esprit

Se former autrement

HONORIS UNITED UNIVERSITIES

Report Data science project

Project:
FARMWISE

Elaborated by **LIBERDATA** group:

Sana khiari

Oumaima Benhaj

Mehdi Bchir

Ines Neji

Mohamed Amine Brahmi

Sarra Bouden

Acknowledgement

As we embark on this project, we would like to express our appreciation for the support and guidance we anticipate receiving from our academic supervisors, mentors, and colleagues. Their insights and expertise will be invaluable in helping us navigate the challenges ahead and refine our approach.

We also acknowledge the importance of the resources, tools, and data that we will be using throughout this project. We look forward to collaborating with various stakeholders, whose contributions will play a crucial role in the development and success of our work.

Finally, we extend our gratitude to our families and friends for their continuous encouragement and motivation as we undertake this project. Their support remains an essential source of strength.

While this report marks the beginning of our journey, we are confident that the collective efforts of all involved will lead to meaningful progress and valuable outcomes.

Content

Table Of figures

Introduction

Agriculture is one of the most important sectors for global food security and economic growth. However, it faces many challenges, such as climate change, limited natural resources, and the increasing complexity of modern farming techniques. Many farmers still depend on their personal experience and intuition rather than data-driven insights, which can sometimes lead to inefficiencies in managing resources, selecting the right crops, and preventing plant diseases.

In response to these challenges, our project was developed at the Private Higher School of Engineering and Technology (ESPRIT) as part of the Integrated Project for Data Science (PIDS) course. This project is designed to help students apply their knowledge in real-world scenarios and develop solutions that have a meaningful impact.

Our project, **FARMWISE** , takes advantage of the latest advancements in artificial intelligence and data science to modernize agricultural practices. The goal is to provide farmers with intelligent recommendations, risk assessments, and predictive insights to help them make better decisions. By integrating AI-driven solutions into farming, **FARMWISE** aims to improve productivity, optimize resource usage, and promote sustainable agriculture.

1. General context

Introduction

This chapter introduces the host organization and the challenges in Tunisia's agricultural sector. It explores existing solutions, their limitations, and the added value of our AI-driven approach. Finally, we present the Team Data Science Process (TDSP) and our contribution to the Sustainable Development Goals (SDGs).

1.1 Host Organisation

The Private Higher School of Engineering and Technology (ESPRIT) is a leading private engineering institution based in Ariana, Tunisia. Founded in 2003, it has grown to become the largest private university in the country, with over 7,000 students and approximately 250 full-time instructors. The school is officially accredited by the Ministry of Higher Education and Scientific Research of Tunisia.

ESPRIT stands out for its strong partnerships with businesses and academic institutions, offering a practical and industry-oriented education that equips students with the skills needed to succeed in the professional world.

In 2020, ESPRIT became part of the Honoris United Universities network, expanding its educational programs and fostering international collaboration. A year later, in 2021, Entreprises Magazine recognized ESPRIT as the best private engineering university in Tunisia, highlighting its commitment to academic excellence and its influential role in the national higher education sector.

Figure 1.1 : Esprit Logo

1.2 Problematic

Agriculture plays a vital role in Tunisia’s economy, employing approximately 15% of the workforce and covering 9.28 million hectares of agricultural land. However, the sector faces several significant challenges. First and foremost, the late detection of crop diseases leads to considerable yield losses and a heavy reliance on pesticides, which exacerbates soil pollution. Additionally, inefficient management of water resources, intensified by climate change, results in significant waste and threatens the profitability of farms. Furthermore, predicting crop yields remains a complex task due to climate fluctuations and the lack of suitable analytical tools. The widespread dependence on chemical inputs not only harms the environment but also poses health risks, with few accessible alternatives. Lastly, the lack of

support and guidance for new farmers severely limits their chances of success and hinders the renewal of the agricultural sector

1.3 Existing solutions

This section explores existing agricultural technologies both within Tunisia and beyond, highlighting their advantages and limitations.

1.3.1 In Tunisia :

- **iFarming :**

iFarming is a precision irrigation solution developed by Agri-Tech Tunisia. It utilizes scientific algorithms to simulate real-time water requirements for various crops, considering factors such as crop type, phenological stage, and local climatic conditions. This approach aims to optimize water usage, potentially achieving water savings exceeding 40%.

