature distribution appears normal, centered around 20°C to 30°C, with a mean of 20.49°C and a standard deviation of 8.64. The skewness (-0.07) and kurtosis (-0.35) indicate a near-normal distribution, and Hartigan's Dip Test reveals no significant evidence of bimodality (p-value = 0.9031). These results suggest a consistent temperature regime, typical for fields experiencing seasonal heating and cooling [37].

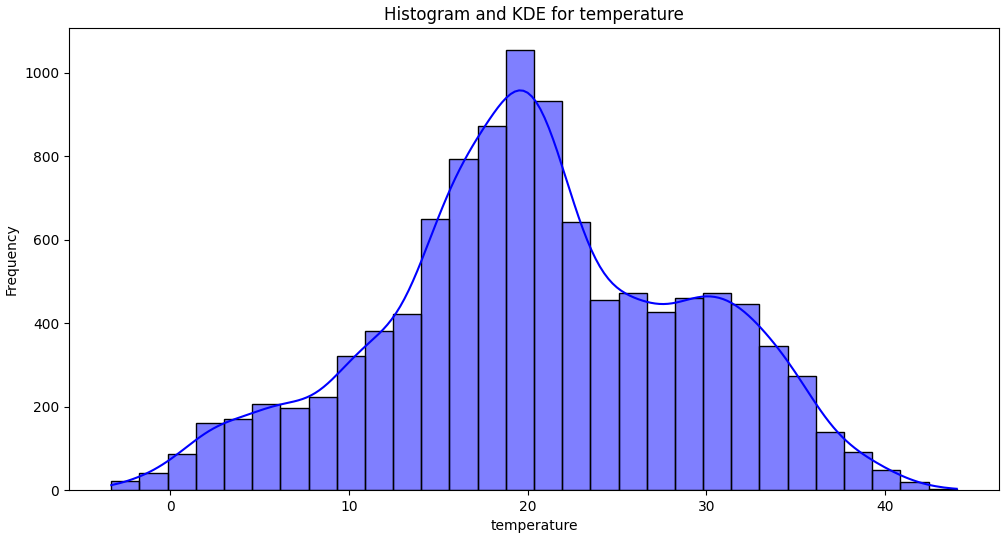



Figure 17 - Histogram and KDE for temperature

Humidity also displays a bimodal distribution, with significant evidence of two distinct states (Dip Value = 0.0138, p-value = 0.0000). This aligns where bimodal distributions were

reported to be observed in agricultural fields because of alternating wet and dry phases [38]. The mean humidity was 75.04% with low skewness (0.008), indicating a balanced distribution with periods of higher and lower humidity so it is concluded that the presence of a bimodal pattern is consistent with seasonal irrigation cycles or natural drying phases, which are characteristic in precision irrigation systems.

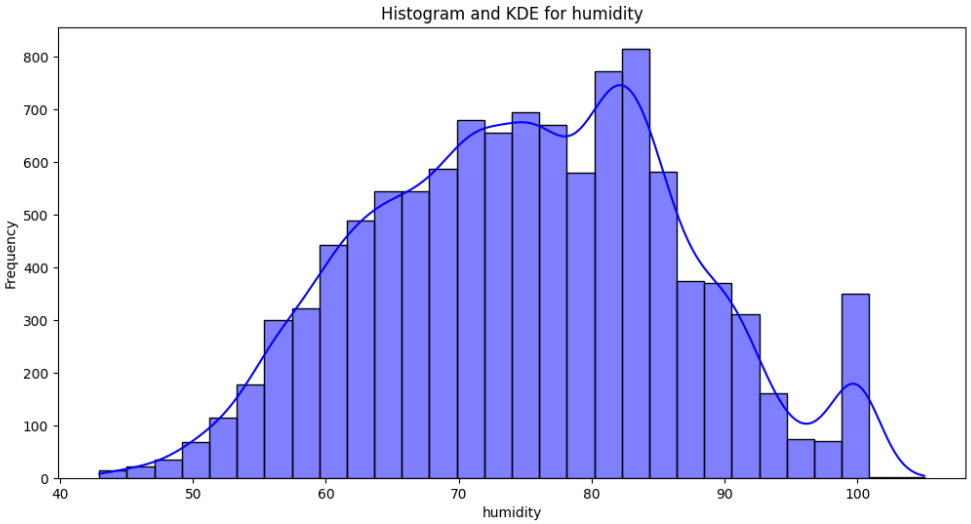



Figure 18 - Histogram and KDE for humidity

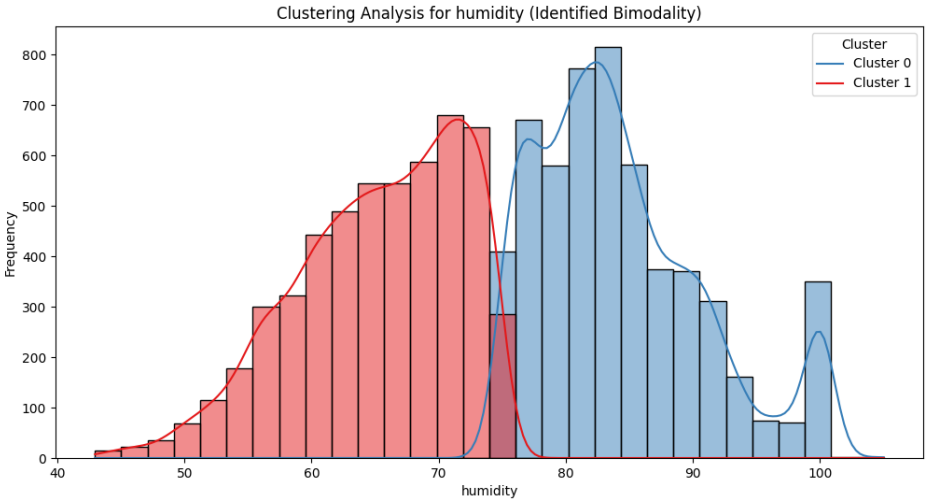



Figure 19 - Clustering analysis for humidity

The soil moisture readings from soil_sensor1_moisture and soil_sensor2_moisture demonstrated clear evidence of bimodality. This observation was statistically confirmed using Hartigan's Dip Test [39], which yielded p-values less than 0.0001 for both sensors, indicating statistically significant evidence of bimodality. Regarding the K Means applied, a centroid is a central point that serves as the center of a cluster in the feature space whereas, the clustering process starts with the selection of random centroids, after which each data point is assigned

to the closest centroid, grouping points according to closeness, It is determined by taking the average location of every data point in the cluster [40]. To minimize within-cluster variation and stabilize the data, it iteratively refines these centroids and reassigns points to clusters until cluster assignments remain constant. Since the data in this case showed a bimodal pattern, suggesting two primary categories, K=2 clusters were selected. The distribution's greater humidity mode is represented by one cluster (Cluster 0, shown in blue), while its lower humidity mode is represented by the other cluster (Cluster 1, shown in red).

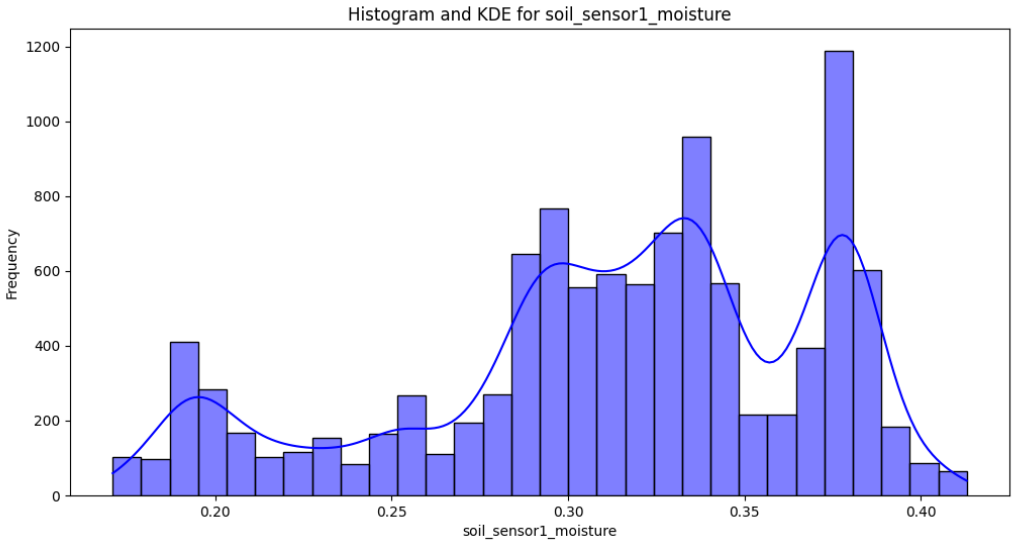



Figure 20 - Histogram and KDE for Soil Sensor 1 Moisture

The Histogram and KDE (Kernel Density Estimate) plot in Figure 20, shows two pronounced peaks, indicative of periods with high and low soil moisture. These peaks represent fluctuations in soil moisture content, likely due to irrigation or rain followed by evapotranspiration. As it can be observed from figure 20, 21, 22 and 23, histograms and KDE plots are effective for understanding soil moisture dynamics, especially in precision agriculture, where identifying shifts between moisture levels helps optimize irrigation strategies.