Figure 1.2 : iFarming Logo

Advantages :

- **Water Efficiency:** By tailoring irrigation schedules to the specific needs of crops and current weather conditions, iFarming promotes significant water conservation, which is crucial in regions facing water scarcity.
- **Scientific Approach:** Utilizing scientific algorithms ensures that irrigation practices are based on empirical data, enhancing the precision and effectiveness of water application.

Limitations :

- **Technological Requirements:** Implementing iFarming may necessitate access to compatible hardware and software, as well as a reliable internet connection, which could be challenging for farmers in remote or under-resourced areas.
- **Learning Curve:** Farmers may need to invest time in understanding and effectively utilizing the system, which could be a barrier for those less familiar with digital tools

- **Smart Farm :**

Smart Farm is a Tunisian startup specializing in precision agriculture solutions aimed at optimizing crop production, conserving water, and enhancing overall farm efficiency. Their offerings include connected soil sensors, decision-support applications, and comprehensive training and support services.

Figure 1.3 : Smart Farm Logo

Advantages:

- **Decision-Support Application:** Smart Farm offers a web and mobile application that visualizes data collected by the sensors. This tool helps farmers anticipate irrigation needs through intuitive dashboards, promoting informed decision-making.
- **Resource Optimization:** By implementing Smart Farm's solutions, farmers can achieve up to a 30% increase in yield, 50% water savings, and 40% energy savings, contributing to both economic and environmental benefits.

Limitations:

- **Initial Investment**: The adoption of precision agriculture technologies may require a significant upfront investment, which could be a barrier for small-scale farmers with limited financial resources.
- **Learning Curve:** Farmers may need to invest time in training to effectively utilize the technology and interpret the data provided by the system, which could be a hurdle for those less familiar with digital tools.

1.3.2 Outside Tunisia :

- **Plantix :**

Plantix is an AI-powered mobile application designed to help farmers diagnose plant diseases, nutrient deficiencies, and pest issues using image recognition. It provides actionable recommendations for improving crop health.

Figure 1.4 : Plantix Logo

Advantages:

- **AI-Powered Image Recognition:** Uses machine learning to accurately detect plant diseases, pests, and deficiencies from smartphone photos.
- **Localized Recommendations:** Provides customized solutions based on the specific region and local agricultural practices. Includes treatment suggestions using chemical, organic, and integrated pest management approaches.

Limitations:

- **Accuracy Depends on Image Quality:** The diagnosis relies heavily on the quality of the uploaded image.Blurry or unclear photos may lead to incorrect results.

- **Limited Disease Database for Some Crops:** While the app covers many crops, it may lack data for less common plants or specific regional crop varieties.

● Bushel Farm:

Bushel Farm is a farm management software designed to assist farmers in efficiently managing their operations and making informed decisions.

Figure 1.5 : Bushel Farm Logo

Advantages:

- **Comprehensive Farm Management:** Bushel Farm offers a unified dashboard that allows farmers to plan, monitor, and market their crops effectively.
- **Financial Tracking and Reporting:** Bushel Farm provides tools for tracking production costs, generating profit and loss statements, and understanding field-level profitability.

Limitations:

- **Feature Limitations in Lower-Tier Plans:** Certain advanced features, such as machine data connections and detailed profit and loss reports, are only available in higher-tier plans. This may limit the functionality for users subscribed to more basic plans.
- **Learning Curve:** New users might require time to fully explore and utilize all features effectively. Adequate training or support may be necessary to maximize the software's potential.

● Climate FieldView:

Climate FieldView is an AI-powered agricultural platform developed by The Climate Corporation (a subsidiary of Bayer). Through yield analysis, predictive analytics, and real-time field monitoring, it offers data-driven insights to assist farmers in optimizing crop management. To improve precision agriculture, the platform combines weather information, soil composition, satellite photography, and machinery sensors.

Figure 1.6 : Climate FieldView Logo

Advantages:

- **Advanced Predictive Analytics:** Predicts potential yields, the best times to plant, and when to harvest using AI-driven models.

  Additionally, it offers farmers up-to-date weather information to aid in decision-making.

- **Remote Field Monitoring via Satellite & IoT Sensors:** Farmers can monitor crop health using high-resolution satellite imagery.

Limitations:

- **Primarily Made for Industrial, Large-Scale Farming:** Smallholder farmers find the platform less accessible due to its high cost and requirement for contemporary farming equipment.
- **Heavy Dependence on Machinery & IoT Sensors:** The platform's full potential is unlocked only when used with high-tech agricultural machinery.