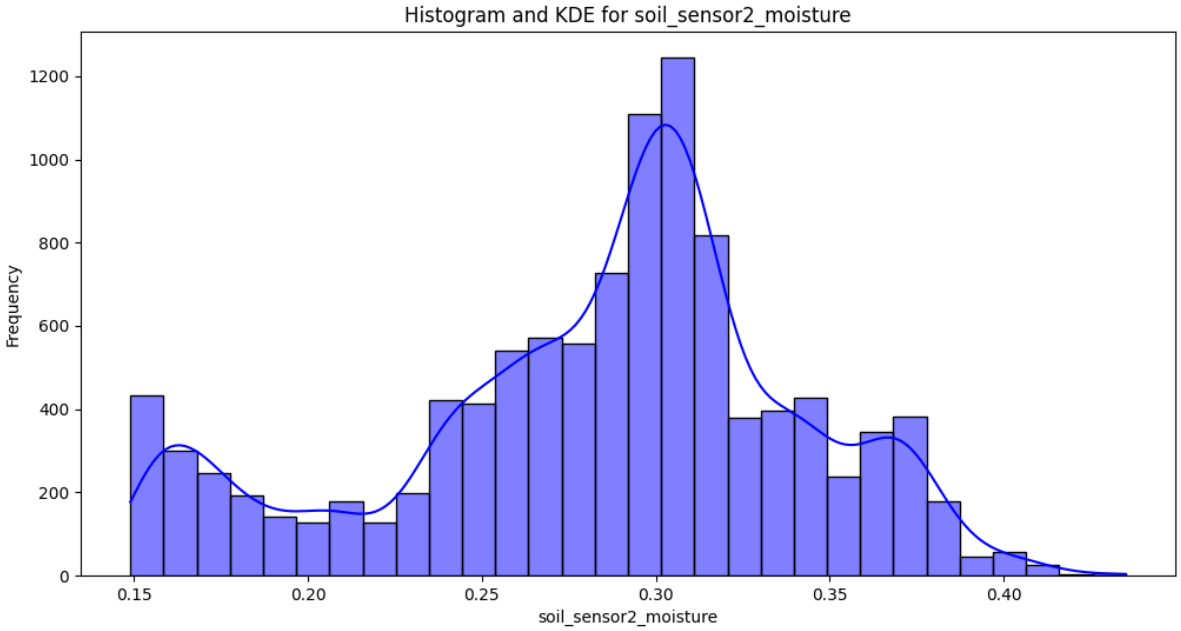



Figure 21 - Histogram and KDE for Soil Sensor 1 Moisture

To further investigate the bimodal distribution observed in soil moisture, K-Means Clustering was implemented using the K-Means function from the scikit-learn library, and visualizations were created with Matplotlib and Seaborn to segment the data into two clusters representing high and low soil moisture states.

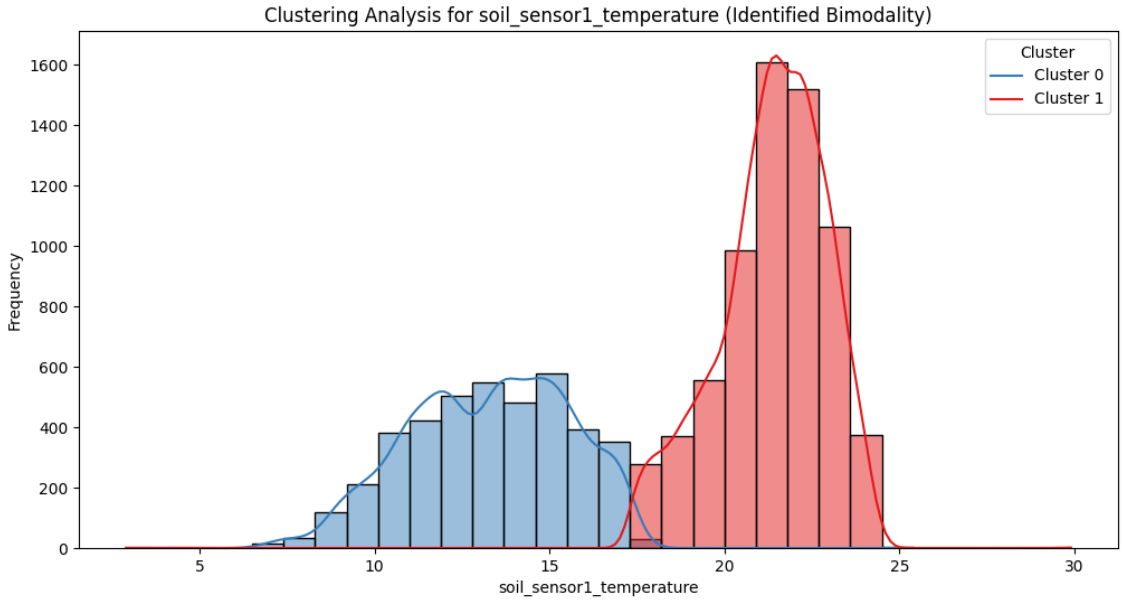



Figure 22 - Clustering analysis for Soil Sensor1 Moisture

Regarding the K-Means, the targeted customization attempts to determine the dividing line between the soil temperature data, which in turn relates to the soil moisture content, particularly when temperature is an influencing factor in moisture retention in your analysis of the setup.

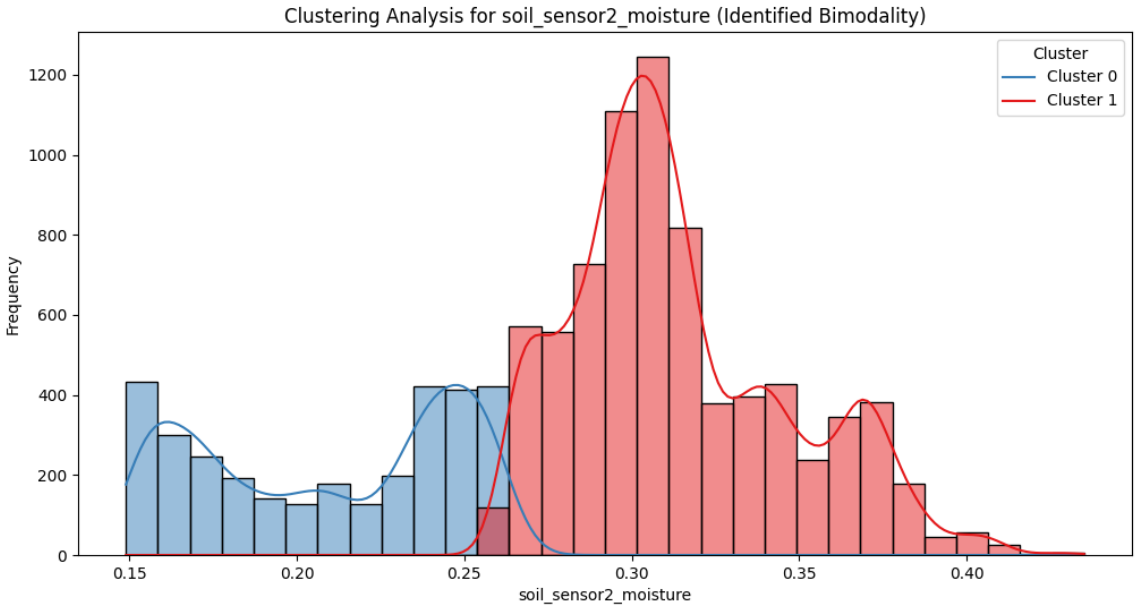



Figure 23 - Clustering analysis for Soil Sensor2 Moisture

This bimodal pattern aligns with findings from recent research, where bimodal distributions in soil moisture have been observed in irrigated agriculture settings, primarily due to frequent and deliberate irrigation events alternating with drying phases [40].

The soil moisture readings for soil_sensor1 and soil_sensor2 show a clear bimodal distribution, with significant evidence from Hartigan's Dip:

Soil_sensor1_moisture: Dip Value = 0.0416, p-value = 0.0000

Soil_sensor2_moisture: Dip Value = 0.0215, p-value = 0.0000

With this it can be said that clustering analysis effectively differentiates between periods immediately following irrigation and periods during which soil moisture decreases due to plant water uptake and evaporation. Without explicit irrigation data in the analyzed dataset, the observed sharp peaks and valleys in soil moisture were interpreted as evidence of irrigation events followed by drying periods. Similar studies, have highlighted that in the absence of rainfall data, changes in soil moisture can serve as a reliable proxy to infer irrigation events whereas the peaks and subsequent decay in moisture, captured through bimodal distributions and K-Means clustering, reflect periods of increased water input followed by gradual water uptake by plants and evaporation [41].