1.4 Proposed solution

It is evident that existing solutions in the agricultural sector often fail to address the key challenges effectively. Many rely on traditional methods or lack integration with advanced technologies, which limits their potential. Our proposed solution aims to combine the best of these existing tools while adding innovative features to create a comprehensive system. The main goal of our project is to enhance the efficiency and sustainability of Tunisian agriculture by developing an intelligent recommendation system that integrates AI for disease prediction, IoT sensors for efficient water management, and drones for precise monitoring of crops. This integrated approach will not only minimize the use of pesticides and water but also provide data-driven insights for farmers to make informed decisions. The system will be designed to be user-friendly, enabling farmers to adapt quickly and optimize their resources, ensuring greater productivity and environmental sustainability.

1.5 Team Data Science Process (TDSP)

The Team Data Science Process (TDSP) is an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently. TDSP helps improve team collaboration and learning. TDSP is a distillation of the best practices and structures from Microsoft and others in the industry that facilitate the successful implementation of data science initiatives that help companies fully realize the benefits of their analytics program [1]

TDSP is divided into the following key phases:

Business Understanding

- Define the problem statement, business objectives, and success criteria.
- Identify key stakeholders and expected project outcomes.

Data Acquisition & Understanding

- Collect and preprocess relevant datasets from various sources.
- Perform exploratory data analysis (EDA) to detect patterns, correlations, and inconsistencies.

Modeling

- Select appropriate deep learning or statistical models based on the problem type.
- Train, evaluate, and optimize models using relevant techniques.

Deployment

- Integrate the trained model into an application .

TDSP Roles

- Solution architect : to design and be responsible for how the entire solution is being operationalized within the organization.
- Project manager : who manages the day by day of the team.
- Project lead is the ultimate technical responsibility for what the team is going to produce.
- Data scientist : role to analyze and pull insights out of the data.