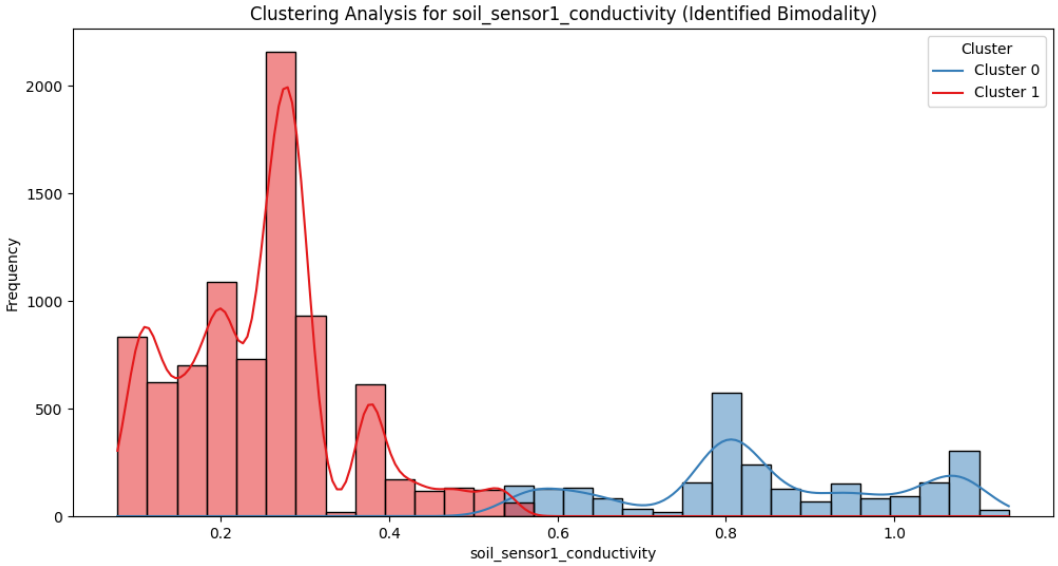



Figure 24 - Clustering analysis for Soil Sensor 1 Conductivity

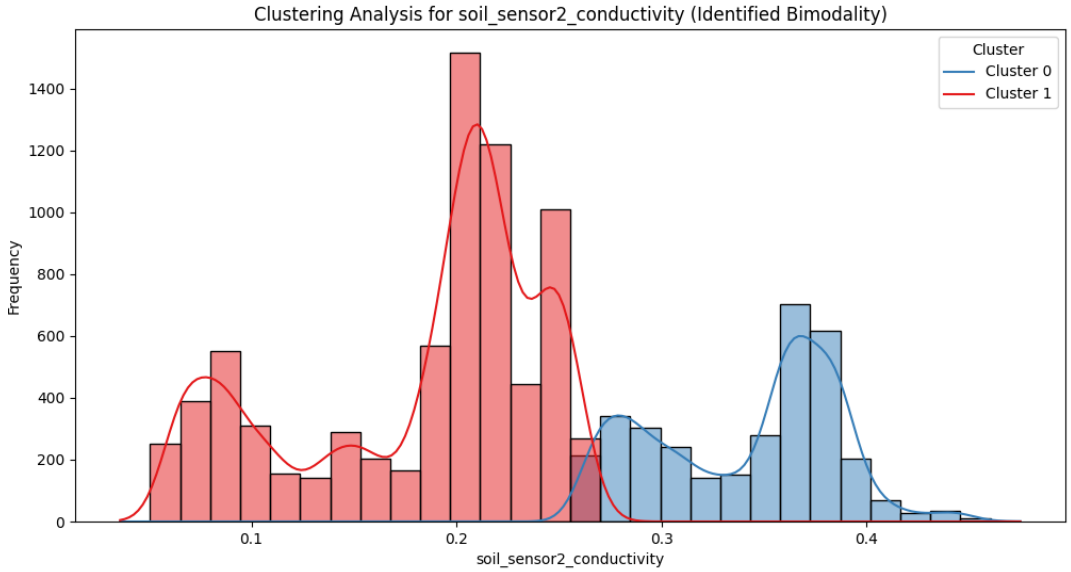



Figure 25- Clustering analysis for Soil Sensor 2 Conductivity

Soil conductivity, measured by soil_sensor1_conductivity and soil_sensor2_conductivity, also showed significant evidence of bimodality:

Soil_sensor1_conductivity: Dip Value = 0.0256, p-value = 0.0000

Soil_sensor2_conductivity: Dip Value = 0.0327, p-value = 0.0000

The bimodal distribution of soil conductivity is an indication of alternating periods of wet and dry soil conditions as agricultural fields, soil conductivity tends to increase with wet conditions due to increased ion mobility and decrease during dry periods. Such behavior is

expected in irrigated fields, where salinity changes with each irrigation cycle [42], thus highlighting the importance of monitoring soil conductivity for effective nutrient and salinity management, especially in precision agriculture. These results highlight the cyclical nature of soil moisture and the usefulness of statistical methods such as K-Means clustering and Hartigan's Dip Test in recognizing various moisture levels in an agricultural setting. Compared to simple rolling means, these techniques offer greater clarity by emphasizing clear behavioral patterns in the data. Given the above statements, it is observed that distinguishing between different states of soil moisture is beneficial for targeted irrigation strategies, leading to improved agricultural outcomes by utilizing methods like clustering and statistical tests provides a deeper understanding of soil moisture dynamics, enabling more informed decisions for irrigation scheduling [43].

### 4.3.2 Summary Statistics

Key summary statistics for variables like temperature, humidity, soil moisture, and conductivity are shown in Tables 3 and 4 both before and after the data preprocessing step. Making educated agricultural decisions requires an extensive understanding of the data distribution, variability, and underlying tendencies, all of which are provided by these statistics. While the median provides a strong measure that is resistant to outliers, the mean is more sensitive to extreme values, therefore including both measures aids in evaluating core tendencies. High skewness and kurtosis values further support the idea that there are outliers or extreme occurrences included in the original dataset, as evidenced by the notable discrepancies between the mean and median for variables like temperature and soil moisture. This underscores the importance of data cleaning, as extreme values can distort overall trends.

*Table 3 - Original Dataset Summary*

| Statistics | temperature | humidity | sensor1_tempera | sensor2_temperat | sensor1_moistu | il_sensor2_moistu | sensor1_conduct | soil_sensor2_conductivity |
|---|---|---|---|---|---|---|---|---|
| Mean | 20,032251 | 77,842921 | 85,948347 | 83,035238 | 0,307794 | 0,238854 | 1,096894 | 0,927614 |
| Median | 20,11 | 75,33 | 20,3 | 20,3 | 0,323 | 0,293 | 0,285 | 0,227 |
| Standard Deviation | 27,40061 | 33,01166 | 484,015556 | 481,170495 | 0,178716 | 0,259816 | 5,110802 | 5,067829 |
| Skewness | -3,164208 | 10,008095 | 8,346549 | 8,674619 | -0,487182 | -1,797715 | 8,371682 | 8,352899 |
| Kurtosis | 81,03922 | 124,061095 | 75,021308 | 81,639125 | 30,274278 | 10,568374 | 75,517139 | 75,010608 |
| Q1 (25th Percentile) | 14,6 | 66,9825 | 14,3 | 14,5 | 0,288 | 0,243 | 0,214 | 0,185 |
| Q3 (75th Percentile) | 26,76 | 83,25 | 22,1 | 22,2 | 0,366 | 0,324 | 0,583 | 0,317 |
| IQR (Interquartile Range) | 12,16 | 16,2675 | 7,8 | 7,7 | 0,078 | 0,081 | 0,369 | 0,132 |
| Min | -327,56 | 0,38 | 0 | 0 | -0,696 | -0,696 | 0 | 0 |
| Max | 304,13 | 645,46 | 6452,9 | 6537,1 | 1,835 | 1,844 | 65,512 | 65,422 |

*Table 4 - Cleaned Dataset Summary*

| Statistics | temperature | humidity | sensor1_tempera | _sensor2_temperat | _sensor1_moist | il_sensor2_moistu | sensor1_conduct | soil_sensor2_conductivity |
|---|---|---|---|---|---|---|---|---|
| Mean | 20,491508 | 75,042718 | 18,322317 | 18,468862 | 0,312517 | 0,283217 | 0,386054 | 0,232088 |
| Median | 20,075 | 75,28 | 20,2 | 20 | 0,322 | 0,294 | 0,279 | 0,221 |
| Standard Deviation | 8,637578 | 11,580227 | 4,342678 | 4,480478 | 0,057056 | 0,05866 | 0,276091 | 0,091881 |
| Skewness | -0,071669 | 0,008618 | -0,605043 | -0,549193 | -0,674796 | -0,518415 | 1,205354 | 0,058294 |
| Kurtosis | -0,345113 | -0,498484 | -0,953031 | -0,845939 | -0,281096 | -0,191346 | 0,233737 | -0,618918 |
| Q1 (25th Percentile) | 15,2625 | 66,4 | 14,5 | 14,9 | 0,287 | 0,252 | 0,201 | 0,188 |
| Q3 (75th Percentile) | 26,73 | 83,22 | 21,9 | 21,8 | 0,358 | 0,317 | 0,496 | 0,287 |
| IQR (Interquartile Range) | 11,4675 | 16,82 | 7,4 | 6,9 | 0,071 | 0,065 | 0,295 | 0,099 |
| Min | -3,31 | 42,95 | 2,9 | 3,2 | 0,171 | 0,149 | 0,077 | 0,036 |
| Max | 44,01 | 105,03 | 29,9 | 27,3 | 0,413 | 0,435 | 1,136 | 0,475 |

Following data preparation, the cleaned dataset shows less kurtosis and skewness, indicating a more regularly distributed dataset (Table 4). This is especially crucial in agricultural settings, where unusual soil moisture levels or sharp temperature swings could be the result of sensor mistakes rather than real environmental circumstances. With skewness levels around zero, these cleaned data patterns imply that the data more accurately depicts normal field circumstances.

The mean provides a measure of central tendency that can be heavily influenced by outliers or extreme values, while the median offers a more robust estimate that minimizes the effect of such outliers, for instance, in the original dataset, the significant divergence between the mean and median in temperature and soil moisture indicates extreme fluctuations. After cleaning, the closer alignment between the mean and median implies a more reliable dataset that is less affected by outliers, which is critical in ensuring accurate modeling and decision-making.

Standard Deviation and IQR (Interquartile Range) help quantify variability and data spread. In the original dataset, the higher standard deviation for variables like soil temperature suggests significant variability, possibly due to sensor malfunctions or environmental outliers, looking at the results provided in the table 3 and 4, after preprocessing, the standard deviation is notably reduced, indicating that the data is now more consistent and reliable. The IQR, which focuses on the central 50% of the data, shows that the data spread in the cleaned dataset is clustered around the median, reflecting fewer extreme variations, thus, this narrowing of the IQR enhances the data's reliability for further analysis and modeling [44].