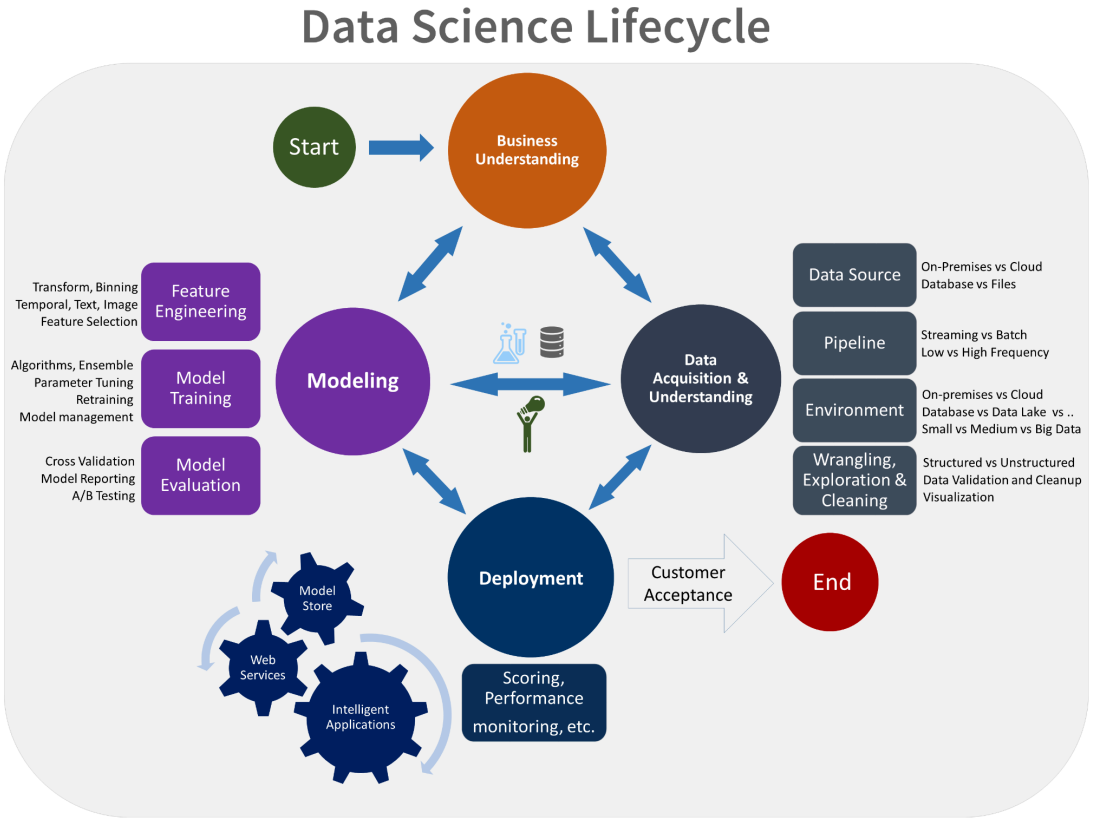


Figure 1.7: TDSP Lifecycle **.**

1.6 Contribution to the Sustainable Development Goals (SDGs)

FARMWISE is committed to addressing key challenges in agriculture by integrating advanced technologies such as AI,deep learning, IoT… . Through these innovations, the project directly contributes to multiple **United Nations Sustainable Development Goals (SDGs)**, aiming to enhance sustainability, resilience, and efficiency in Tunisia's agricultural sector.

Among the SDGs that this project contributes to, we can highlight the following:

- **SDG 3 – Good Health and Well-Being**

Figure 1.8: SDG 3

How?

- Reducing the **use of harmful pesticides** through the recommendations that offer our solution .

Impact:

- **Healthier food production** with lower chemical residues.
- **Reduced occupational hazards** for farmers working in extreme conditions.

- **SDG 8 – Decent Work and Economic Growth**

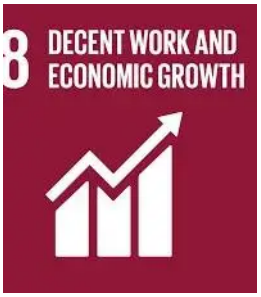


Figure 1.9: SDG 8

How?

- Creating **new job opportunities** in agri-tech, AI, and data-driven farming.

- Increasing **farmers' income** through higher productivity and cost savings.

Impact:

- Empowering **rural communities** with sustainable agricultural practices.
- Driving **economic growth** through innovation in farming.

❖ **SDG 12 – Responsible Consumption and Production**

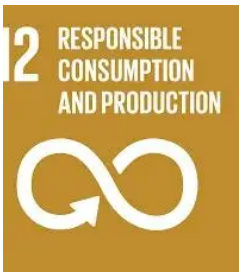


Figure 1.10: SDG 12

How?

- Reduce overuse of resources.
- Encouraging **sustainable supply chain management** in agriculture.

Impact:

- Lower **environmental impact** of farming practices.
- More **efficient use of natural resources**.

❖ **SDG 15 – Life on Land**

Figure 1.11 : SDG 15

How?

- Helping farmers use land in a way that keeps it healthy.

Impact:

- Keeping the soil rich and fertile, preventing it from turning into desert.

- Making sure nature and farming can exist together for a long time.

Through its innovations, FARMWISE aligns with these SDGs to help Tunisia's agricultural sector become more sustainable, resilient, and productive, while contributing to broader global goals for sustainable development.

1.7 GLOBAL METRICS OF THE FARMWISE PROJECT

The FarmWise Project aims to revolutionize modern agriculture through precision farming techniques and AI-driven solutions. Below are the key global metrics that highlight its impact:

- **Economic & Business Impact:** Increased profitability for farmers by optimizing resource usage and reducing operational costs.
- **Agricultural Productivity & Efficiency**: Enhanced crop yields through automated weed detection and management, leading to more efficient farming practices.
- **Risk Management and Mitigation**: Reduced dependency on chemical herbicides and minimized crop damage, ensuring sustainable farming practices.
- **High Compliance Rate for Export Regulations:** Adherence to strict agricultural standards, facilitating easier access to international markets.
- **Reduction in Resource Waste:** Significant decrease in water, pesticides, and fertilizers used, promoting environmentally friendly and cost-effective farming solutions.

This comprehensive approach ensures that the FarmWise Project contributes to both economic growth and sustainable agriculture.

Conclusion:

In this chapter, we began with an overview of the host organization, followed by a discussion of the problem that led to this project and the existing solutions. We then presented our proposed solution, the TDSP methodology, and finally, our contribution to the Sustainable Development Goals.

2.Business understanding

Introduction

In this chapter, we will define the key objectives of the FARMWISE project, starting with the project's overall aim to enhance agriculture in Tunisia . We will then discuss the business objectives . Finally, we will outline the data science objectives, emphasizing the role of data-driven insights in optimizing farming practices and driving innovation in the sector.

2.1 Project Objectives

The FARMWISE project aims to revolutionize agriculture in Tunisia by using advanced technologies to provide farmers with data-driven insights. It helps optimize farming practices, improve productivity, and promote sustainability, ultimately enhancing food security and economic stability in the sector.