High skewness and kurtosis values in the original dataset were detected, especially for temperature and soil moisture, indicating the presence of unreal data and non-symmetric distributions. After data cleaning, these values are closer to zero, reflecting a more natural

distribution. Reducing skewness and kurtosis is essential in predictive modeling, to create a solid base, as it enables more accurate forecasting by reducing the distortion caused by extreme values [45].

In precision agriculture, the variability in temperature and soil moisture has direct implications for crop management, irrigation scheduling, and environmental monitoring. As shown in Table 4, the cleaned dataset offers a more accurate representation of typical field conditions, with reduced outliers and variability, making it easier to predict future trends and plan accordingly, particularly the reduction in skewness, kurtosis, and standard deviation, along with the narrowing of the IQR, illustrate the effectiveness of data preprocessing. These improvements result in a dataset that more accurately represents typical field conditions, offering a more reliable foundation for modeling and analysis in precision agriculture.

## 4.4 Bivariate Analysis

Bivariate analysis is one of the simplest forms of quantitative statistical analysis, It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them [32].

The Pearson correlation coefficient was selected for this analysis as it measures the strength and direction of the linear relationship between two continuous variables [47]. This makes it well-suited for the current dataset, where the goal is to quantify linear dependencies between environmental variables such as soil moisture, temperature, and humidity. This is where Pearson correlation assumes that the relationships are linear, and that the data is normally distributed or at least symmetrically distributed, which holds reasonably true for this dataset after cleaning. While alternatives such as Spearman's rank correlation are available for non-linear or rank-based relationships, the focus here is on linear trends, which Pearson is best suited for. If there were significant non-linear relationships in the data, which were not present in this instance after cleaning, Spearman's correlation would be more appropriate. A thorough summary of the connections between the main variables, humidity, temperature, soil moisture (from two sensors at different depths), soil temperature, and soil conductivity (from two sensors at different depths) is provided by the Pearson correlation matrix, which is shown in Figure 26.

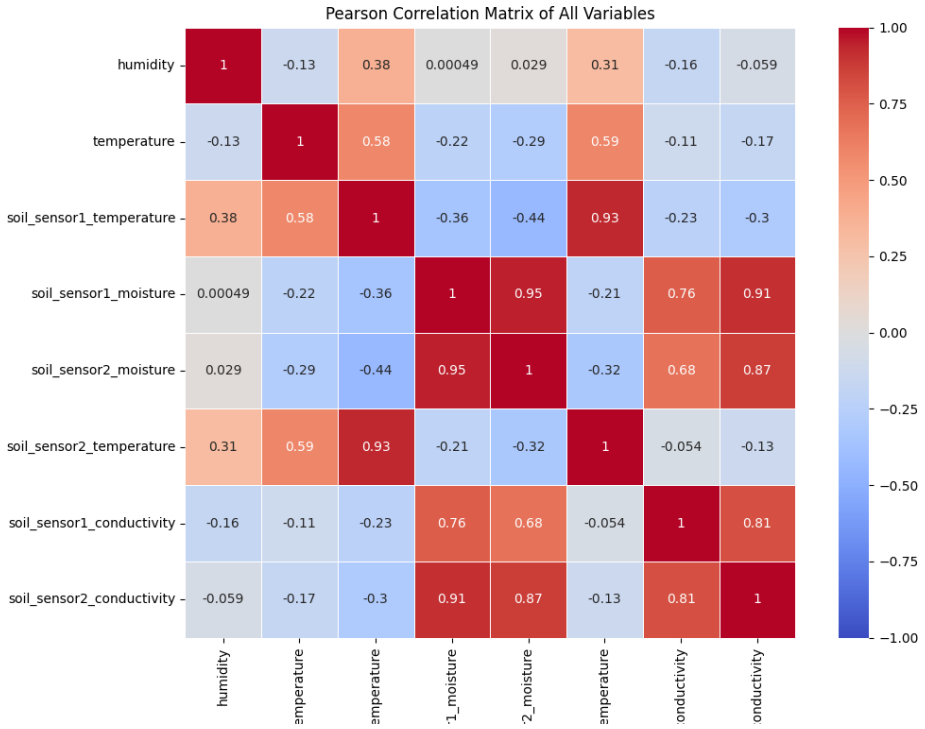



Figure 26 - Pearson correlation matrix of all variables

The matrix reveals several important insights about how these variables interact, which are essential for understanding the environmental dynamics of the field and optimizing agricultural management strategies where the heatmap uses color gradients to indicate the strength of the correlations, with redder tones showing strong positive correlations and bluer tones indicating negative or weak correlations, using a heatmap from the seaborn library.

This was calculated by determining the correlation coefficient for each pair of selected variables, measuring the strength and direction of linear relationships between them. Although time was not directly used in the correlation calculations, the cleaned dataset was loaded into a DataFrame from a SQL Server database, with the 'time' column converted to datetime format to ensure accurate handling of time-based data. For every pair of variables, the Pearson correlation coefficient was calculated using the df_cleaned[variables_to_correlate].corr(method='pearson') function in the developed script that deals with data analysis, resulting in values that range from +1 to -1, where +1 represents a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 signifies no correlation.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Figure 27 - Pearson's correlation coefficient

One of the most noticeable observations from the matrix is the strong positive correlation between soil moisture and soil conductivity across both sensors, mainly, soil_sensor1_moisture and soil_sensor1_conductivity show a correlation of 0.76, while soil_sensor2_moisture and soil_sensor2_conductivity have a correlation of 0.87. These strong correlations suggest that moisture levels directly influence soil conductivity, which is a critical consideration for irrigation management and soil health monitoring. Another important observation is the negative correlation between soil moisture and temperature, soil_sensor1_moisture correlates with temperature at -0.22, and soil_sensor2_moisture correlates with temperature at -0.29. This negative relationship aligns with expected environmental dynamics, where higher temperatures lead to increased evaporation, reducing soil moisture.

This observation has important implications for understanding how temperature variations may necessitate adjustments in irrigation to maintain optimal soil moisture for crops. As expected, the temperature captured by the temperature variable correlates strongly with both soil temperature sensors (soil_sensor1_temperature at 0.58 and soil_sensor2_temperature at 0.59). These correlations suggest that atmospheric temperature strongly influences the soil temperature, but not in a direct 1:1 relationship, which is likely due to factors such as soil depth, shading, and the thermal properties of the soil. Regarding Humidity and Soil Temperature, Humidity shows a moderate positive correlation with soil temperature, particularly with soil_sensor1_temperature (0.38) and soil_sensor2_temperature (0.31), this relationship may reflect how atmospheric conditions impact the heat retention of the soil and the microclimate around the soil surface.

Notably, there is an extremely high correlation between soil_sensor1_moisture and soil_sensor2_moisture is 0.95, while the correlation between soil_sensor1_temperature and soil_sensor2_temperature is 0.93, this high value between the two sensors validates the reliability of the data collected, indicating that the sensors were functioning properly and captured consistent environmental conditions across the field, given that the installation on

different depths. The results of this bivariate analysis provide critical insights into the relationships between environmental factors in the dataset while the use of Pearson correlation allowed for a clear, quantitative understanding of linear dependencies between variables such as temperature, humidity, and soil moisture, the strong correlations between soil moisture and conductivity, as well as the negative correlation between soil moisture and temperature, highlighting the interconnected nature of these variables in an agricultural setting, therefore, these observations and values will be valuable in developing predictive models for irrigation management and environmental monitoring, ensuring optimal growing conditions for crops and confirmed the validity of the sensor data, as indicated by the strong consistency between readings from different sensors, as this bivariate analysis will serve as a foundation for more advanced predictive modeling.

# Chapter 5
# **Predictive Model**

The model development phase is a critical component of this thesis, focusing on establishing predictive models tailored to the objectives defined in prior chapters. The models aim to address primary goals predicting soil moisture based on seasonal indicators and generating synthetic data that simulate real-time environmental conditions. These models sustain the foundation for informed decision-making in precision agriculture, where accurate forecasting and scenario simulation are essential for optimizing resource usage and crop management. The choice of modeling techniques, therefore, was guided by both the nature of the data and the practical requirements of the agricultural domain.

## 5.1 Model Selection Process

Considering the goals of predicting soil moisture, the model selection process prioritized methods capable of handling complex, multivariate interactions that reflect real-world agricultural dynamics, for instance, timeseries data. At various depths, it was found that soil moisture and soil conductivity interacted strongly with variables such as temperature, humidity, and conductivity. These relationships, however, revealed a non-linear and temporally dependent structure, indicating that conventional linear models might not adequately capture these complexities.

Initial testing included Linear and Polynomial Regression models due to their simplicity and ease of interpretability where, soil moisture readings displayed dependencies on a range of other factors, including temperature, humidity, and conductivity, which interact in a non-linear manner. Linear models struggle to capture complex, multivariate interactions, particularly in environmental data that is inherently dynamic and interdependent [48]. To evaluate and compare model performance, two key metrics were used Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). These metrics provide valuable insights into the accuracy of the model predictions by quantifying the error between predicted and actual values.

Mean Absolute Error (MAE) is the average of the absolute differences between the predicted values and the actual values, it is calculated as [49]:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Figure 28 - Mean absolute error equation

where $y_i$ is the actual value, $\widehat{y_i}$ is the predicted value and $n$ is the total number of observations. MAE provides an intuitive measure of average error in the same units as the target variable, making it easy to interpret. Lower MAE values indicate better model accuracy.

Root Mean Squared Error (RMSE) is the square root of the average squared differences between the predicted and actual values, calculated as [49]:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Figure 29 - Root mean square deviation equation

Significant differences between expected and actual values are given more weight by RMSE because it is sensitive to larger errors because of the squaring term. In applications like precision agriculture, where notable departures from actual soil moisture or conductivity could affect decision-making, RMSE is especially pertinent since it can be used to identify models that produce large errors less frequently.

- Linear Regression

  Train MAE: 0.029, Test MAE: 0.029, Test RMSE: 0.036

  The comparatively high MAE and RMSE values of linear regression indicate that it has trouble capturing the intricate relationships present in the dataset. The simplicity of the model restricts its predictive ability for non-linear interactions in soil moisture data, even though the errors are comparable between training and testing, suggesting little overfitting.

- Polynomial Regression (degree 2):

  Train MAE: 0.017, Test MAE: 0.017, Test RMSE: 0.022

  Polynomial Regression improves upon Linear Regression by introducing non-linearity, which lowers both MAE and RMSE values. This improvement suggests that the model captures more of the data's complexity. However, there remains a notable error, indicating that further model sophistication is necessary.

- Random Forest:

  Train MAE: 0.001, Test MAE: 0.003, Test RMSE: 0.007

  Significantly low MAE and RMSE values show that the Random Forest model predicts soil moisture with high accuracy. It stands out as the best option due to its capacity to capture non-linear dependencies between moisture levels and influencing variables such as conductivity, soil temperature, and humidity.

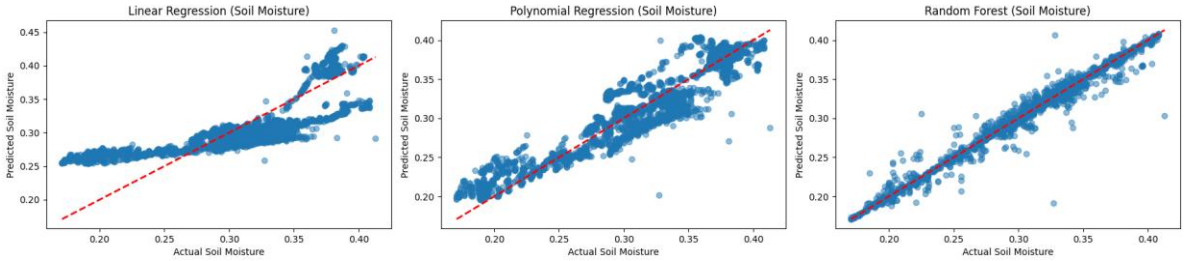



Figure 30 - Comparison between models in Soil Moisture

For conductivity prediction similar trends were observed.

- Linear Regression for Soil Conductivity:

  Train MAE: 0.145, Test MAE: 0.145, Test RMSE: 0.173

  High MAE and RMSE for conductivity suggest that a simple linear relationship cannot adequately predict soil conductivity from other features, underscoring the need for a more complex model.

- Polynomial Regression (degree 2) for Soil Conductivity:

  Train MAE: 0.094, Test MAE: 0.093, Test RMSE: 0.119

  Polynomial Regression moderately improves accuracy for conductivity prediction, further supporting the need for non-linear modeling techniques.

- Random Forest:

  Train MAE: 0.007, Test MAE: 0.018, Test RMSE: 0.043

  For soil conductivity prediction, Random Forest achieves low MAE and RMSE values, confirming its superiority over linear and polynomial models for this dataset.

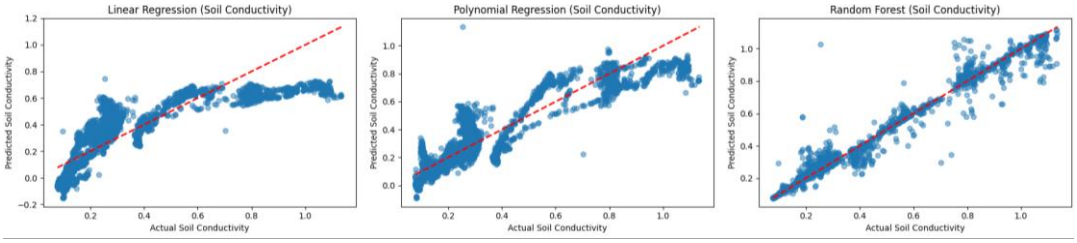



Figure 31 - Comparison between models in Conductivity

## 5.2 Random Forest

The process of selecting an effective model mainly for soil moisture prediction required careful consideration of the non-linear and multivariate nature of environmental data. The aim was to evaluate several modeling approaches, Linear Regression, Polynomial Regression, and Random Forest to understand their capacity for handling the complex interdependencies in this dataset. Unlike regression models that assume a particular structure in data, Random Forests are ensemble methods that do not rely on linear assumptions, making them well-suited for agricultural datasets characterized by interdependent environmental factors. The ability of Random Forests to capture complex feature interactions through an ensemble of decision trees, allows for accurate predictions even with minimal data preprocessing or transformation [51]. It is further supported by insights derived from the bivariate correlation analysis whereby applying Pearson's correlation matrix identified significant positive correlations between soil moisture and conductivity (0.76 for sensor 1 and 0.87 for sensor 2), as well as between soil temperature and atmospheric temperature. These relationships underscore the complexity of environmental interactions, with multiple variables influencing moisture and conductivity in ways that linear models cannot easily capture. Random Forests excel in identifying and leveraging non-linear dependencies within data, crucial for accurately predicting soil properties that are affected by various environmental factors.

In summary, the Random Forest model outperformed Linear and Polynomial Regression by a significant margin, demonstrating robust predictive accuracy for both soil moisture and conductivity. It is a good option for complex agricultural data prediction because, on the surface, its low MAE and RMSE values demonstrate its ability to capture complex relationships within the dataset, nevertheless, because this environmental dataset contains time-dependent variables, it is crucial to capture sequential dependencies.

## 5.3 Long Short-Term Memory (LSTM)

Given the time-dependent nature of environmental data, Long Short-Term Memory (LSTM) networks are considered ideal for generating synthetic data to simulate realistic soil moisture patterns over time. LSTM is a type of Recurrent Neural Network (RNN), that excels at capturing long-term dependencies, which are crucial for modeling how soil moisture evolves over daily and seasonal cycles, effectively addressing the vanishing gradient problem prevalent in

sequential data, allowing the retaining of information over extended periods. LSTM's ability to manage temporal dependencies makes it highly suitable for agricultural data, which often fluctuates based on weather, seasonal changes, and soil conditions [52]. Long Short-Term Memory (LSTM) networks are preferred over other time-series models because of their capacity to efficiently handle multivariate inputs, which is crucial for simulating intricate, interdependent environmental factors like temperature, humidity, and soil moisture. In contrast to traditional time-series models, can simultaneously capture complex relationships among multiple variables, which is essential for comprehending the dynamic interactions between environmental factors.

Moreover, the correlations found in the dataset underscored the need for models that can handle non-linearity and multivariate relationships effectively. Random Forests were chosen to meet this requirement due to their inherent capacity to model complex interactions, enabling the identification of relationships between multiple environmental factors such as soil moisture, temperature, and conductivity. Their ability to manage and interpret multivariate complexity ensures that the intricacies of soil conditions are well-represented, providing reliable predictive outcomes suitable for practical agricultural use.

By leveraging the strengths of these models, this study can provide a robust data-driven approach to agricultural decision-making, enabling researchers and farmers to simulate, predict, and better understand the conditions impacting crop production.

## 5.4 Model Training and Tuning

### 5.4.1 Data Preparation and Feature Engineering

The dataset used for this work was sourced from the PreparedData table, containing time-series data of environmental variables, including temperature, humidity, soil moisture, and soil temperature, since temporal dependencies are crucial in this type of data which is mentioned on chapter 4. Cyclical features were introduced to capture daily patterns in the form of hour of the day transformed into sine and cosine components this is effective in representing features like hours or months, ensuring that relationships between times, such as midnight and midday, are preserved in the data model [53].

Standardizing input features soil moisture is important for preparing features for machine learning models, especially for Long Short-Term Memory (LSTM) networks. The

StandardScaler from sklearn.preprocessing library, normalizes features to have a mean of zero and a standard deviation of one, was used to accomplish this, because LSTMs are sensitive to the size of inputs, therefore as mentioned, feature scaling is essential enabling a strong neural network learning effectiveness, especially in time-series models like LSTMs, where performance can be greatly impacted by magnitude variations [52].

### 5.4.2 Model Architecture

The LSTM network was designed to predict soil moisture using a sliding window approach, thus overlapping sequences of time steps, allowing the model to learn dependencies across these sequences. Each input sequence consists of 10 consecutive time steps, which were by trial, and error chosen to balance between short-term and longer-term dependencies. The output is the prediction for soil moisture at the subsequent time step.

The first LSTM layer, with 50 units, captures complex temporal dependencies in the data and provides outputs to the subsequent LSTM layer, for the second layer, it has the remaining 50 units and aggregates information from the first layer. Doing this as a stacked LSTM architecture, improves the model's capacity to learn complex temporal relationships [54].

### 5.4.3 Training and Evaluation

Overall, the dataset was split into 80% training and 20% testing, a common practice in machine learning for time-series data, ensuring that the model learns patterns from historical sequences while retaining some data for evaluating performance [56].The LSTM was trained for 100 epochs with a batch size of 50 units, balancing computational efficiency and the ability to generalize from the data. Dropout layers were also added between the LSTM layers to mitigate overfitting, this is a phenomenon in machine learning where a model learns not only the underlying patterns in the training data but also the noise and random fluctuations [55], ensuring that the model generalizes well on new, unseen data. The performance of the LSTM was evaluated using metrics such as the MSE on training and test data, a comparison between training and testing errors helps in identifying overfitting issues.