2.2 The Business Objectives

BO1: Assist farmers in making informed, real-time decisions.

BO2: Optimize resource use (water, fertilizers, pesticides).

BO3: Predict risks: disasters, diseases, or anomalies.

BO4: Suggest cure for infected plants.

BO5: Predict the appearance of parasitic herbs.

BO6: Segmentation of Farmers and Land for Agricultural Businesses.

2.3 Data Science Objectives

DSO1: Train predictive models using soil, climate, and yield data to generate personalized recommendations for:

-an optimized irrigation schedules based on:

- Soil moisture levels (IoT soil sensors, remote sensing).
- Weather forecasts (rain prediction, temperature, humidity)**.**

-Fertilization & Soil Health Management based on

- Soil nutrient composition (NPK levels, pH, organic matter…)
- Crop nutrient needs (based on growth stage)

-Suggest best crop(s) to plant based on:

- Soil suitability (type, nutrients, acidity).

- Climate compatibility (temperature, rainfall, humidity).
- Market demand (crop prices & profitability trends).

-Revenue Estimation:

- Estimate **total expected yield** based on climate & soil conditions.
- Suggest **best selling window** based on price trends.

DSO2: Develop models to predict the optimal amount of resources needed based on soil data, weather conditions, and crop types.

-Implement optimization algorithms to maximize resource efficiency while minimizing waste..

DSO3: Develop a deep learning model using image data to classify plant diseases with high accuracy along with a suggestions system for:

-Organic treatment options (e.g., neem oil, biological pest control)**.**

-Severity assessment (low/medium/high risk) to prioritize intervention.

-Suggest **preventive measures** based on disease type and weather patterns.

DSO4: Implement decision support systems to recommend treatments or preventive measures based on the detected disease.

-Build classification models to diagnose plant diseases based on image data, environmental factors, and symptoms.

DSO5: Develop predictive models for the appearance of parasitic herbs

- Use historical data, environmental factors (e.g., temperature, humidity, soil conditions), and satellite imagery to train machine learning models that predict the likelihood and timing of parasitic herb infestations.
- Apply image recognition techniques on crop and field images to detect early signs of parasitic herbs.

DSO6: Use clustering models for:

-Farmers segmentation: Categorize farmers into meaningful groups to optimize services & recommendations(Market orientation segmentation,Experience level segmentation,Smallholder vs. large-scale farmers…)

-Land Segmentation & Classification: Classify land areas based on:

- Soil health & fertility (pH, NPK levels, organic matter)
- Climate suitability (rainfall, temperature, drought risk)
- Crop adaptability (land best suited for wheat, maize, vegetables, etc.)
- Irrigation type (rain-fed, drip, canal-based, dryland farming)

Agricultural Business Optimization: Match segmented farmers & lands with:

- Best supply chain partners (seed suppliers, fertilizer distributors).
- Agro-business investment opportunities based on land profitability.

Conclusion

The FARMWISE project harnesses advanced data science techniques to optimize agriculture in Tunisia. By integrating predictive modeling, deep learning, and clustering, it provides farmers with intelligent recommendations for irrigation, fertilization, crop selection, and disease management. This data-driven approach enhances productivity, sustainability, and economic stability, paving the way for a smarter and more resilient agricultural sector.

3.Data Acquisition, understanding and preparation

Introduction

Data acquisition is a crucial step in any data science project. It involves collecting data from various sources and understanding its structure, quality, and relevance. This step ensures that the data is reliable and ready for further processing analysis and prediction .

3.1 Definition:

Data Acquisition

Data acquisition is the process of collecting raw data from various sources, including databases, sensors, APIs, satellite images, or manual records. This step ensures that relevant, high-quality data is gathered for analysis and modeling.

Key Steps:

- Identifying sources (e.g., weather databases, soil monitoring, market reports).
- Extracting raw data (e.g., downloading CSV files, scraping APIs, collecting satellite images).
- Validating data integrity (checking completeness and accuracy at the source).

Data Understanding

Definition:
Data understanding involves exploring, analyzing, and assessing the quality of acquired data. It helps in identifying patterns, inconsistencies, and potential biases that may affect modeling outcomes.

Key Steps:

- Data exploration (summary statistics, distributions, missing values).
- Data visualization (histograms, correlation matrices, scatter plots).
- Checking for inconsistencies (duplicate records, incorrect formats, class imbalances).