### 5.4.4 Synthetic Data Generation

After training the LSTM model, it was deployed to generate synthetic data in real-time, to complement the sensors measures from temperature and humidity real-time temperature and humidity data were obtained from the OpenWeatherMap API, specifically for Lisbon,

Portugal [57], thus enhancing the quality of the synthetic data, making it as realistic as possible.

As stated in section 5.3, LSTM works on predicting the soil moisture sensor measures, whilst derived sensor values were calculated using observed linear relationships between soil temperature, moisture, and conductivity. To provide additional utility, an anomaly alert mechanism was included in the synthetic data, if the predicted soil moisture fell below a predefined threshold, an alert was generated, then later, given the architecture this alert is sent to the ThingsBoard platform. The synthetic data generated is stored in a table named SyntheticData_RealTime, serving as a repository for all generated values, including temperature, humidity, soil temperature, and moisture, the SQLAlchemy library provided seamless integration with the database, allowing for continuous data storage by the minute.

## 5.5 Discussion and Insights

When dealing with time-dependent agricultural data, the results of this analysis highlight the benefits of utilizing Long Short-Term Memory (LSTM) networks for soil moisture prediction. This is concluded by the initial testing of Random Forest and linear and polynomial regression models, which found patterns in soil moisture. Although Random Forest was good at identifying non-linear relationships in multivariate data, its static nature made it difficult to model temporal sequences that are essential for forecasting changing soil conditions.

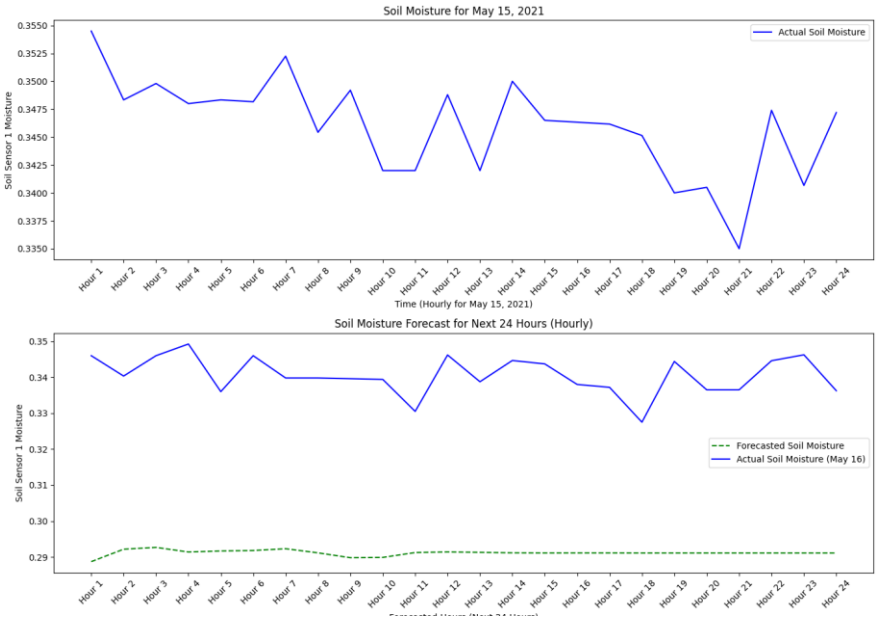



Figure 32 - Random Forest prediction for the next day vs actual next day

The LSTM model, on the other hand, showed an improved ability to manage temporal dependencies and multivariate inputs simultaneously through iterative training and tuning. By effectively learning from sequences, the model was able to identify cyclical patterns in the environment that the Random Forest was unable to predict. The LSTM's design for temporal predictions is validated by the evaluation metrics, which showed that it generalized well without overfitting, including lower Mean Squared Error (MSE) on both training and testing data.

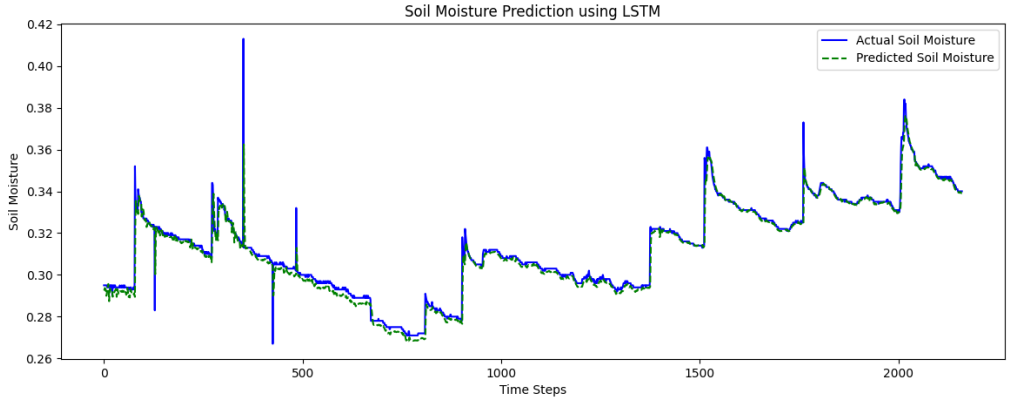



Figure 33 – LSTM over soil moisture

The close alignment between the actual (blue line) and predicted (green dashed line) values in the Soil Moisture Prediction using LSTM plot shows how accurate the model is at forecasting soil moisture levels over time. This accuracy required a few crucial actions. The dataset was first thoroughly preprocessed, with missing values eliminated and input features standardized using MinMax scaling to guarantee consistency between training and testing stages, then the model can learn dependencies between past and future soil moisture levels by creating time-series sequences with a 24-hour (24 Time Steps) look-back window. The LSTM model provides a strong instrument for precision agriculture by bridging the gap between historical trends and future forecasts, providing insights crucial for maximizing resource allocation, which was not possible with Random Forest. Additional enhancements could involve adjusting the LSTM model's hyperparameters, experimenting with various time window sizes, and employing attention mechanisms to improve the model's interpretability and prediction quality. The method shown here offers insightful information for enhancing agricultural water resource management and demonstrates how deep learning models can be used to provide ongoing, real-time farming solutions.

## 5.6 Data Visualization

As can be observed in figure 32, the developed home dashboard for the “Soil IoT” project platform serves as a central hub for monitoring environmental sensors deployed in agricultural fields and to act based on the alerts provided. This interface is designed to present key sensor data in an organized and accessible format, enabling users to effectively manage the sensor network.

A comprehensive overview of every sensor is given by the provisioning table, which is a prominent feature of the dashboard. It shows details like the sensor's name, type, status, and environmental measurements. Furthermore, latitude and longitude values are used to represent the geographic position of each sensor, enabling accurate geospatial awareness of sensor placements. A map view, which provides a real-time satellite perspective of the deployment area and indicates the precise locations of all sensors, further supports this integration of spatial data. Better contextual analysis of the environmental conditions is made possible by this mapping feature, which allows users to visually correlate the sensor data with the corresponding regions in the field. The dashboard also includes an alarms section, which allows users to monitor system alerts and quickly address potential issues related to sensor function or environmental thresholds. The combination of these components ensures that users have access to both a macro-level visualization and granular control of sensor activities, contributing to effective and proactive field management.