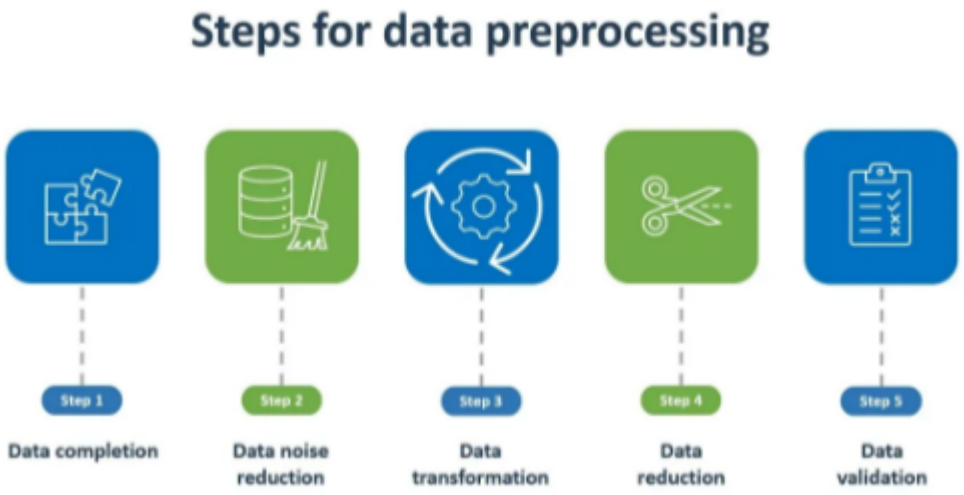
Data Preparation

Data preparation is the process of cleaning, transforming, and structuring raw data into a usable format for modeling. It ensures data consistency, enhances model performance, and reduces biases.

Key Steps:

- Data Cleaning (handling missing values, removing duplicates).
- Feature Engineering (creating new variables, aggregating data).

- Data Transformation (normalization, encoding categorical data).
- Splitting datasets (train, validation, test sets).
- Data augmentation (**for image datasets to improve model generalization).**



3.2 Data Types

In this project, we will be using two different types of data: numerical data and images.

- ### Numerical data

Numerical data refers to information that is represented by numbers and can be used for mathematical calculations, statistical analysis, and modeling. It includes integers, floating-point numbers, and can represent various types of measurements such as age, temperature, or revenue.

Loading data

```
df = pd.read_csv("Crop_Recommendation (1).csv")
df.head()
```

	N	P	K	temperature	humidity	ph	rainfall	label
0	90	42	43	20.879744	82.002744	6.502985	202.935536	rice
1	85	58	41	21.770462	80.319644	7.038096	226.655537	rice
2	60	55	44	23.004459	82.320763	7.840207	263.964248	rice
3	74	35	40	26.491096	80.158363	6.980401	242.864034	rice
4	78	42	42	20.130175	81.604873	7.628473	262.717340	rice

Figure 3.1: Crop_Recommendation dataset.

- **Image data**

Image data refers to visual information represented in the form of pictures or graphics. It consists of pixels, each containing color values, and can be used for tasks such as object detection, image classification, and pattern recognition. Image data is typically stored in formats like JPEG, PNG, or TIFF.

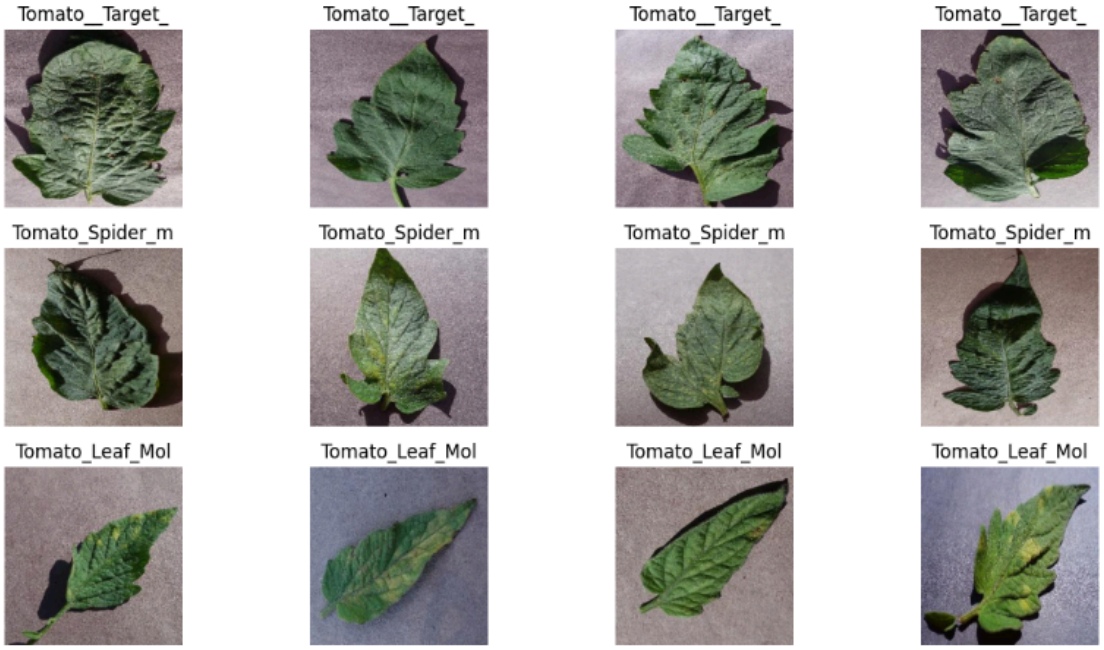


Figure 3.2: images of dataset PlantVillage.

3.3 Data Acquisition, understanding and preparation for Each Business Objective

BO1: Assist Farmers in Making Informed, Real-Time Decisions

Dataset: Climate & Soil Data (Tunisia Meteorological Database)

Data Acquisition

- **Source**:
  - Tunisia Meteorological Database
  - OpenWeather historical datasets
  - Remote sensing soil data
- **Purpose**:
  - Provide real-time **weather and soil conditions** for precision agriculture.
  - Help farmers decide **when to plant, irrigate, or harvest** based on climate.
- **Data Collected**:
  - **Climate Data 1** (62,680 rows, 14 columns) → **Historical weather trends**
  - **Climate Data 2** (450 rows, 6 columns) → **Real-time station data**

- **Soil Data** (62,680 rows, 21 columns) → **Soil quality and composition**

Data Understanding

- **Analyzed temperature variations** and their correlation with rainfall.
- **Identified missing station readings** and inconsistent timestamps.
- **Plotted histograms** of soil pH, moisture, and nitrogen content.
- **Checked multicollinearity** (correlation matrix) between soil attributes.

Data Preparation

- **Data Merging**: Combined climate datasets using **timestamps as the key**.
- **Handling Missing Values**:
  - Used **linear interpolation** for climate readings.
  - Used **mean imputation** for missing soil moisture values.
- **Feature Engineering**:
  - Created **"Soil Suitability Index"** based on pH, moisture, and nutrients.
  - Introduced **"Heat Stress Index"** to measure extreme temperature effects.
- **Data Encoding**:
  - One-hot encoded categorical variables (soil type, region).
- **Standardization**:
  - Applied **MinMaxScaler** to scale humidity, temperature, and moisture.

BO2: Optimize Resource Use (Water, Fertilizers, Pesticides)

Datasets: Fertilizer Prediction, Pesticide Price Database, Water Consumption

Data Acquisition

- **Fertilizer Prediction Dataset**:
  - **Source**: IEEE DataPort
  - **Purpose**: Optimize **fertilizer recommendations** based on soil conditions.
- **Pesticides Price Database**:
  - **Source**: Aggregated pesticide prices from local markets.
  - **Purpose**: Help farmers optimize pesticide purchases.
- **Water Consumption Data**:
  - **Source**: FAO Tunisia Water Resources Database.
  - **Purpose**: Monitor and **optimize irrigation schedules** based on usage trends.

Data Understanding

- **Checked data completeness** (fertilizer dataset had no missing values).
- **Visualized water usage trends** per governorate in Tunisia.
- **Analyzed correlation** between **fertilizer application and yield output**.

Data Preparation

- **Encoding & Normalization**:
  - Used **Label Encoding** for **crop types**.

    - Applied **MinMaxScaler** for **nitrogen, phosphorus, and potassium** variables.
- **Feature Engineering**:
    - Created **"Optimal Fertilizer Cost per Hectare"** as a derived feature.
    - Generated **"Irrigation Deficiency Score"** for dry regions.
- **Data Aggregation**:
    - Merged **water consumption with soil moisture data** to create **regional water efficiency scores**.