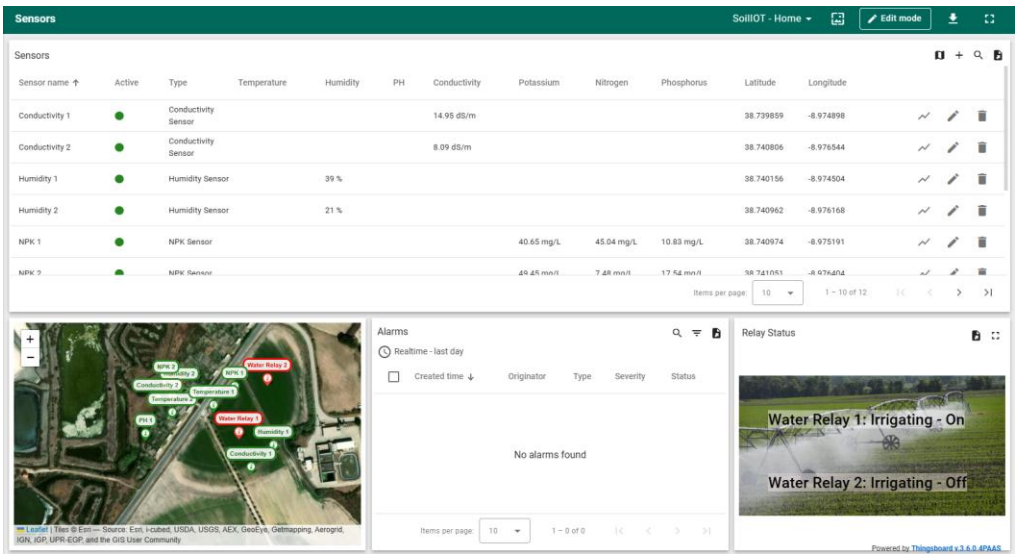



Figure 34 - Developed Home Dashboard

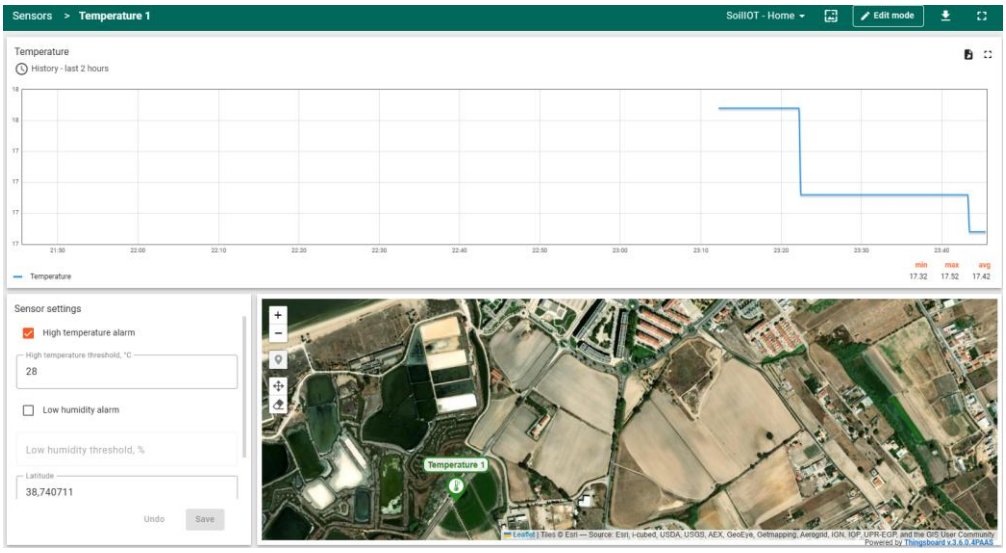



Figure 35 - Device page

The user can also see the real location of the device given the coordinates and create alarms for each device value, which includes a graph displaying temperature history over time. Users can set alarms for specific thresholds, such as high-temperature alerts, and adjust sensor settings directly from these pages, as can be observed in figure 35.

The combination of these components ensures that users have access to both a macro-level visualization and granular control of sensor activities, contributing to effective field management.

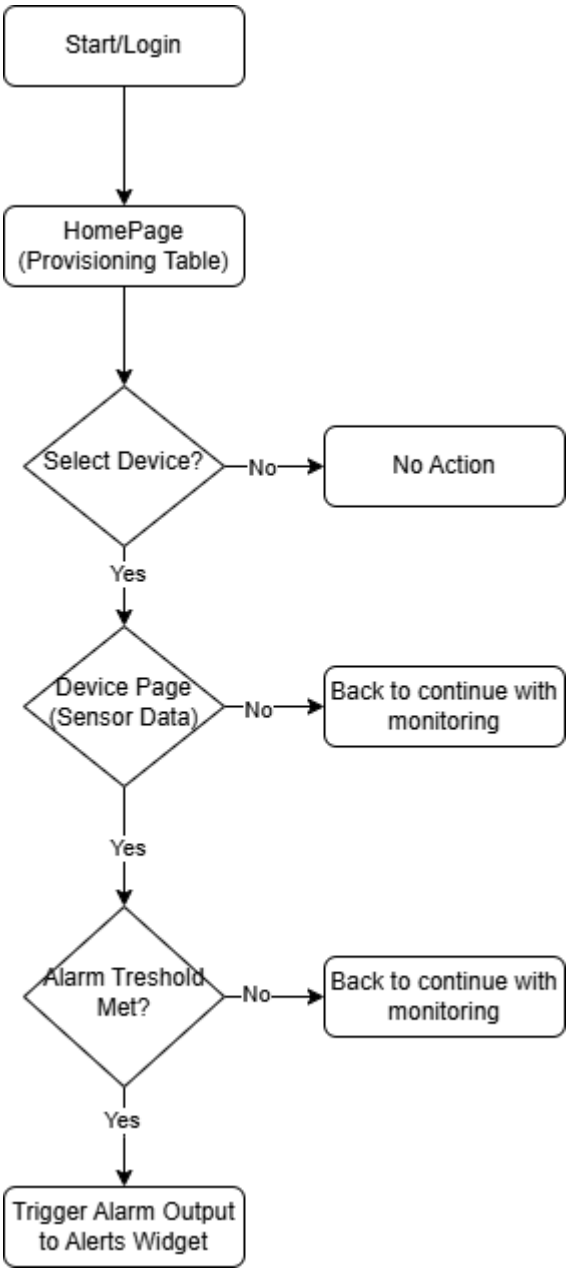



Figure 36 - Thingsboard platform main fluxogram

In figure 34 when the user clicks on a device, the page shows the metrics regarding that device only, being able to pinpoint its location on the map and edit its alarm rules. Alongside the provisioning table and map view, the dashboard also features individual sensor detail pages, such as the temperature sensor interface.

# Chapter 6
# Conclusions and Future Work

## 6.1 Conclusions

To address the issues of climate change, water scarcity, and the demand for sustainable farming methods, the Soil IoT project represents the application of data analysis and IoT to precision agriculture. The project develops a thorough model for tracking soil health, soil moisture, by anticipating hydric stress by mimicking real-time data collection from a network of virtual sensors measuring soil moisture, humidity, temperature, and conductivity. This system's predictive capabilities for soil moisture are improved by its ability to analyze data using machine learning models like LSTM. Synthetic real-time data generation is based on actual datasets, this offers realistic simulations that help users to make well-informed decisions for the most effective decision-making. Sensor-based networks can provide critical insights into soil conditions, enabling resource conservation and environmental sustainability, showing the importance of advanced machine learning models for handling multivariate and time-dependent environmental data. By capturing complex interactions between variables like soil moisture, temperature, and conductivity, the project lays the foundation for more sophisticated data-driven approaches in agricultural management.

With the help of tools for device management, data visualization, and automation, users can visualize, manage, and react to data insights in real time thanks to the project's integration with the ThingsBoard IoT platform. Because of its open-source nature and flexibility, the platform can be customized to meet a variety of agricultural needs, whereas the dashboard's mapping capabilities, provisioning table, and alert systems offer an intuitive user interface that facilitates proactive field management, enabling users to keep an eye on field conditions, spot trends, and effectively address anomalies

Through the integration of IoT and machine learning, the "Soil IoT" project offers a scalable solution for precision agriculture, as it supports sustainable farming methods by providing a framework for in-the-moment monitoring and decision-making. The architecture of the project serves as an example of how technology can improve sustainable agriculture and solve environmental issues. This project establishes the foundation for upcoming advancements in smart agriculture by streamlining data interpretation with adaptable dashboards and offering comprehensive user support.

## 6.2 Future Work

Future work on the Soil IoT project should focus on expanding and enhancing the data analysis capabilities to better support agricultural decision-making like including more parameters to be predicted. Implementing more sophisticated predictive models will enable the system to identify patterns and trends, providing valuable insights such as early warnings for soil degradation or optimal planting times.

Another critical area of future work involves improving data accuracy and sensor integration. Adding more advanced sensors, such as those capable of detailed nutrient analysis, can enhance the precision of the data collected, leading to better analytics. Expanding the integration of the Soil IoT platform with other IoT systems is another valuable future direction, incorporating data from weather stations, drone-based imaging, and pest monitoring systems, the Soil IoT platform can provide a more complete view of the agricultural environment, giving better predictions and resource management. This comprehensive data integration will lead to better insights, supporting not only soil management but also overall farm management and optimization.

Finally, enhancing the machine learning models used in the Soil IoT project for anomaly detection on all measures received, and data reliability will be crucial, the very process of keeping the data ready to analyze by having an outlier removing mechanism and data preprocessing, would in fact provide its value, as it protects the predictive models overfitting while learning the provided datasets, making the system more resilient and less dependent on manual oversight.

# References


[1] National Library of Belarus, "Water is the basis of life," 19 03 2021. [Online]. Available: https://www.nlb.by/en/news/Book-exhibitions/water-is-the-basis-of-life/. [Accessed 14 01 2023].

[2] A. Tomaz, P. Palma, S. Fialho, A. Lima, P. Alvarenga, M. Potes and R. Salgado, *Spatial and temporal dynamics of irrigation water quality under drought conditions in a large reservoir in Southern Portugal*, p. 17, 06 01 2017.

[3] PRIMA, "Drought, the silent enemy of the Mediterranean," 2022 03 2022. [Online]. Available: https://prima-med.org/drought-the-silent-enemy-of-the-mediterranean/. [Accessed 05 01 2023].

[4] McKinsey Global Institute, "Climate risk and response: Physical hazards and socioeconomic impacts," *A Mediterranean basin without a Mediterranean climate?*, p. 24, 28 05 2020.

[5] A. Ravesa and A. S. Shabir, *Precision agriculture using IoT data analytics and machine learning*, p. 17, 05 06 2021.

[6] C. Spencer, "What is NPK Fertilizer? And What Does NPK Do for Plants?," 02 06 2022. [Online]. Available: https://simplysmartgardening.com/what-is-npk/. [Accessed 02 05 2023].

[7] M. Dholu and K. A. Ghodinde, "Internet of Things (IoT) for Precision Agriculture," in *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2018.

[8] Wikipedia, "Scientific method," 10 2023. [Online]. Available: https://en.wikipedia.org/wiki/Scientific_method. [Accessed 3 09 2023].

[9] V. J. P. S. S. Goswami, *Review on How IIoT Has Revolutionized Greenhouse, Manufacturing and Medical Industries*, 2022.

[10] M. Linaza, J. Posada, J. Bund, P. Eisert, M. Quartulli, J. Döllner, A. Pagani, I. G. Olaizola, A. Barriguinha, T. Moysiadis and e. al., *Data-Driven Artificial Intelligence Applications for Sustainable Precision Agriculture, Agronomy, vol. 11, no. 6*, p. 1227, Jun 2021.

[11] A. S. Pradeep Kumar Singh, *An intelligent WSN-UAV-based IoT framework for precision agriculture application*, 2022.

[12] Portland State University, "Research for Thesis & Dissertation Literature Reviews," 2021. [Online]. Available: https://guides.library.pdx.edu/literaturereviews. [Accessed 23 06 2023].

[13] Haifa Group, "Crop Guide: Onion," 2021. [Online]. Available: https://www.haifa-group.com/crop-guide-onion. [Accessed 16 5 2023].

[14] S. K. Abhishek Khanna, "Evolution of Internet of Things (IoT) and its significant impact in the field of Precision Agriculture," *Computers and Electronics in Agriculture, Volume 157*, pp. 218-231, 2019.

[15] A. BENEDEK, T. ROKICKI and A. SZEBERÉNYI, "Energies (19961073)," *Bibliometric Evaluation of Energy Efficiency in Agriculture*, p. p. 5942, 2023.

[16] F. e. a. SÁNCHEZ MILLÁN, "Sensors (14248220)," *Development of Smart Irrigation Equipment for Soilless Crops Based on the Current Most Representative Water-Demand Sensors*, p. p. 3177, 2023.

[17] L. WU and H. ZHENG, "Regional Climate Effects of Irrigation under Central Asia Warming by 2.0 °C," *Remote Sensing*, p. p. 3672, 2023.

[18] T. B. e. a. SHAHI, "Recent Advances in Crop Disease Detection Using UAV and Deep Learning Techniques," *Remote Sensing*, p. p. 2450, 2023.

[19] E. PALOMAR-COSÍN and M. GARCÍA-VALLS, "Flexible IoT Agriculture Systems for Irrigation Control Based on Software Services," *Sensors (14248220)*, p. p. 9999, 2022.

[20] Lora Alliance, "LoRa Specification," *LoRaWAN Specification*, 2015.

[21] Connectivity Standards Alliance , "Zigbee Specification," *Zigbee Document 05-3474-23* , March 2023.

[22] A. Z. J. A. J. G. &. M. R. Ghosh, *Fundamentals of LTE*, 2010.

[23] V. e. a. KRIŽANOVIĆ, "An Advanced Energy-Efficient Environmental Monitoring in Precision Agriculture Using LoRa-Based Wireless Sensor Networks," *Sensors (14248220)*, p. p. 6332, 2023.

[24] A. &. A. A. &. A. L. &. K. A. &. D. K. Khalifeh, "A machine learning-based weather prediction model and its application on smart irrigation," *Journal of Intelligent & Fuzzy Systems*, pp. 43. 1-8, 2022.

[25] F. e. a. JIAO, "Prediction Model of Rice Seedling Growth and Rhizosphere Fertility Based on the Improved Elman Neural Network," *Computational Intelligence & Neuroscience*, p. p. 1–7, 2022.

[26] A. Agrawal, "Agile Methodology: Incremental and Iterative way of development," 04 12 2019. [Online]. Available: https://medium.com/@ashutoshagrawal1010/agile-methodology-incremental-and-iterative-way-of-development-a6614116ae68.

[27] Catchpoint, "MQTT Broker," [Online]. Available: https://www.catchpoint.com/network-admin-guide/mqtt-broker. [Accessed 02 10 2024].

[28] ThingsBoard, "What is ThingsBoard?," 2024. [Online]. Available: https://thingsboard.io/docs/getting-started-guides/what-is-thingsboard/. [Accessed 3 10 2024].

[29] Thingsboard, "Using RPC capabilities," 2023. [Online]. Available: https://thingsboard.io/docs/user-guide/rpc/. [Accessed 30 10 2023].

[30] Emqx, "Mosquitto MQTT Broker: Pros/Cons, Tutorial, and a Modern Alternative," 21 08 2023. [Online]. Available: https://www.emqx.com/en/blog/mosquitto-mqtt-broker-pros-cons-tutorial-and-modern-alternatives. [Accessed 31 10 2023].

[31] K. T. K. I. H. I. C. N. P. L. P. T. &. P. P. Guoqing Li, "The Convergence of Container and Traditional Virtualization: Strengths and Limitations," *SN Computer Science*, 11 05 2023.

[32] Purdue University, "Dataset 2021," 2021. [Online]. Available: https://purduewhin.ecn.purdue.edu/dataset2021/. [Accessed 01 05 2024].

[33] J. S. a. D. Syed, "Techniques to deal with missing data," *2016 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA)*, pp. 1-4, 2016.

[34] M. K. J. P. J. Han, "Getting to Know Your Data," *Data Mining (Third Edition)*, pp. 39-82, 2012.

[35] Wikipedia, "Univariate (statistics)," 06 2024. [Online]. Available: https://en.wikipedia.org/wiki/Univariate_(statistics)#cite_note-1. [Accessed 10 2024].

[36] Wikipedia, "Moving average," 03 10 2024. [Online]. Available: https://en.wikipedia.org/wiki/Moving_average. [Accessed 20 10 2024].

[37] K. C. K. A. L. &. T. K. R. DeJonge, "Simulating soil temperature effects on soil water and crop water productivity," *Computers and Electronics in Agriculture 113,* p. 32–45, 2015.

[38] M. S. &. I. S. Kukal, "Climate-driven crop yield and yield variability and climate change impacts on the U.S.," *Scientific Reports, 8, Article 3450,* 2018.

[39] J. A. H. a. P. M. Hartigan, "The Dip Test of Unimodality," *The Annals of Statistics, vol. 13, no. 1,* pp. 70-84, 1985.

[40] L. N. a. F. Cicirelli, "Fast and Accurate K-means Clustering Based on Density Peaks," *dvances in Data-Driven Computing and Intelligent Systems, 1st ed.,* 2023.

[41] M. S. K. a. S. Irmak, "Climate-Driven Crop Yield and Yield Variability and Climate Change Impacts on the U.S. Great Plains Agricultural Production,," *Scientific Reports, vol. 8, no. 1, Art. no. 3450,,* pp. 1-10, 2018.

[42] M. T. a. D. Or, "Generalized Soil Water Retention Equation for Adsorption and Capillarity," *Vadose Zone Journal, vol. 4, no. 2,* p. 190–207, 2005.

[43] D. R. R. a. T. M. S. H. Blanco-Canqui, "Soil and crop response to precision irrigation in the Great Plains: Benefits, challenges, and future directions," *Agronomy Journal, vol. 113,* p. 3697–3711.

[44] S. O. I. a. C. A. Madramootoo, "Sensitivity of spectral vegetation indices for monitoring water stress in tomato plants," *Computers and Electronics in Agriculture, vol. 163,* 2019.

[45] C. C. A. a. P. S. Yu, "Outlier Detection for High Dimensional Data," *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 2001),* p. 37–46, 2001.

[46] T.-H. K. a. H. White, "On More Robust Estimation of Skewness and Kurtosis," *inance Research Letters, vol. 1, no. 1,* p. 56–73, 2004.

[47] Wikipedia, "Bivariate analysis," 12 2023. [Online]. Available: https://en.wikipedia.org/wiki/Bivariate_analysis. [Accessed 10 2024].

[48] J. Wang, "Pearson Correlation Coefficient," *Encyclopedia of Systems Biology,* 2013.

[49] A. G. a. J. Hill, "Data Analysis Using Regression and Multilevel/Hierarchical Models," *Cambridge, U.K.: Cambridge Univ. Press,* 2007.

[50] Wikipedia, "Mean absolute error," 2024. [Online]. Available: https://en.wikipedia.org/wiki/Mean_absolute_error. [Accessed 24 09 2024].

[51] Wikipedia, "Root mean square deviation," 2024. [Online]. Available: https://en.wikipedia.org/wiki/Root_mean_square_deviation. [Accessed 23 09 2024].

[52] A. L. a. M. Wiener, "Classification and Regression by randomForest," *R News, vol. 2, no. 3,* p. 8–22, 2002.

[53] Y. B. a. A. C. I. Goodfellow, "Deep Learning," *Cambridge, MA: MIT Press,* 2016.

[54] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," p. 849, 2019.

[55] I. S. a. O. V. W. Zaremba, "Recurrent Neural Network Regularization," p. 8, 2014.

[56] J. Brownlee, "Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-To-End," *Machine Learning Mastery,* 2017.

[57] G. H. A. K. I. S. a. R. S. N. Srivastava, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting,," *Journal of Machine Learning Research, vol. 15,* p. 929–1958, 2014.

[58] OpenWeather, "OpenWeather API Documentation," [Online]. Available: https://openweathermap.org/api. [Accessed 03 04 2024].

[59] Joint Research Centre (JRC), the European Commission’s science and knowledge service, "Drought in western Mediterranean 2022," *GDO Analytical Report,* p. 37, 2022.

[60] Food And Agriculture Organization Of The United Nations, [Online]. Available: https://www.fao.org/land-water/databases-and-software/crop-information/onion/en/. [Accessed 03 05 2023].

[61] D. Corwin and S. Lesch, "Computers and Electronics in Agriculture," *Apparent soil electrical conductivity measurements in agriculture,* pp. 11-43, March 2015.

[62] H. M. A. Cherine Fath, "A Secure IoT-Based Irrigation System for Precision Agriculture Using the Expeditious Cipher," *Sensors,* p. 23(4), 5 January 2023.

[63] Wikipedia, "Desenvolvimento ágil de software," 10 2021. [Online]. Available: https://pt.wikipedia.org/wiki/Desenvolvimento_%C3%A1gil_de_software. [Accessed 02 2022].