BO3: Predict Risks: Disasters, Diseases, or Anomalies

Datasets: Landslide4Sense (Satellite Data) & PlantVillage (Crop Diseases)

Data Acquisition

- **Landslide4Sense Dataset**:
    - **Source**: Sentinel-2 satellite data
    - **Purpose**: Predict **landslides and soil degradation risks**.
- **PlantVillage Dataset**:
    - **Source**: Kaggle (annotated images of diseased and healthy plants).
    - **Purpose**: Identify **early symptoms of plant diseases**.

Data Understanding

- **Checked missing satellite metadata** (some image timestamps missing).
- **Balanced class distribution** (healthy vs. diseased plants).
- **Examined spectral indices** (NDVI, soil moisture) in landslide-prone areas.

Data Preparation

- **Image Processing**:
    - Rescaled **satellite images** to **512x512** for uniformity.
    - Applied **segmentation to isolate landslide areas**.
- **Data Augmentation (PlantVillage)**:
    - Random **rotations, brightness changes, and contrast adjustments**.
- **Feature Selection**:
    - Kept **NDVI, soil moisture, and temperature** as key predictors.
- **Dataset Splitting**:
    - **Train (70%)**, **Validation (15%)**, **Test (15%)**.

BO4: Suggest Cure for Infected Plants

Dataset: Plant Disease Treatment & Symptoms Database

Data Acquisition

- **Source**: Agricultural research reports, public health datasets.

- **Purpose**: Recommend **effective treatments** for plant diseases.

Data Understanding

- **Mapped symptoms to diseases** based on pathology research.
- **Checked chemical treatment effectiveness** against specific pathogens.

Data Preparation

- **Data Cleaning**:
  - Removed **duplicate disease-treatment pairs**.
- **Feature Engineering**:
  - Created **"Treatment Effectiveness Score"** based on past application results.
- **Standardization**:
  - Converted **chemical dosages to uniform ppm (parts per million)**.

BO5: Predict the Appearance of Parasitic Herbs

Dataset: Toxic & Non-Toxic Parasitic Plants Database

Data Acquisition

- **Source**: Kaggle
- **Layout:** Custom dataset of **10,000 plant images** from agricultural monitoring.
- **Purpose**: Detect **harmful weeds and parasitic plants** that can be toxic to crops (induce diseases) and weeds that will just use the ressources.

Data Understanding

- **Checked class distribution** between toxic and non-toxic plants.
- **Identified duplicates** (using hash comparisons).
- **Explored variations** in plant species.

Data Preparation

- **Duplicate Image Removal**:
  - Used **hashing techniques** to find and delete **redundant images**.
- **Data Augmentation**:
  - Applied **random flipping, cropping, and brightness adjustments**.
- **Normalization**:
  - Standardized **pixel values** for deep learning models.

BO6: Segmentation of Farmers and Land for Agricultural Businesses

Datasets: Agricultural Land Prices & Farmer Segmentation Database

Data Acquisition

- **Source**:
    - **Land prices dataset** (Scrapped from Tunisie Annonce).
    - **Farmer segmentation data** (CTAB Tunisia).
- **Purpose**:
    - Identify **high-value land**.
    - Segment **farmers based on agricultural activity**.

Data Understanding

- **Checked missing price values** and inconsistencies in location names.
- **Plotted land price distributions** across Tunisia.
- **Mapped farm categories to farmer demographics**.

Data Preparation

- **Address Standardization**:
    - Cleaned location names and standardized formats.
- **Feature Engineering**:
    - Created **"Soil Quality Index"** based on historical yield data.
- **Clustering Analysis**:
    - Applied **K-Means Clustering** for farmer segmentation.

Conclusion

This part outlined the **data acquisition, understanding, and preparation** process for multiple agricultural datasets, ensuring they are clean, structured, and ready for predictive modeling.

We collected data from **satellite imagery, weather databases, soil research, market reports, and agricultural studies**, then analyzed distributions, correlations, and missing values to assess quality. Through **data cleaning, feature engineering, normalization, encoding, and augmentation**, we optimized datasets for **real-time decision-making, resource optimization, risk prediction, disease treatment recommendations, and land segmentation**.

By ensuring high-quality data, we have built a **strong foundation for AI-driven agricultural solutions**, enabling **smarter farming, improved productivity, and sustainable agricultural practices**. 🚀🌱

References

[1] https://github.com/Azure/Microsoft-TDSP/blob/master/Docs/README.md

[2] https://www.kaggle.com/datasets/arifmia/agricultural-land-suitability-and-soil-quality

[3] https://www.kaggle.com/datasets/tarundalal/dangerous-insects-dataset?resource=download